Backhaul-Aware Joint Traffic Offloading and Time Fraction Allocation for 5G HetNets

Peng-Yong Kong, *Senior Member, IEEE*, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*—In fifth-generation (5G) wireless networks, heterogeneous networks (HetNets) are used to improve capacity and users' densification. However, to rapidly deploy a large number of femto base stations (BSs) in a cost-efficient manner, they should be connected to the core network through residential broadband links. As such, different femto BSs may have different backhaul capacities, which should be taken into consideration when macro BSs offload traffic to femto BSs. Furthermore, users offloaded to femto BSs can be severely interfered by the macro BSs if macro and femto BSs transmit at the same time in the downlink. This interference can be avoided by proper transmission time scheduling. In view of the problem, this paper proposes a novel scheme to jointly perform traffic offloading and time fraction allocation (TOTFA) to ensure proportional fairness in throughput among the users. The proposed scheme, called TOTFA maximizes the weighted sum of the logarithm of user throughput, subject to constraints imposed by the backhaul capacity. Simulation results show that TOTFA can improve aggregated throughput up to 100% compared with a baseline scheme when the backhaul link capacity is 1.25 Mb/s. Over a range of different backhaul capacities, the average throughput improvement is 45%.**

***Index Terms*—Backhaul, heterogeneous networks (HetNets), time fraction allocation, traffic offloading.**

I. Introduction

We have witnessed a rapid growth in telecommunication network traffic, particularly in the mobile networks sector. In a recent report [1], CISCO has claimed that global mobile data traffic has reached 2.5 exabytes per month at the end of 2014, and it will continue to grow to 24.3 exabytes per month by 2019. This rapid traffic growth is driven by the proliferation of wireless smart devices and the popularity of media-rich wireless applications. For example, there is an increasing trend for video transmissions over mobile networks [2]. At the same time, an explosive number of sensors and automated devices will be connected to the Internet, in the vision to form a smart city in the future [3]. Supporting the anticipated huge traffic demand and a massive number of communicating devices in a cost-efficient manner is one of the key tasks of the fifth-generation (5G) wireless systems [4]. This task can be accomplished through a large-scale and dense deployment of small cells, which leads to the formation of heterogeneous networks (HetNets) [5]–[7].

In HetNets, a traditional macro base station (BS) operates in the same spectrum band with small BSs such as pico and femto BSs. These small BSs are generally low cost and low power and have a very small coverage range. In such a co-deployment architecture, each type of BSs forms a different tier [8]–[10]. In the highest tier, macrocells provide a wide-area coverage umbrella, whereas in the lower tiers, small cells are located in a more targeted but unplanned manner to alleviate coverage dead zones or to form hot spots. According to [11], this multitier network can improve capacity because, in the absence of interference, each new small BS can add extra spectral efficiency without degrading the coverage probability, which is independent of the BS densities and their transmission power values.

In the future 5G two-tier networks, the femto BSs will be installed by individual residents so that a large number of them can be rapidly deployed at a low cost. These femto BSs will be configured to provide open access instead of closed access and are connected to the core network through residential broadband links [12]. The service provider will pay for the hardware and installation cost, whereas the residents will pay for the monthly energy consumption and broadband connection. In return for the residents' contributions, the service provider will offer them priority access to the femto BSs and discount their subscription fees. This business model allows the service provider to nurture the operator–user partnership in building a better network. At the same time, the different premises may have a different type of residential broadband connections, and this will lead to a situation where the different femto BSs may have a different backhaul capacity. In general, backhaul capacity will have a significant impact on the overall network performance, and in [13], the capacity of a wireless backhaul network under saturation conditions is analyzed. Unfortunately, the results from [13] cannot be directly used in this paper because we assume that the femto backhaul links are existing residential broadband connections, whereas in [13], the recently proposed IEEE 802.11ac configuration has been assumed.

As long as the capacity of the femto backhaul link is sufficient, it is desirable to offload traffic from a heavily loaded macro BS to a nearby lightly loaded femtocell. This process

Manuscript received June 22, 2015; revised September 29, 2015 and December 12, 2015; accepted January 6, 2016. Date of publication January 13, 2016; date of current version November 10, 2016. The review of this paper was coordinated by Prof. W. Song.

P.-Y. Kong is with the Department of Electrical and Computer Engineering, Khalifa University of Science, Technology and Research, Abu Dhabi, United Arab Emirates (e-mail: pengyong.kong@kustar.ac.ae).

G. K. Karagiannidis is with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54636 Thessaloniki, Greece, and also with the Department of Electrical and Computer Engineering, Khalifa University of Science, Technology and Research, Abu Dhabi, United Arab Emirates (e-mail: geokarag@auth.gr).

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/TVT.2016.2517671

improves the throughput for both the offloaded user equipment (UE) and the other UEs that stay in the macrocell. This improvement is achieved as a result of the better load balancing, such that all UEs can access more resources that may appear in the forms of transmission time slots, bandwidth, etc. In handling offloaded UEs, it was shown in [12] that allocating dedicated bandwidth to femto BSs can lead to a better performance, as compared with sharing a common bandwidth between different network tiers.

In real networks, traffic offloading is performed by adding a cell association bias to the received signal reference power of a femto BS, such that a UE selects the femto BS for association, even if its received signal power is weaker than that of the macro BS. In the literature, this traffic offloading mechanism is called *cell range expansion* because this bias has an effect of expanding the coverage range of a femtocell [14]. With the association bias, traffic offloading comes with a side effect, where UEs are no longer necessarily connected to the BS with the highest signal strength. This exposes femto UEs to severe interference from macro BSs in the downlink when femto and macro BSs transmit simultaneously [15]. To deal with this issue, enhanced intercell interference coordination (eICIC) techniques must be used [16], [17].

According to [18], eICIC can be performed by scheduling transmissions at different time slots, such that interference at receivers is mitigated. For instance, almost blank subframes (ABSFs) can be scheduled in downlink transmissions [19]. Within each ABSF, macro BSs are silent so that the femto BS can transmit to its associated UEs without interference from the macro BSs. This technique is similar to transmission time fraction allocation, where each transmitter is allocated a fraction of a frame duration period, within which interference from other transmitters can be avoided. However, focusing only on time fraction allocation can lead to serious unfairness among different UEs, i.e., a UE with a better link quality is allocated a larger time fraction to maximize aggregated throughput. This unfairness is demonstrated in [20] in the context of wireless local area networks, where time fraction allocation has been done in a distributed manner through the IEEE 802.11 configuration. In view of the unfairness issue, instead of solely maximizing the throughput, one should aim to achieve proportional fairness among all UEs because doing so will implicitly find a balance between maximizing the throughput and being fair [21], [22].

In this paper, we aim to achieve proportional fairness among UEs in a HetNet while allocating time fractions for transmissions and adjusting cell association biases for traffic offloading. Furthermore, we aim to achieve this goal, assuming different femto BSs may have different backhaul link capacities due to their individual choices of residential broadband connections.

A. Related Literature

In the literature, in [23], the benefits of spectrum sensing by femto BSs in a two-tier HetNet, is analyzed, and it is concluded that it is important to look beyond system-level performance. In addition to the overall network performance, it is important to understand the performance of the individual components such as different network tiers and individual users. This is consistent with the goal of this paper, which is not to only maximize the system throughput but also to be fair to individual UEs so that each of them can attain a fair share of throughput.

Traffic offloading can be also performed by adding association bias to the received signal-to-interference-and-noise ratio (SINR) [24]. This is different from the case we consider in this paper, where the offloading decision is based on the actual received signal reference power plus an association bias. In practice, a good cell selection scheme should take into account both signal strength and signal quality (e.g., SINR) but not only just one of them. For example, in a case where the signal qualities from two candidate BSs are similar, their signal strengths will be needed to further differentiate the two. While signal quality has a more direct effect on the data of a link, there is no fundamental difference between the concept of applying association bias to signal strength and signal quality, and we do not differentiate them while surveying related literature.

Through extensive simulations, in [25], the effect of different association bias settings is studied, and it is shown that only a moderate bias is effective in enhancing capacity and throughput. This is because an overly large bias value may affect the transmissions in control channels, which has a significant impact on the overall performance. Within an analytical framework, in [26], the optimal cell association bias was determined by solving a network-wide utility maximization problem using convex optimization. In a separate work, in [27], an algorithm was proposed to adjust cell association bias to minimize energy consumption while supporting multiple traffic classes with different packet delay requirements. However, in [25]–[27], time fraction allocation to ensure fairness among different UEs within a HetNet was not considered.

For time fraction allocations in eICIC, in [28], the number of ABSFs required in each frame for a given cell association bias has been studied. Ideally, given the interplay between association bias and interference coordination, traffic offloading and time fraction allocation (TOTFA) should be performed jointly but not separately for different objectives. By adjusting both transmission time fraction and cell association bias for a common performance goal, we expect to achieve better fairness and a higher throughput at the same time.

In [29], optimal ABSF density has been determined after jointly considering UE association and time fraction allocation. In this paper, UE association has been done individually on each user but not on each BS. This contradicts the convention described earlier [14], where cell association bias is a parameter assigned to each BS but not UE. Without using the cell association bias, it is unclear what is the scalable practical mechanism that can associate a large number of UEs to various BSs. Through a semianalytical approach, in [30], the cumulative distribution function (CDF) of the downlink SINR has been used to capture the collective effects of association bias and time fraction allocation. As the CDF changes in response to adjustment of the bias value and transmission time allocation, the authors have suggested a guideline to set the association bias for a desired capacity and fairness. In [30], fairness has been measured by the sum of link capacity of some UEs but not the per-UE proportional fairness. For a different objective to maximize the sum utility of long-term data rate, in [31], a joint user association and resource-allocation scheme has been proposed. However, similar to [29], user association in [31] has been done individually on each UE and not according to the per-BS association bias.

The joint user association and resource-allocation scheme in [32] has aimed to achieve the per-UE proportional fairness. However, the resource is not the time fraction but the BS transmission power. In addition, similar to [29] and [31], association has been performed based on each UE but not per-BS. The work in [33] has jointly performed traffic offloading and transmission time fraction allocation for achieving proportional fairness. In a broad sense, it investigates the same problem as in this paper. Through optimization based on a real-world radio coverage map, it has been shown in [33] that significant throughput improvement can be achieved, as compared with that of using a uniform association bias for all femto BSs. However, in [33], the effects of limited residential broadband backhaul link capacity have not been considered.

B. Motivation and Contribution

Based on the related literature presented in the previous section, there is a clear need to develop a backhaul-aware method to jointly control cell association bias and transmission time fraction in achieving proportional fairness among UEs. The proposed method should follow the convention, where traffic offloading is done through association bias assigned to each BS but not individual UEs.

In this paper, we propose a novel scheme, called TOTFA, to jointly perform traffic offloading and transmission time fraction allocation in 5G HetNets while ensuring proportional fairness among UEs, subject to constraints imposed by the backhaul link capacity. In summary, the proposed scheme

- jointly controls cell association bias for traffic offloading and transmission time fraction allocation for fairness;
- exploits spatial reuse in time fraction allocation, such that multiple noninterfering femto BSs can transmit in a same time slot;
- considers the limited backhaul link capacity, particularly for unplanned femto BSs, which are likely deployed by home users, by utilizing their own residential broadband connections;
- adopts a unique combination of heuristics in deciding cell association biases and Lagrangian approach in finding the optimal time fraction allocation.

C. Structure

The remainder of this paper is organized as follows. Section II describes the system model. In Section III, the proposed TOTFA scheme is described in detail. Section IV presents evaluation results and discussions before this paper ends with concluding remarks in Section V.

II. System Model

We consider a two-tier HetNet and focus on the downlink transmissions. For simplicity, we assume that the network consists of a macrocell and multiple open-access femtocells, although the method developed in this paper can be easily extended to the case of multiple macrocells and more femtocells. For scalability, this extension to cover a bigger network will be achieved in such a way that a copy of the proposed algorithm takes care of only a macrocell and its femtocells while multiple neighboring macrocells operate in different frequency bands with sufficient frequency reuse factor to avoid cochannel interference. This way, there will be multiple copies of the computer program running at a same time at different macro BSs. When cochannel interference cannot be completely eliminated in a scenario with more random macrocells deployment, each copy of the proposed algorithm will take care of a cluster of macrocells, and different copies of the program need to coordinate among themselves to avoid interference between clusters.

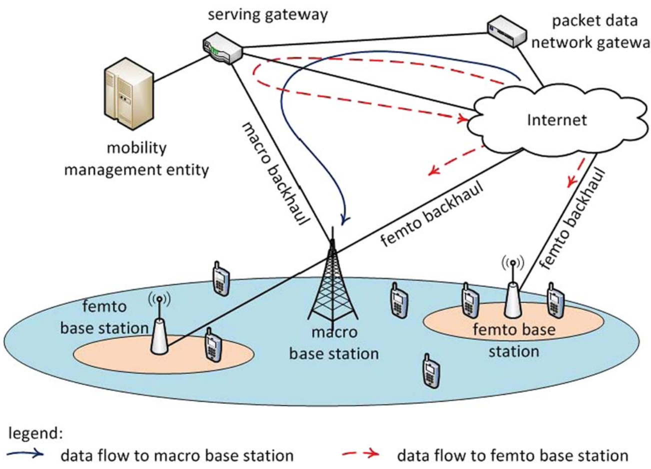


Fig. 1. System model for one macro BS and multiple femto BSs. Data traffic that is directed to femto BSs needs to re-enter the Internet.

In our model, within the macrocell, femtocell locations are randomly distributed, whereas a BS is located at the center of each cell. Let $\mathcal{F}$ be the set of femto BSs, where the cardinality $|\mathcal{F}|$ indicates the number of femto BSs within the macrocell. Then, BS 0 refers to the macro BS, and BS $i = 1, 2, \ldots, |\mathcal{F}|$ refers to the ith femto BS in $\mathcal{F}$. In addition, the set $\{0, \mathcal{F}\}$ includes all BSs.

As shown in Fig. 1, the traffic from the Internet enters the HetNet through the packet data network gateway. Within the network, data traffic is directed to BSs through the serving gateway, which is also connected to the mobility management entity that keeps track of the associated BSs of each UE. The serving gateway is connected to each macro BS through a dedicated backhaul link, which is controlled by the network operator. On the contrary, there is no dedicated link between the serving gateway and each femto BS. Traffic heading to femto BSs needs to re-enter the Internet before reaching them. Each femto BS is connected to the Internet using individual residential broadband connections.

Different residents may have subscribed to different broadband services and, thus, have different types of connections. We define b_i as the backhaul link capacity for the ith BS. Since the femto backhaul is not a dedicated link, its data rate can be time varying. Thus, we assume that b_i is the guaranteed component of the time-varying capacity. The idea of separating a time-varying link capacity into a guaranteed and a nonguaranteed

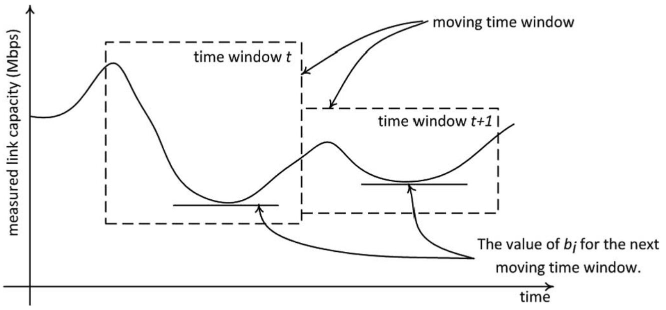

measured link capacity (Mbps)
moving time window
time window t
time window $t+1$
The value of b_i for the next moving time window.
time

Fig. 2. Measurement of backhaul link capacity for femto BSs.

component has been explored in a different context in [34] to provide delay guarantee to real-time packets. As shown in Fig. 2, we determine b_i as the minimum value of multiple measurements of the link capacity within a moving time window. Each measurement can be performed using the single-end probe method [35].

We use P_i to denote the transmission power of BS i, and therefore, $P_0 > P_i \ \forall i \in \mathcal{F}$ because femto BSs transmit at lower power compared with macro BSs. Let $l_{i,j}$ be the distance between UE i and BS j and α_j be the path loss exponent for BS j. The received signal strength at UE i from BS j is given by $P_j l_{i,j}^{-\alpha_j}$. When a UE has to select a suitable BS for association, it may choose the one with the maximum received signal strength. This simple selection strategy can lead to underutilization of femtocells because their BSs transmit at very low power compared with the macro BS. To provide flexibility and control in offloading traffic from the macrocell to femto BSs, Long-Term Evolution-Advanced (LTE-A) has introduced a control parameter called cell association bias [14], [24]. This bias adds an offset to the received power so that the perceived value becomes higher than the actual value. Let y_j (in decibels) be the cell association bias for BS j and α (in decibel-milliwatts) be the minimum power level needed to correctly detect the signal. In practice, α must be at least 3 dB above the measured noise power. Then, UE i will select BS j^* for association where j^* is determined as

$$j^* = \arg \max_{j\in\{0,\mathcal{F}\}} \left\{ 10 \log \frac{10^3 P_j}{l_{i,j}^{\alpha_j}} + y_j \;\middle|\; 10 \log \frac{10^3 P_j}{l_{i,j}^{\alpha_j}} > \alpha \right\} \tag{1}$$

where the operation is conditional to ensure that a UE is not associated with a BS from which it does not receive a sufficiently strong signal power. According to (1), we can offload more traffic to BS j by using a larger y_j. This is because a larger y_j means that the UE is more likely to be connected to BS j although the actual received signal power from BS j is not as high as the signal power received from other BSs. Equation (1) also implies that each UE can only be associated with one BS at any one moment.

Let $\mathcal{U}_j$ be the set of all UEs associated with BS j. The BS allocates a nonoverlapping transmission time fraction x_i to each UE $i \in \mathcal{U}_j$, such that the transmission time fraction of BS j can be determined as

$$T_j = \sum_{i\in\mathcal{U}_j} x_i. \tag{2}$$

In the context of the LTE-A, T_j is the time fraction of a frame allocated to BS j, where each frame consists of ten subframes and each subframe has a length of 1 ms. The total transmission time fraction allocated to all femto BSs makes up the proportion of ABSF in a frame. If there is no spatial reuse, only one femto BS is allowed to transmit at each time and the allocated time fractions for different femto BSs must not overlap. Therefore

$$\sum_{j\in\{0,\mathcal{F}\}} T_j \leq 1. \tag{3}$$

Based on $\mathcal{U}_j$, we further define $\mathcal{U} = \cup_{j\in\{0,\mathcal{F}\}} \mathcal{U}_j$ as the set of all UEs in the system, and the cardinality $|\mathcal{U}|$ is the total number of UEs.

III. Traffic Offloading and Time Fraction Allocation

As aforementioned, the goal of this paper is to achieve proportional fairness in throughput by controlling the backhaul-aware cell association bias of each BS and allocating the transmission time fraction for each UE. Here, we describe the proposed scheme, i.e., TOTFA, that selects the set of optimal association biases and time fractions.

In TOTFA, proportional fairness is achieved by maximizing the weighted sum of the logarithm of individual UE's throughput as

$$\max \sum_{i\in\mathcal{U}_j; j\in\{0,\mathcal{F}\}} w_i \log(u_i) \tag{4}$$

where w_i represents a weight associated with UE i, and u_i is the throughput for UE i with

$$u_i = x_i r_i \tag{5}$$

where x_i has been defined earlier as the transmission time fraction allocated to UE i, and r_i is the data rate of UE i. The data rate can be written as

$$r_i = \sum_{j\in\{0,\mathcal{F}\}} \phi_{i,j} W_j \log_2(1+\gamma_{i,j}) \tag{6}$$

where W_j is the bandwidth for BS j, $\gamma_{i,j}$ is the received SINR at UE i from BS j, and $\phi_{i,j}$ is a BS selection parameter. Here, $\phi_{i,j}$ takes the value 1 if UE i is associated with BS j and takes the value 0, otherwise. Parameter $\phi_{i,j}$ depends on y_j as given in (1). By changing y_j, we can alter which BS that a UE is connected to. Then, the BS selection parameter is

$$\phi_{i,j} = \begin{cases} 1, & \text{if } j = j^* \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The SINR $\gamma_{i,j}$ depends on the UE's location relative to the BS. Let N_o be the noise power. Then, the respective SINR is

$$\gamma_{i,j} = \frac{P_j l_{i,j}^{-\alpha_j}}{\sum_{k\neq j} P_k l_{i,k}^{-\alpha_k} + N_o}. \tag{8}$$

In (8), the interference comes from concurrently transmitting femto BSs.

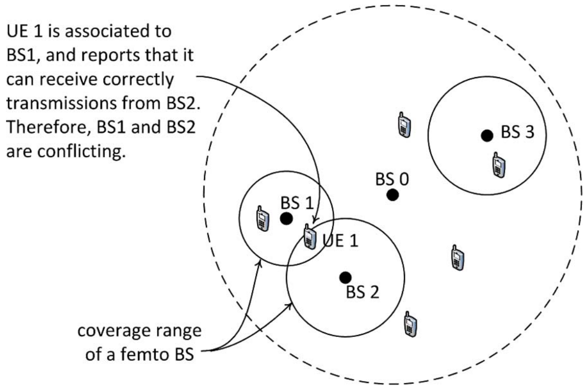

UE 1 is associated to BS1, and reports that it can receive correctly transmissions from BS2. Therefore, BS1 and BS2 are conflicting.
BS 3
BS 0
BS 1
UE 1
BS 2
coverage range of a femto BS

Fig. 3. Formation of nonconflicting groups, i.e., $\mathcal{G}_1 = \{0\}$, $\mathcal{G}_2 = \{1, 3\}$, and $\mathcal{G}_3 = \{2\}$. All BSs in a group can transmit at a same time without causing interference to each other. For example, BS 1 and BS 3 can transmit concurrently.

After finding r_i in (6), we still need to know the transmission time allocation before we can determine the throughput using (5). To allocate downlink transmission time for all UEs, we exploit the fact that multiple femto BSs may simultaneously transmit as long as they are sufficiently separated far apart and do not cause significant interference to each other. We propose grouping femto BSs into nonconflicting groups, where all BSs in a group can transmit in the same time slot. For downlink transmissions, the interference is measured at the UE but not at the BS. Therefore, as shown in Fig. 3, two BSs are considered conflicting if there is a UE located such that it can receive sufficiently strong signals from both BSs.

Practically, each UE i needs to report to its associated BS j a list $\mathcal{L}_{j,i}$ of all other BSs from which it can correctly receive a transmission. The BS j will merge the reported lists from all its UEs such that $\mathcal{L}_j = \cup_{i \in \mathcal{U}_j} \mathcal{L}_{j,i}$. A nonempty list $\mathcal{L}_j$ contains the conflicting BSs of BS j. The list is further reported by the femto BS j to the macro BS. This reporting is done regularly at the same timescale as the one we use to estimate backhaul link capacity in Fig. 2. With the lists from all femto BSs, the macro BS forms the nonconflicting groups by using Algorithm 1. Let $\mathcal{G}_i$ be the ith nonconflicting group, and each BS can be a member of one and only one group. The algorithm reserves the first group, i.e., $\mathcal{G}_1$, for the macro BS only. All BSs without a group assignment are candidates of a new group. A group is formed by recursively transferring a new candidate BS into the group if and only if the candidate does not conflict with all its existing members. This conflict is verified based on $\mathcal{L}_j$, which has been reported by each BS j. When a group is formed but there are remaining candidates, a new group is started. Therefore, Algorithm 1 will naturally stop when there is no more candidate BS without a group assignment.

Algorithm 1 Nonconflicting group forming algorithm

```
1: Initialize $\mathcal{F}'$ to include all femto BSs. Initialize $i = 1$.
2: Set $\mathcal{G}_1 = \{0\}$.
3: If $\mathcal{F}'$ is empty, do step 8. Otherwise, continue with step 4.
4: Set $i = i + 1$.
5: For each member $j \in \mathcal{F}'$, do step 6.
6: If $j \notin \mathcal{L}_k \; \forall \, k \in \mathcal{G}_i$, do the following:
       $\mathcal{G}_i = \mathcal{G}_i \cup \{j\}$.
       $\mathcal{F}' = \mathcal{F}' \setminus \{j\}$.
7: Go to step 3.
8: Stop.
```

For each ith nonconflicting group $\mathcal{G}_i$, we define an effective transmission time fraction τ_i as

$$\tau_i = \max_{j \in \mathcal{G}_i} \{T_j\}. \tag{9}$$

Let $\mathcal{G}$ be the set of all nonconflicting groups, such that its ith member is $\mathcal{G}_i$. Then, since all BSs in a nonconflicting group may transmit at the same time, (3) can be rewritten as

$$\sum_{i \in \mathcal{G}} \tau_i \leq 1. \tag{10}$$

With the relations among user throughput, cell association bias, and transmission time fraction defined so far in this paper, we can now state the proportional fairness optimization problem. First, we define a vector $\mathbf{x}$ to include transmission time fractions for all UEs such that $\mathbf{x} = [x_1, x_2, \ldots, x_{|\mathcal{U}|}]$. In addition, we define another vector $\mathbf{y}$ to include association biases for all BSs such that $\mathbf{y} = [y_0, y_1, \ldots, y_{|\mathcal{F}|}]$. Then, the optimization problem can be written as

$$\max_{\mathbf{x},\mathbf{y}} \quad \sum_{i \in \mathcal{U}} w_i \log(x_i r_i) \tag{11}$$

$$\begin{aligned} \text{s.t.} \quad & 0 \leq x_i \leq 1 \; \forall \, i \in \mathcal{U} \\ & y_j \geq 0 \; \forall \, j \in \{0, \mathcal{F}\} \\ & \sum_{i \in \mathcal{U}_j} x_i r_i \leq b_j \; \forall \, j \in \{0, \mathcal{F}\} \\ & \sum_{k \in \mathcal{G}} \max_{j \in \mathcal{G}_k} \sum_{i \in \mathcal{U}_j} x_i \leq 1. \end{aligned} \tag{12}$$

The third constraint of (12) is necessary to ensure that all the UEs of a BS are not allocated transmission time fractions, which will result in an aggregated throughput, that is larger than the BS's backhaul link capacity.

The optimization problem (11) is not easy to solve because r_i depends on $\mathbf{y}$ through $\phi_{i,j}$, which is discrete. TOTFA deals with the problem by breaking it into two steps. In the first step, we find a vector $\mathbf{y}^*$ through heuristics so that the second step does not depend on the discrete variables and, thus, can find the optimal transmission time fractions more easily.

To find $\mathbf{y}^*$, we let the association bias of the macro BS, i.e., $y_0 = 0$. This is a realistic consideration because the purpose of introducing association bias is to offload traffic to femtocells, and thus, there is no need for the macro BS to compete with the femto BSs for more UEs by increasing its association bias. For other BSs in the system, their association biases are determined using Algorithm 2. In this algorithm, given the association bias vector $\mathbf{y}$, the function

$$f_j(\mathbf{y}) = \sum_{i \in \mathcal{U}_j} r_i(\mathbf{y}) \tag{13}$$

determines the aggregated data rate of all UEs associated to BS j.

Now, to fairly distribute traffic load among different femto BSs with different backhaul capacities, the aggregated data rate at a BS j must not exceed a fraction, i.e., $b_j / \sum_{i\in\{0,\mathcal{F}\}} b_i$, of the total aggregated data rates of all BSs, i.e., $\sum_{i\in\{0,\mathcal{F}\}} f_i(\mathbf{y})$. Therefore, in a round-robin manner, Algorithm 2 successively increases the association bias for as much as Δ, for a BS j, only if doing so does not cause the adjusted $f_j(\mathbf{y})$ to exceed a backhaul-dependent proportion ℓ_j, which is determined as

$$\ell_j = \frac{b_j \sum_{i\in\{0,\mathcal{F}\}} f_i(\mathbf{y})}{\sum_{i\in\{0,\mathcal{F}\}} b_i}. \tag{14}$$

Since each BS is examined once in each round, the algorithm stops when none of the BSs has its adjusted aggregated rate $f_j(\mathbf{y})$ lower than or equal to ℓ_j.

Algorithm 2 Cell association bias algorithm

```
1: Initialize $y_0 = 0$ and $y_i = 0; \forall\, i \in \mathcal{F}$.
2: For each member $j \in \mathcal{F}$, do steps 3 to 5.
3: Set $\mathbf{y}' = [y_0, \ldots, y_j + \Delta, \ldots, y_{|\mathcal{F}|}]$.
4: Compute $\ell_j = b_j \sum_{i\in\{0,\mathcal{F}\}} f_i(\mathbf{y}') / \sum_{i\in\{0,\mathcal{F}\}} b_i$.
5: If $f_j(\mathbf{y}') \leq \ell_j$, do the following:
     $y_j = y_j + \Delta$.
6: If none of the $y_i; \forall\, i \in \mathcal{F}$ has been updated, continue with
   step 7. Otherwise, go to step 2.
7: Set $\mathbf{y}^* = [y_0, y_1, \ldots, y_{|\mathcal{F}|}]$.
8: Stop.
```

With $\mathbf{y}^*$ determined in Algorithm 2, we can now rewrite (11) so that it can be solved in the second step. Let r_i^* be the data rate of UE i given the association bias vector $\mathbf{y}^*$. Then, the optimization problem is given by

$$\max_{\mathbf{x}} \quad \sum_{i\in\mathcal{U}} w_i \log\left(x_i r_i^*\right) \tag{15}$$

$$\text{s.t.} \quad 0 \leq x_i \leq 1\; \forall\, i \in \mathcal{U}$$

$$\sum_{i\in\mathcal{U}_j} x_i r_i^* \leq b_j\; \forall\, j \in \{0, \mathcal{F}\} \tag{16}$$

$$\sum_{j\in\mathcal{E}} \sum_{i\in\mathcal{U}_j} x_i \leq 1 \tag{17}$$

$$\sum_{i\in\mathcal{U}_j} x_i \leq \sum_{i\in\mathcal{U}_{\mathcal{E}_k}} x_i\; \forall\, j \in \mathcal{G}_k. \tag{18}$$

Constraint (17) is an alternative representation of the final constraint of (11). In (17), $\mathcal{E}$ is a set where its ith member, i.e., $\mathcal{E}_i$ is a BS from the ith nonconflicting group $\mathcal{G}_i$, and this BS is determined as

$$\mathcal{E}_i = \arg\max_{j\in\mathcal{G}_i} \{T_j\}. \tag{19}$$

Constraint (18) is necessary to ensure that the transmission time fraction of a BS is upper bounded by the largest transmission time fraction of its nonconflicting group.

The optimization problem in (15) can be solved by the augmented Lagrangian approach [36]. We define the Lagrangian as

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \eta) = &\sum_{i\in\mathcal{U}} w_i \log\left(x_i r_i^*\right) + \sum_{j\in\{0,\mathcal{F}\}} \lambda_j \left(b_j - \sum_{i\in\mathcal{U}_j} x_i r_i^*\right) \\ &+ \sum_{k\in\mathcal{G}} \sum_{j\in\mathcal{G}_k} \mu_{k,j} \left(\sum_{i\in\mathcal{U}_{\mathcal{E}_k}} x_i - \sum_{i\in\mathcal{U}_j} x_i\right) \\ &+ \eta \left(1 - \sum_{j\in\mathcal{E}} \sum_{i\in\mathcal{U}_j} x_i\right) \end{aligned} \tag{20}$$

where time fraction allocations $\mathbf{x}$ are called the primal variables. On the other hand, λ_j, $\mu_{k,j}$, and η are dual variables, and they are also called the Lagrange multipliers. All Lagrange multipliers must be positive real numbers. We use $\lambda = [\lambda_0, \lambda_1, \ldots, \lambda_{|\mathcal{F}|}]$ to denote the vector of λ_j. Similarly, we use $\boldsymbol{\mu}$ to denote the vector of $\mu_{k,j}$. For simplicity, we further use $\mathbf{z}$ to represent the vector of all dual variables such that $\mathbf{z} = [\boldsymbol{\lambda}, \boldsymbol{\mu}, \eta]$.

The original constrained optimization problem can now be transformed into the following nonconstrained optimization problem:

$$D(\mathbf{z}) = \min_{\mathbf{z}\geq 0} \max_{\mathbf{x}\in\mathbf{X}} L(\mathbf{x}, \mathbf{z}) \tag{21}$$

where $\mathbf{X}$ is the feasible space for the primal variables. All the solutions to $D(\mathbf{z})$ are forced to lie within $\mathbf{X}$. The choice of the space can affect the solution and the speed in finding the solution. We define the space $\mathbf{X}$ as

$$\mathbf{X} = \left\{\mathbf{x} : 0 \leq x_i \leq 1\; \forall\, i \in \mathcal{U}, \sum_{i\in\mathcal{U}} x_i \leq 1\right\}. \tag{22}$$

This feasible solution space is larger than that constrained by (17) alone so that we can explore a solution compromising various constraints when there is a duality gap.

The problem $D(\mathbf{z})$ can be iteratively solved until a convergence is achieved, where each iteration consists of two steps, namely, greedy primal update step and subgradient descent dual update step. In the greedy primal update, the primal variable in iteration t, i.e., $\mathbf{x}_t$, is updated as

$$\mathbf{x}_t = \arg\max_{\mathbf{x}\in\mathbf{X}} L(\mathbf{x}_{t-1}, \mathbf{z}_{t-1}). \tag{23}$$

Equation (23) requires finding the value of each x_i that maximizes the Lagrangian, given other primal and dual variables' values from the previous iteration. If the Lagrangian is differentiable by x_i, the maximum value is achieved when its derivative is zero. The partial derivative is determined as

$$\frac{\partial L(\mathbf{x}, \mathbf{z})}{\partial x_i} = \frac{w_i}{x_i} - \left(\lambda_j r_i^* + \mu_{k,j} + \eta\right). \tag{24}$$

This simple form of the partial derivative can be obtained with the fact that each UE i can only be associated with one BS j.

In (24), $\mu_{k,j}$ indicates that UE i is associated to BS j, which is a member of the nonconflicting group $\mathcal{G}_k$. Therefore, $\mu_{k,j} = 0$ when $j = \mathcal{E}_k$. In addition, $\eta = 0$ when $j \neq \mathcal{E}_k$. Let $x_i(t)$ be the value of the time fraction for iteration t. Similarly, $\lambda_j(t)$, $\mu_{k,j}(t)$, and $\eta(t)$ are respectively the dual variable values at iteration t. Then, we can find the primal value that maximizes the Lagrangian at iteration $t+1$ as

$$x_i(t+1) = \begin{cases} \dfrac{w_i}{\lambda_j(t) r_i^* + \eta(t)}, & \text{if } j = \mathcal{E}_k \\ \dfrac{w_i}{\lambda_j(t) r_i^* + \mu_{k,j}(t)}, & \text{otherwise.} \end{cases} \tag{25}$$

In subgradient descent dual update [37], the updated primal values from (25) are used together with the dual values from previous iteration as

$$\mathbf{z}_t = [\mathbf{z}_{t-1} + \delta \mathbf{g}(\mathbf{x}_t)]^+ \tag{26}$$

where $\mathbf{g}(\mathbf{x})$ is a vector of the three constraints (16)–(18) written in such a way that $\mathbf{g}(\mathbf{x}) \leq 0$. Defining $\mathbf{g}(\mathbf{x})$ in such a way is essential because it is desired that the dual variables are to be updated to bring down the value of the Lagrangian only when the constraints are enforced. In (26), δ is the step size for updating the dual variables, and $[\cdot]^+$ is componentwise projection into the space of nonnegative real numbers. Due to the fact that each UE i can only be associated with a BS j, we can update the dual variable $\lambda_j(t+1)$ at iteration $t+1$ as

$$\lambda_j(t+1) = \left[\lambda_j(t) + \delta\left(\sum_{i \in \mathcal{U}_j} x_i(t+1)\frac{r_i^*}{b_j} - 1\right)\right]^+ \tag{27}$$

where we use $\sum_{i \in \mathcal{U}_j} x_i r_i^*/b_j - 1$ instead of $\sum_{i \in \mathcal{U}_j} x_i r_i^* - b_j$ to ensure a similar dynamic range for all dual variables. This is important because all the dual updates share the same step size δ.

In addition to λ_j, each BS $j \neq \mathcal{E}_k$ of a nonconflicting group $\mathcal{G}_k$ has to update another dual variable at iteration $t+1$ as

$$\mu_{k,j}(t+1) = \left[\mu_{k,j}(t) + \delta\left(\sum_{i \in \mathcal{U}_j} x_i(t+1) - \sum_{i \in \mathcal{U}_{\mathcal{E}_k}} x_i(t+1)\right)\right]^+. \tag{28}$$

On the other hand, for a BS $j = \mathcal{E}_k$, $\mu_{k,j}$ is not needed and the other dual variable is updated at iteration $t+1$ as

$$\eta(t+1) = \left[\eta(t) + \delta\left(\sum_{j \in \mathcal{E}} \sum_{i \in \mathcal{U}_j} x_i(t+1) - 1\right)\right]^+. \tag{29}$$

The proposed TOTFA scheme uses Algorithm 3 to perform the greedy primal update and subgradient descent dual update. After initializing the algorithm, primal and dual variables are iteratively updated until the value of the Lagrangian converges within a small variation ϵ. Alternatively, the algorithm will be carried out for a maximum number of iterations. The outcome of this algorithm is the set $\mathbf{x}^*$ of the optimal transmission time fraction allocations for all UEs. Together with $\mathbf{y}^*$ determined earlier, TOTFA has now completed the job of finding the optimal cell association bias and transmission time fraction to achieve proportional fairness in a HetNet.

Algorithm 3 Primal and dual variables updating algorithm

```
1: Initialize primal variables $x_i \ \forall i \in \mathcal{U}$ to any feasible value
   within $\mathbf{X}$. Initialize each dual variable to a random nonnegative
   real number. Initialize iteration $t = 0$.
2: If the number of maximum iterations is not exceeded, do
   steps 3 to 6. Otherwise, go to step 7.
3: Set $t = t + 1$.
4: Primal update:
   For each $i \in \mathcal{U}$,
      If $i$ is associated with BS $j \in \mathcal{E}_k \ \forall k$
         $x_i(t) \leftarrow \frac{w_i}{\lambda_j(t-1) r_i^* + \eta(t-1)}$.
      If $i$ is associated with BS $j \notin \mathcal{E}_k \ \forall k$
         $x_i(t) \leftarrow \frac{w_i}{\lambda_j(t-1) r_i^* + \mu_{k,j}(t-1)}$.
5: Dual update:
   For each BS $j \in \{0, \mathcal{F}\}$
      $\lambda_j(t) \leftarrow \left[\lambda_j(t-1) + \delta\left(\sum_{i \in \mathcal{U}_j} x_i(t)\frac{r_i^*}{b_j} - 1\right)\right]^+$.
   For each nonconflicting group $\mathcal{G}_k$, identify $\mathcal{E}_k$ as follows:
      $\mathcal{E}_k \leftarrow \arg\max_{j \in \mathcal{G}_k} \left\{\sum_{i \in \mathcal{U}_j} x_i(t)\right\}$.
      If BS $j = \mathcal{E}_k$
         $\eta(t) \leftarrow \left[\eta(t-1) + \delta\left(\sum_{j \in \mathcal{E}} \sum_{i \in \mathcal{U}_j} x_i(t) - 1\right)\right]^+$.
      If BS $j \neq \mathcal{E}_k$
         $\mu_{k,j}(t) \leftarrow \left[\mu_{k,j}(t-1) + \delta\left(\sum_{i \in \mathcal{U}_j} x_i(t) - \sum_{i \in \mathcal{U}_{\mathcal{E}_k}} x_i(t)\right)\right]^+$.
6: Check for convergence:
   Diff = $|L(\mathbf{x}(t), \mathbf{z}(t)) - L(\mathbf{x}(t-1), \mathbf{z}(t-1))|$.
   If Diff $\leq \epsilon$, go to step 7. Otherwise, go to step 2.
7: Set $\mathbf{x}^* = \mathbf{x}$.
8: Stop.
```

In practice, TOTFA is regularly executed at the timescale we use to determine the backhaul link capacity in Fig. 2. After each moving time window, backhaul link capacities are estimated at the same time Algorithm 1 is performed to construct nonconflicting groups. Then, Algorithm 2 is carried out to determine the association bias vector before Algorithm 3 finds the time fraction allocation vector. Since femtocells are most suited for slow-moving users in a home environment and wired home broadband link does not change rapidly, we may perform the optimization process once every few minutes.

TABLE I

COMPARISON OF SOLUTIONS FROM TOTFA AND EXHAUSTIVE SEARCH. (a) $\alpha_0 = 2$, $\alpha_1 = \alpha_2 = \alpha_3 = 3$, AND $b_0 = b_1 = b_2 = b_3 = 10$ Mb/s; (b) $\alpha_0 = 2$, $\alpha_1 = \alpha_2 = \alpha_3 = 3$, $b_0 = 10$ Mb/s, AND $b_1 = b_2 = b_3 = 100$ Mb/s; (c) $\alpha_0 = \alpha_1 = \alpha_2 = \alpha_3 = 2$, $b_0 = 10$ Mb/s, AND $b_1 = b_2 = b_3 = 100$ Mb/s

	y_0	y_1	y_2	y_3	x_1	x_2	x_3	x_4	x_5	$\sum w_i \log(x_i r_i)$
Exhaustive search	0	0	0	0	0.1800	0.2600	0.1800	0.1800	0.1600	34.30
TOTFA	0	0	0	0	0.1937	0.1937	0.1945	0.1937	0.1521	34.25

(a)

	y_0	y_1	y_2	y_3	x_1	x_2	x_3	x_4	x_5	$\sum w_i \log(x_i r_i)$
Exhaustive search	0	0	0	0	0.4400	0.4400	0.1800	0.2200	0.1600	35.00
TOTFA	0	0	0	0	0.3285	0.3285	0.1815	0.3285	0.1614	34.93

(b)

	y_0	y_1	y_2	y_3	x_1	x_2	x_3	x_4	x_5	$\sum w_i \log(x_i r_i)$
Exhaustive search	0	0	0	0	0.4400	0.4400	0.1800	0.2200	0.1600	35.35
TOTFA	0	0	0	0	0.3285	0.3285	0.1815	0.3285	0.1614	35.28

(c)

IV. PERFORMANCE EVALUATION

We have evaluated the proposed TOTFA through simulations. In the simulations, we set the transmission power of macro BS, i.e., $P_0 = 20$ W, and the transmission power of femto BS, i.e., $P_j = 0.2$ W $\forall j \in \mathcal{F}$. In addition, the path loss exponent for macrocell is $\alpha_0 = 2$, whereas the path loss exponent for femtocells is $\alpha_j = 3$ $\forall j \in \mathcal{F}$. Each macro and femto BS has a bandwidth $W_j = 5$ MHz $\forall j \in \{0, \mathcal{F}\}$. All transmissions are affected by a noise power $N_o = 10^{-6}$ W, and the threshold for useful signal power $\alpha = 2N_o$. For simplicity, we assume that all UEs have the same proportional fairness weight such that $w_i = 1$ $\forall i \in \mathcal{U}$. For Algorithm 2, $\Delta = 0.001$ dB. For Algorithm 3, $\delta = 0.05$ and $\epsilon = 0.0005$. In addition, the maximum number of iterations allowed in Algorithm 3 is 10 000. These typical values must be assumed, unless stated otherwise.

First of all, we want to verify the optimality of solution $\{\mathbf{x}^*, \mathbf{y}^*\}$ identified by TOTFA. For this purpose, we benchmark the performance of TOTFA against that of an exhaustive search within a feasible space defined by $\mathbf{X}$ and $\mathbf{Y}$. While $\mathbf{X}$ has been defined earlier, $\mathbf{Y}$ is given as

$$\mathbf{Y} = \left\{\mathbf{y} : 0 \leq y_j \leq 2 \times 10^{-2} \, \forall j \in \{0, F\}\right\}. \qquad (30)$$

The boundary of $\mathbf{Y}$ is set such that the search space is not too big for a desktop computer to handle, and the typical system values previously given do not suggest an optimal solution beyond the space. For the benchmark, we create a network topology that consists of a macro BS and three femto BSs. There are only five UEs randomly distributed within the system. The small number of UEs is necessary because the number of searches needed increases very rapidly with UEs. Nevertheless, a small number of UEs are acceptable because our objective here is solely to assume the outcome of exhaustive search as a truly optimal solution for comparison. Table I compares the solutions obtained from TOTFA with those obtained from the exhaustive search. First of all, each exhaustive search needs about 24 h for a desktop computer (3.4-GHz Intel i7 Core CPU, 8-GB RAM, 64-bit Window 7) while each TOTFA execution takes less than a minute. Given the significant difference in the computational requirement, TOTFA's results are close to those of the exhaustive search, particularly in terms of the weighted sum of logarithm of throughput. The obvious difference in the achieved time fraction allocations is mainly due to the 0.02 step size that has been adopted to speed up the exhaustive search. Comparing subtable (b) and subtable (c), there is no difference in the cell association biases and time fraction allocations, but there is still a difference in the achieved weighted sum. This is because the changes in path loss exponents result in changes in data rates, which can directly affect the throughput. Based on Table I, we find that TOTFA can achieve optimality in proportional fairness by adjusting association bias and allocating time fractions.

With the optimality of TOTFA verified, we now want to establish its performance benefit. For performance comparison, we have adopted the weighted equal allocation scheme as the baseline for benchmark. In the baseline scheme, all cell association biases are set to zero, and UE i is allocated a transmission time fraction according to its weight w_i as

$$x_i = \frac{w_i}{\sum_{i \in \mathcal{U}} w_i}. \qquad (31)$$

In addition to the baseline scheme, we also compare TOTFA with another scheme that allocates time fractions to achieve proportional fairness [see (4)]. This proportional fairness scheme differs from the baseline scheme only in terms of time fraction allocation in achieving a different performance objective. Neither comparison scheme controls the association bias, whereas such a control is done in TOTFA. Although there is no need to adjust association bias in the proportional fairness scheme, finding its time fraction allocations for UEs is still a constrained optimization problem that we have solved using the MATLAB toolbox.

We first create a system with one macro BS located at the center and three femto BSs randomly located within 1-km coverage radius of the macro BS. The backhaul link capacity of the macro BS is 25 Mb/s, whereas the backhaul link capacities of femto BSs are varying parameters. For simplicity, we assume that the system has only ten randomly distributed UEs after noticing that a larger number of UEs do not affect the performance trend.

Fig. 4 shows the aggregated throughput of all UEs given different femto backhaul link capacities. In the figure, TOTFA can achieve a much higher aggregated throughput than the baseline scheme. The throughput improvement is higher when the

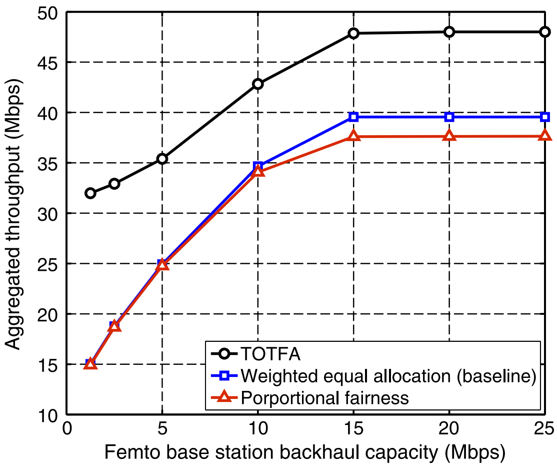


Fig. 4. Aggregated throughput of all UEs at different femto backhaul link capacities.

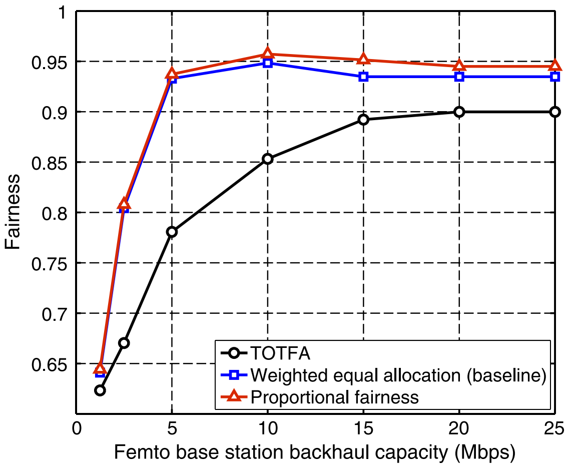


Fig. 5. Jain's fairness index at different femto backhaul link capacities.

femto backhaul link capacity is lower. Specifically, TOTFA's aggregated throughput is more than 100% higher than that of the baseline scheme when the femto backhaul link capacity is 1.25 Mb/s. When the femto backhaul link is 25 Mb/s, TOTFA's throughput is only 20% higher. Over different backhaul link capacities (1.25–25 Mb/s), the average throughput improvement is 45%. At the smaller backhaul link capacities, the larger throughput improvements are due to backhaul awareness and spatial reuse, as compared with only spatial reuse that contributes to the improvement at a larger backhaul link capacity. Recall that TOTFA is backhaul aware because it adjusts the association biases of different BSs according to their respective backhaul link capacities so that UEs will be intelligently distributed among the BSs to avoid bottlenecks. The effect of such backhaul awareness is diminishing when the backhaul link capacity increases and poses a smaller bottleneck. There is a bottleneck in the backhaul link of a BS when the aggregated throughput of all its associated UEs is larger than the link capacity. In this simulation configuration, such a bottleneck occurs at different values of backhaul capacities at different BSs. These different backhaul values are observed to be lower than 15 Mb/s.

Fig. 4 also shows that, when there is a bottleneck in the backhaul ($b_j < 15$ Mb/s), increasing femto backhaul link capacity can lead to a higher aggregated throughput. On the other hand, when there is no bottleneck in the backhaul ($b_j \geq 15$ Mb/s), further increasing femto backhaul link capacity does not produce noticeably higher aggregated throughput. Through careful examination of log files, we notice that the performance trends are due to the fact that not all frames are fully occupied when the limited femto backhaul link capacity constrains the amount of time fraction that can be allocated to all UEs. The portion of unused frame is reduced with a higher backhaul link capacity, and this leads to a higher throughput. Following the same understanding, when all the frames are fully occupied, a further increase in femto backhaul link capacity can hardly change the throughput.

TOTFA aims to achieve proportional fairness by maximizing the weighted sum of logarithm of throughput. While this is a conventional approach, the weighted sum itself is not a measure of fairness among different UEs. In the literature, Jain's fairness index F is a widely used measure for fairness [38], and it is defined as

$$F = \frac{\left(\frac{\sum_{i\in\mathcal{U}} x_i r_i}{w_i}\right)^2}{|\mathcal{U}| \sum_{i\in\mathcal{U}} \left(\frac{x_i r_i}{w_i}\right)^2}. \tag{32}$$

Fig. 5 compares the fairness index between TOTFA and the baseline scheme. According to the figure, TOTFA is less fair compared with the baseline scheme. This is understandable because the baseline scheme does a strict equal time fraction allocation among all the UEs. On the other hand, TOTFA maximizes the weighted sum of logarithm of throughput, and doing so allows TOTFA to exploit varying data rates among UEs for a higher aggregated throughput but at the expense of a lower equality. While TOTFA scores lower in fairness, its value is still high in the range of 0.9, and this slight unfairness is the price TOTFA pays for its higher throughput.

Given the significant improvement in throughput and a slight compromise in fairness, TOTFA confirms the importance of being backhaul-aware and the advantages of controlling cell association and allocating time fraction in a joint fashion. In addition, TOTFA is a suitable scheme for HetNets with femto BS connections through residential broadband links because TOTFA can adapt to variability in the backhaul links among different femto BSs.

In Figs. 4 and 5, the baseline scheme (weighted equal allocation) and the proportional fairness scheme show a similar performance trend. In fact, the difference in performance is only noticeable when there is no bottleneck in backhaul links. This is because, unlike TOTFA, the two schemes are not backhaul aware, and thus, they cannot adjust association bias to alleviate the effects of backhaul bottleneck. Given that time fractions are the only controllable variables and the need to keep total time fraction upper bounded by the backhaul capacity, there are only slight differences in the allocated time fractions among the two schemes. When the backhaul bottlenecks disappear at the higher link capacities, the proportional fairness scheme can trade part of its throughput for better fairness, as compared with the baseline scheme. Since there is no significant difference in

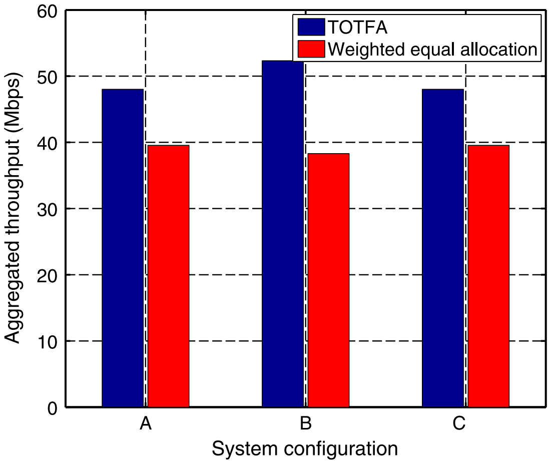


Fig. 6. Aggregated throughput of all UEs in three different configurations. (A) $b_0 = b_1 = b_2 = b_3 = 25$ Mb/s. (B) $b_0 = 10$ Mb/s, and $b_1 = b_2 = b_3 = 25$ Mb/s. (C) $b_0 = 100$ Mb/s, and $b_1 = b_2 = b_3 = 50$ Mb/s.

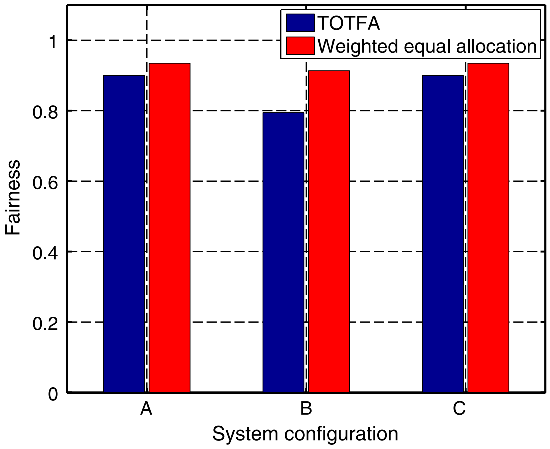


Fig. 7. Jain's fairness index in three different configurations. (A) $b_0 = b_1 = b_2 = b_3 = 25$ Mb/s. (B) $b_0 = 10$ Mb/s, and $b_1 = b_2 = b_3 = 25$ Mb/s. (C) $b_0 = 100$ Mb/s, and $b_1 = b_2 = b_3 = 50$ Mb/s.

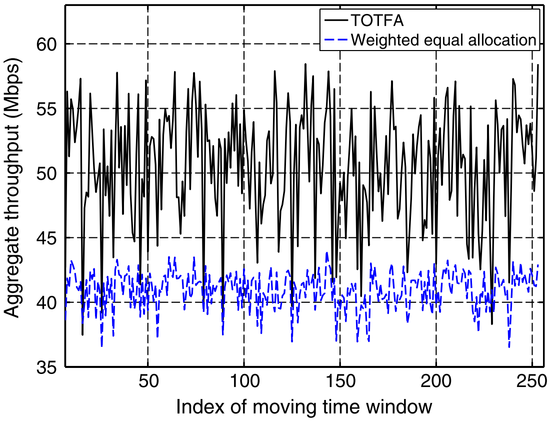


Fig. 8. Aggregated throughput when each femto backhaul link capacity is time varying while macro backhaul is fixed at 25 Mb/s.

the performance trend between the two benchmark schemes, we only compare the baseline scheme to TOTFA hereafter.

So far, we have fixed the macro backhaul link capacity while varying the femto backhaul link capacity. Fig. 6 compares the aggregated throughput with different macro backhaul capacities in three different configurations. In configuration A, $b_0 = b_1 = b_2 = b_3 = 25$ Mb/s. In configuration B, $b_0 = 10$ Mb/s, and $b_1 = b_2 = b_3 = 25$ Mb/s. In configuration C, $b_0 = 100$ Mb/s, and $b_1 = b_2 = b_3 = 50$ Mb/s. Comparing the aggregated throughput for configurations A and B, we notice that, when there is a bottleneck in macro backhaul, the system can achieve a higher throughput as long as there is no bottleneck in the femto backhaul. In other words, limitation in macro backhaul capacity has forced TOTFA to allocate a larger time fraction for femto BSs, and the system benefits from the higher data rate with femto BSs. This is confirmed with the corresponding drop in fairness in Fig. 7. At the first glance, the macro backhaul bottleneck should be dealt with through cell association bias by assigning a larger bias to femto BSs. However, we notice that this is not the case here because Algorithm 2 finds that a larger femto association bias may cause femto BSs overloading as determined by (14) because the data rates of their existing UEs are high. Instead, TOTFA finds that aggregated throughput can be improved by allocating a smaller time fraction to macro UEs. This is a piece of evidence that traffic offloading must be performed jointly with time fraction allocation for the best outcome when offloading alone cannot help in improving performance. When the macro backhaul capacity is limited, there is no point to allocate more transmission time to its associated UEs.

Comparing configurations A and C in Figs. 6 and 7 also shows that an increase in the backhaul link capacity may not improve throughput or fairness. A detailed examination of the simulation log files reveals that the invariance in the performance is due to the fact that there is no difference in the UE associations between the two configurations. In addition to backhaul links, UE association also depends on the user location, which will further affect the link data rate. TOTFA is capable of taking into account all these factors in deciding the association bias and time fraction allocation. Therefore, there is no need to blindly increase the femto backhaul capacity or to worry about the suitability of connecting femto BSs through residential broadband links. In TOTFA, traffic offloading to femto BSs is implicitly done in accordance to the available femto backhaul capacity so that a bottleneck can be avoided as much as possible. It is not useful to increase cell association bias of a femto BS if its backhaul link is already constrained. In addition, since TOTFA is backhaul aware, a larger time fraction will be automatically allocated to a BS that has more associated loads.

Fig. 8 shows the aggregated throughput when each femto backhaul link capacity is randomly and independently varying from time to time within the range [10, 25] Mb/s. As such, different femto backhauls may have different capacities at a given time while the macro backhaul link capacity is fixed at 25 Mb/s. In this result, there are three femto BSs and 50 UEs distributed within the coverage of a macro BS. The result indicates that TOTFA is capable of adapting to time-varying backhaul capacities and delivering a higher aggregated throughput compared with the baseline scheme despite the challenging environment.

Fig. 9 presents the actual times taken to execute TOTFA for different configurations. Each point on the graph is an average value for multiple executions on the desktop computer we have used earlier for the exhaustive search in Table I. The figure shows that TOTFA requires a longer computational time for

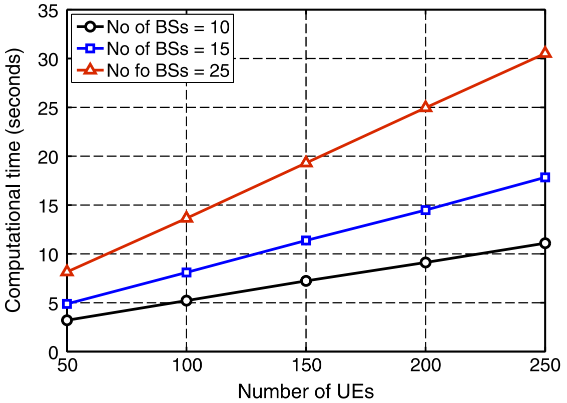


Fig. 9. Computational times required to perform the proposed TOTFA algorithms at different configurations.

more UEs and femto BSs. The computational time increases much faster with increasing number of BSs, as compared with an increasing number of UEs. Considering a macrocell that has 25 femto BSs and supports a total of 250 UEs, the time required to execute TOTFA once is only about 30 s. This time is far below our consideration of running TOTFA once every few minutes. As such, despite the sophistication, computational requirement is a not a limitation in implementing TOTFA in a HetNet.

V. Conclusion

This paper has proposed a novel scheme called TOTFA to perform traffic offloading and transmission time fraction allocation to achieve proportional fairness among UEs in the future 5G two-tier HetNets. This scheme exploits spatial reuse to allow concurrent transmissions of multiple femto BSs and considers backhaul link capacity in allocating resources. Compared with a baseline scheme, TOTFA can improve aggregated throughput for as much as 100% when there is a bottleneck in femto backhaul links. The improvement is lower at about 20% when the bottleneck does not exist, and the average attainable improvement is 45%. The superior aggregated throughput comes with the price of lower fairness. Nevertheless, the fairness is still acceptably high at about 0.9 in most cases. Performance results also reveal that it is essential to perform jointly transmission time fraction allocation and cell association bias control. In addition, it is critical to consider the femto backhaul link capacity such that a bottleneck can be avoided as much as possible.

References

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

[2] C. Singhal, S. De, and H. M. Gupta, "User heterogeneity and priority adaptive multimedia broadcast over wireless," in *Proc. IEEE ICC*, Sydney, NSW, Australia, Jun. 2014, pp. 1699–1704.

[3] M. Batty *et al.*, "Smart cities of the future," Centre Adv. Spatial Anal., Univ. College London, London, U.K., Working Papers Ser., Oct. 2012.

[4] "The 5G infrastructure public private partnership: The next generation of communication networks and services," Centre Adv. Spatial Anal., Univ. College London, London, U.K., 5G Vision, Feb. 2015. [Online]. Available: http://www.5g-ppp.eu

[5] A. Ghosh *et al.*, "Heterogeneous cellular networks: From theory to practice," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 54–64, Jun. 2012.

[6] P.-Y. Kong and A. Sluzek, "Average packet delay analysis for a mobile user in a two-tier heterogeneous cellular network," *IEEE Syst. J.*, to be published.

[7] P.-Y. Kong, "A Markov chain model for packet queueing delay analysis of a mobile user in HetNets," in *Proc. IEEE WCNC*, New Orleans, LA, USA, Mar. 2015, pp. 1990–1995.

[8] Y. Song, P.-Y. Kong, and Y. Han, "Power-optimized vertical handover scheme for heterogeneous wireless networks," *IEEE Commun. Lett.*, vol. 18, no. 2, pp. 277–280, Feb. 2014.

[9] P.-Y. Kong, "Power consumption and packet delay relationship for heterogeneous wireless networks," *IEEE Commun. Lett.*, vol. 17, no. 7, pp. 1376–1379, Jul. 2013.

[10] P.-Y. Kong and G. K. Karagiannidis, "Minimizing power consumption in HetNets with packet delay constraints," in *Proc. IEEE PIMRC*, Washington, DC, USA, Sep. 2015, pp. 1803–1807.

[11] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Mar. 2012.

[12] C. Coletti *et al.*, "Heterogeneous deployment to meet traffic demand in a realistic LTE urban scenario," in *Proc. IEEE VTC—Fall*, Quebec City, QC, Canada, Sep. 2012, pp. 1–5.

[13] R. Liao, B. Bellalta, J. Barcelo, V. Valls, and M. Oliver, "Performance analysis of IEEE 802.11ac wireless backhaul networks in saturated conditions," *EURASIP J. Wireless Commun. Netw.*, vol. 2013, no. 1, pp. 1–14, Sep. 2013.

[14] P. Okvist and A. Simonsson, "LTE HetNet trial—Range expansion including micro/pico indoor coverage survey," in *Proc. IEEE VTC—Fall*, Quebec City, QC, Canada, Sep. 2012, pp. 1–5.

[15] R. Madan *et al.*, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1479–1489, Dec. 2010.

[16] S. H. Wong and Z. Z. Lei, "Inter-cell interference management for heterogeneous networks," in *Heterogeneous Cellular Networks*, R. Q. Hu and Y. Qian, Eds. Oxford, U.K.: Wiley, Apr. 2013, pp. 93–117.

[17] Y. L. Lee, T. C. Chuah, J. Loo, and A. Vinel, "Recent advances in radio resource management for heterogeneous LTE/LTE-A networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2142–2180, Aug. 2014.

[18] D. Lopez-Perez *et al.*, "Enhanced inter-cell interference coordination challenges in heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 22–30, Jun. 2011.

[19] J. Oh and Y. Han, "Cell selection for range expansion with almost blank subframe in heterogeneous networks," in *Proc. IEEE PIMRC*, Sydney, NSW, Australia, Sep. 2012, pp. 653–657.

[20] T. D. Lagkas and P. Chatzimisios, "Performance and fairness analysis of a QoS supportive MAC protocol for wireless LANs," in *Proc. IEEE ICC*, Kyoto, Japan, Jun. 2011, pp. 1–6.

[21] A. Banchs, P. Serrano, and H. Oliver, "Proportional fair throughput allocation in multirate IEEE 802.11e wireless LANs," *Wireless Netw.*, vol. 13, no. 5, pp. 649–662, Oct. 2007.

[22] T. Bu, L. E. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. IEEE INFOCOM*, Barcelona, Spain, Apr. 2006, pp. 1–12.

[23] H. ElSawy and E. Hossain, "Two-tier HetNets with cognitive femtocells: Downlink performance modeling and analysis in a multichannel environment," *IEEE Trans. Mobile Comput.*, vol. 13, no. 3, pp. 649–663, Mar. 2014.

[24] S.-S. Sun, W. Liao, and W.-T. Chen, "Traffic offloading with rate-based cell range expansion offsets in heterogeneous networks," in *Proc. IEEE WCNC*, Istanbul, Turkey, Apr. 2014, pp. 2833–2838.

[25] K. Okino, T. Nakayama, C. Yamazaki, H. Sato, and Y. Kusano, "Pico cell range expansion with interference mitigation toward LTE-Advanced heterogeneous networks," in *Proc. IEEE ICC Workshop*, Kyoto, Japan, Jun. 2011.

[26] S. Corroy, L. Falconetti, and R. Mathar, "Dynamic cell association for downlink sum rate maximization in multi-cell heterogeneous networks," in *Proc. IEEE ICC*, Ottawa, ON, Canada, Jun. 2012, pp. 2457–2461.

[27] Y. Song, P.-Y. Kong, and Y. Han, "Minimizing energy consumption through traffic offloading in a HetNet with 2-class traffic," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1394–1397, Aug. 2015.

[28] M. Cierny, H. Wang, R. Wichman, Z. Ding, and C. Wijting, "On Number of Almost Blank Subframes in Heterogeneous Cellular Networks. [Online]. Available: arxiv.org/abs/1304.2269

[29] Y. Jin and L. Qiu, "Joint user association and interference coordination in heterogeneous cellular networks," *IEEE Commun. Lett.*, vol. 17, no. 12, pp. 2296–2299, Dec. 2013.

[30] I. Guvenc, "Capacity and fairness analysis of heterogeneous networks with range expansion and interference coordination," *IEEE Commun. Lett.*, vol. 15, no. 10, pp. 1084–1087, Oct. 2011.

[31] H. Boostanimehr and V. K. Bhargava, "Unified and distributed QoS-driven cell association algorithms in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1650–1662, Jan. 2015.

[32] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1100–1113, Jun. 2014.

[33] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets," *IEEE Trans. Netw.*, vol. 22, no. 1, pp. 137–150, Feb. 2014.

[34] P.-Y. Kong, K. C. Chua, and B. Bensaou, "MultiCode-DRR: A packet scheduling algorithm for delay guarantee in a MultiCode-CDMA network," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2694–2704, Nov. 2005.

[35] S. Chakravarty, A. Stavrou, and A. D. Keromytis, "LinkWidth: A method to measure link capacity and available bandwidth using single-end probes," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-002-08, 2008.

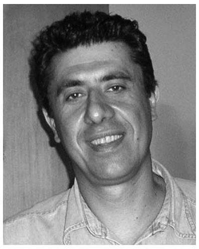
[36] M. Bazaraa, H. Sherali, and C. Shetty, *Nonlinear Programming: Theory and Algorithms*. New York, NY, USA: Wiley, 1993.

[37] D. P. Bertsekas, *Nonlinear Programming Programming*. Cambridge, U.K.: Athena Scientific, 1999.

[38] R. Jain, D. M. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer system," Digit. Equip. Corp. (DEC), Hudson, MA, USA, Tech. Rep. 301, Sep. 1984.

Peng-Yong Kong (M'03–SM'12) received the B.Eng. (first-class honors) degree in electrical and electronic engineering from Universiti Sains Malaysia, Gelugor, Malaysia, in 1995, and the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2002.

Prior to his Ph.D. studies, he was an Engineer with Intel Malaysia. He was previously an Adjunct Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, and concurrently a Research Scientist with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Khalifa University of Science, Technology and Research, Abu Dhabi, United Arab Emirates. His research interests are in wireless networking and wireless network protocols.



George K. Karagiannidis (M'96–SM'03–F'14) was born in Pythagorion, Samos Island, Greece. He received the University Diploma (five years) and Ph.D. degrees in electrical and computer engineering from the University of Patras, Patras, Greece, in 1987 and 1999, respectively.

From 2000 to 2004, he was a Senior Researcher with the Institute for Space Applications and Remote Sensing, National Observatory of Athens, Greece. In June 2004, he joined the faculty of Aristotle University of Thessaloniki, Thessaloniki, Greece, where he is currently a Professor with the Department of Electrical and Computer Engineering and the Director of Digital Telecommunications Systems and Networks Laboratory. He is also with the Department of Electrical and Computer Engineering, Khalifa University of Science, Technology and Research, Abu Dhabi, United Arab Emirates. He is also an Honorary Professor with Southwest Jiaotong University, Chengdu, China. His research interests are in the broad area of digital communication systems with emphasis on wireless communications, optical wireless communications, wireless power transfer and applications, molecular communications, communications and robotics, and wireless security. He is the author or coauthor of more than 400 technical papers published in scientific journals and presented at international conferences. He is also the author of the Greek edition of a book *Telecommunications Systems* and a coauthor of the book entitled *Advanced Optical Wireless Communications Systems* (Cambridge, 2012).

Prof. Karagiannidis has been involved as a General Chair, a Technical Program Chair, and a member of technical program committees in several IEEE and non-IEEE conferences. He was an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, a Senior Editor of the IEEE COMMUNICATIONS LETTERS, an Editor of the European Association for Signal Processing (EURASIP) *Journal of Wireless Communications and Networks*, and a Guest Editor several times for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. From 2012 to 2015, he was the Editor-in-Chief of the IEEE COMMUNICATIONS LETTERS. He was selected as a Highly Cited Researcher by Thomson Reuters in 2015.