Mobile Netw Appl (2016) 21:453–466
DOI 10.1007/s11036-015-0650-z

Throughput Optimization via Association Control in Wireless LANs

Heyi Tang[1] · Liu Yang[2] · Jiang Dong[3] · Zhonghong Ou[3] · Yong Cui[1] · Jianping Wu[1]

Published online: 22 September 2015

Abstract With the rapid development of the mobile computing, accessing the Internet everywhere is important for mobile device users. Wireless LAN is a stable and reliable technique to provide network access for mobile devices. The Wireless LAN Access Points(APs) have been densely deployed so that a user can access the Internet almost everywhere. However, this fact brings some new challenges. Since the regular AP association strategy is signal-based when a user receive the signals of multiple APs. The APs with strong signal will be too overloaded while the bandwidth resource in other APs is wasted. The throughput of the whole WLAN is not optimized. Moreover, the diverse bandwidth demands among users further exacerbate the situation. In this paper, aiming at optimizing the throughput over the whole WLAN, a joint AP association and bandwidth allocation problem is formulated. The different users' bandwidth demands are added as new constraints. We comprehensively analyze the solution space and prove the problem NP-hard. Our trace-driven evaluations show that the throughput is improved about 23.1 % compared to the conventional schemes.

✉ Yong Cui
cuiyong@tsinghua.edu.cn

1 Tsinghua University, Beijing, China

2 Beijing University of Posts and Telecommunication, Beijing, China

3 Aalto University, Espoo, Finland

1 Introduction

Recently, the widely deployed Wireless Local Area Networks (WLANs) based on 802.11 protocol are constantly developing to meet the growing demand on Internet access of users. By default, the user in a WLAN associates with the AP that has the best Received Signal Strength Indicator (RSSI). Since the users are usually not uniformly distributed in the area, the RSSI-based association strategy often leads to the load imbalance among APs and decreases the aggregated throughput.

Many researchers focus on the design of optimized association schemes with their own specific objectives. Rather than considering the RSSI only, they introduce more parameters at the user side and propose local-view association strategies [7, 8, 10, 15, 17, 26, 32]. The others consider the parameters of the whole network and give global-view association strategies [1, 9, 12, 13, 16, 31].

After the optimal association strategies are given, another important issue is the fair bandwidth allocation of AP. Many bandwidth allocation strategies are proposed to achieve the proportional fairness, where the users are allocated bandwidth according to their priority [19, 23–25, 28]. In [22, 29, 30], the users are allocated the same bandwidth, which achieves the max-min fairness among the users.

Existing AP association and bandwidth allocation strategies are proposed based on the assumption that any user has a huge bandwidth demand. The user demand cannot be satisfied, no matter how to allocate the bandwidth of AP to its associated users. Then the strategies allocate bandwidth equal or proportional to the transmission rate of users and assume the users can run out their allocated bandwidths. However, it is common that the bandwidth demand of each user differs vastly due to their preference on different network applications. The allocated bandwidth may be wasted if users just enjoy surfing the web-sites; while the bandwidth is not enough, if users watch videos. Hence, the allocation strategies without consideration of the user bandwidth demand are not optimal.

To address this challenge, we introduce the bandwidth demand of users as a new constraint to formulate the Joint AP Association and bandwidth Allocation (JAA) problem. The introduction of bandwidth demand fundamentally changes the nature of both the AP association and bandwidth allocation problems, and essentially calls for a separate study. Inspired by existing works in which the number of associated users is used to measure the AP load, we consider the joint effect of the number of users and each user's demand. In the single AP case, the conventional schemes that allocate bandwidth equal or proportional to the transmission rate may result in bandwidth waste. Thus we give a comprehensive analysis on the solution space of optimal bandwidth allocation while satisfying the proportional fairness among users in this new situation. As to the multi-AP case, we formulate the JAA problem as a mixed-integer nonlinear programming problem. By reducing the well-known partition problem [14], we prove that the JAA problem is NP-hard.

To solve the problem, we give an approximate Maximum Aggregated Bandwidth Utility (MABU) algorithm to achieve load balance by estimating on the AP utilization. With a detailed analysis on the properties of the optimal bandwidth allocation, we design the fair bandwidth allocation strategies for APs. We prove that the approximation factor of MABU is 1/2 under our imported settings. Our performance evaluations based on the real data traces demonstrate that MABU increases the user's throughput by up to 23.1 % and achieves better user fairness compared with conventional schemes.

The rest of the paper is organized as follows. We summarize the related work in Section 2. In Section 3, we formulate the JAA problem. The MABU algorithm is proposed in Section 4 to solve the problem. Section 5 gives the performance evaluations. We conclude the paper in Section 6.

2 Related work

Improving the throughput via optimizing AP association and balancing the load has been extensively studied. Rather than the default RSSI-only strategy, some studies enhance the association strategy by considering more parameters based on the user's view. In [17], the transmission based on the Distributed Coordination Function (DCF) mechanism is analyzed to estimate the load on AP and then a distributed association selection is made. Wendong et al. [15] propose an optimal AP selection scheme while considering both throughput and energy consumption. In [10], the differentiated access service selection based on the users' running applications is presented. Mingming et al. [26] model the AP selection problem as a matching problem and study the localized solutions. In [8], the access point (AP) selection problem is investigated using game theory with only local information.

However, these strategies proposed with the user's local view are not globally optimal. Thanks to the appearance of network controller [3], the centralized scheduling with parameters of the whole network becomes possible. Bejerano et al. [4] schedule the AP association while considering the power of APs, inspired by the experience of cell breathing technic in cellular network. In [16], the impact of legacy 802.11a/b/g clients is considered, and the optimal AP association is given to maximize the minimum MAC efficiency. In [9], in order to balance the load of the whole network, the One-AP-Multiple-Interface (OAMI) technic is proposed to achieve the seamless handoff among different interfaces in one AP. Guangtao et al. [31] propose a novel AP association scheme, in which the users with intense social relationships are assigned to different APs. Thus jointly departure of those users would have minor impact on the load balance of APs. In [13], the hidden terminal effect is considered as the main reason of AP performance degradation.

Besides improving the throughput by solving the problem of load imbalance, the fair bandwidth allocation among the users should be also considered. Some studies consider the bandwidth allocation problem while giving the optimal association strategies. They investigate how to improve the throughput while achieving user fairness. In [23–25], to balance the throughput increase and fairness guarantee, the time-based fair scheduling has been proposed. The user is allocated transmission time according to its priority. Ouldooz et al. [19] investigate a collaborative association problem, i.e., a user can obtain transmission time from multiple APs simultaneously. The priority of user is different when associated with different APs. Xie et al [28] design a dynamic priority-based online algorithm to allocate bandwidth for the users.

Another group of work also focuses on the joint AP association and bandwidth allocation problem ,but proposes different bandwidth allocation criterions. In these works, the allocated bandwidth to each user is equal. In [29, 30], the AP selection problem is studied in a decentralized manner, with the objective of maximizing the minimum bandwidth. Stratos et al. [22] propose innovative metrics on which the association should be based, such as contention, transmission rate and backlogged traffic. In these research, the allocated transmission time is the reciprocal of its transmission rate, to ensure the equality of allocated bandwidth.

The above research contributes a lot to solve the problem of load imbalance and fair bandwidth allocation. However, the bandwidth demand of users, a key parameter that influences the load balance is usually ignored in the existing research. It is mentioned in [5], but not analyzed in detail. The benefit of load balance is degraded without consideration of user bandwidth demand. After the different bandwidth demands are brought in, existing bandwidth allocation schemes cannot work well. Simply allocating the bandwidth equally or according to the priority may cause the significant bandwidth waste.

3 Problem formulation

3.1 Network model

Our network topology models an IEEE 802.11-based multirate WLAN that consists of multiple APs operating at orthogonal channels. Each AP has the limited coverage area and serves users in its area. Overlapping coverage areas of adjacent APs exist. The investigation focuses on the down link data sent to the users, as it produces the dominated traffic for the real applications [2, 20]. We assume that each AP transmits data with the same power, and that each user is covered by at least one AP. We further assume that a period of time T, in which the network is stable, with no user joins or leaves, is to be considered. This means that the allocable transmission time of an AP equals to T. Each AP assigns fractional transmission time to its associated users, and a user is allowed to associate to only one AP within T.

We use the coefficient x_{ij} to indicate the association relationship between user i and AP j. The coefficient is defined as follows:

$$x_{ij} = \begin{cases} 1, & \text{if user } i \text{ is associated to AP } j, \\ 0, & \text{otherwise,} \end{cases}$$

where $i = 1, 2, \cdots, M$, $j = 1, 2, \cdots, N$. The allocated transmission time of user i from AP j is denoted as t_{ij}. A list of symbols used in this paper are summarized in Table 1.

Definition 1 The transmission rate of user i associated to AP j is the size of data that the user can download from AP j during per unit of time when the AP is fully occupied by user i, denoted by r_{ij}.

r_{ij} is determined by the experienced Signal to Interference plus Noise Ratio (SINR) of user i. The SINR of user i associated to AP j is given by [11]:

$$\gamma_{ij} = \frac{P_j d_{ij}^{-\alpha}}{N_0}. \tag{1}$$

Table 1 Notations

M	The number of users
N	The number of APs
T	The allocable transmission time of AP
γ_{ij}	The SINR of user i associated to AP j
N_0	The power of Gaussian white noise
P_j	The transmission power of AP j
d_{ij}	The physical distance between user i and AP j
S_i	The size of data required by user i
B_i	The bandwidth demand of user i
x_{ij}	The association coefficient of user i and AP j
r_{ij}	The transmission rate of user i and AP j
T_{ij}	The transmission time demand of user i on AP j
t_{ij}	The allocated transmission time of user i by AP j
b_{ij}	The allocated bandwidth of user i by AP j
$\widetilde{b}_i$	The total allocated bandwidth of user i

P_j is the transmission power of AP j. d_{ij} is the physical distance between AP j and user i. α is the shadowing factor ranging from 2.0 to 5.0 [27]. N_0 is the power of Gaussian white noise in the environment. Since the channels taken by different APs are orthogonal, the user is not interfered by other APs and the noise only includes Gaussian white noise.

The relationship between r_{ij} and γ_{ij} in an 802.11g based network is shown in Table 2 [25]. After the association is decided, the transmission rate of user i is given by $\widetilde{r}_i = \sum_{j=1}^{N} x_{ij} r_{ij}$.

Definition 2 The bandwidth demand of user i is the size of data that the user requires during per unit of time, denoted by B_i. It is formulated by

$$B_i = \frac{S_i}{T}. \tag{2}$$

S_i is the size of data required by user i. It elicits the concept of transmission time demand.

Definition 3 The transmission time demand of user i associated to AP j is the length of time that the user needs to download the required data from AP j, which can be formulated as

$$T_{ij} = \frac{S_i}{r_{ij}} = \frac{B_i}{r_{ij}} T. \tag{3}$$

Table 2 Relationship between γ_{ij} and r_{ij}

$\gamma_{ij}(dB)$	6-7.8	7.8-9	9-10.8	10.8-17	17-18.8	18.8-24	24-24.6	24.6-
$r_{ij}(Mbps)$	6	9	12	18	24	36	48	54

The bandwidth demand of user i is fixed and has no relationship with its associated AP. However, the transmission time demand of user i is variable, since the transmission rate is related to the associated AP. After the association is decided, the transmission time demand of user i is given by $\widetilde{T}_i = \sum_{j=1}^{N} x_{ij} T_{ij}$.

Definition 4 The allocated bandwidth of user i from AP j is the average number of bits that the user downloads from AP j during per unit of time, when the AP is partly occupied by user i. It is given by

$$b_{ij} = \frac{r_{ij} t_{ij}}{T}. \tag{4}$$

After the association is decided, the allocated bandwidth of user i is given by $\widetilde{b}_i = \sum_{j=1}^{N} x_{ij} b_{ij}$.

The JAA problem is to find the optimal association decisions $\{x_{ij}\}$ and allocated user bandwidth $\{b_{ij}\}$, which maximize the aggregated throughput of all the users while ensuring the user fairness.

3.2 Bandwidth allocation of single AP

In this part, we analyze the bandwidth allocation problem with the demand as constraints, and give the properties of optimal solution. In a single AP, the allocated bandwidth is decided by the transmission rate and the allocated transmission time. Since the transmission rate is fixed after the association is decided, the problem of bandwidth allocation transforms to the problem of time allocation.

Here we denote the allocated transmission time of user i on the AP as $\widetilde{t}_i$, the transmission time demand of user i as $\widetilde{T}_i$, the transmission rate of user i as $\widetilde{r}_i$. Thus the allocated bandwidth of user i in this situation is $\widetilde{b}_i = \widetilde{r}_i \cdot \widetilde{t}_i$.

Note that in the multi-rate network, only maximizing the aggregated throughput may result in the starvation of the low-rate users. Thus taking the user fairness into account is necessary. Our goal is to allocate the transmission time of AP to the users in a proportional-fairness manner. Maximizing the aggregated bandwidth utility is used to achieve the fairness [21], i.e., max $\sum_{i=1}^{m} \log(\widetilde{b}_i)$. m is the number of users associated to the AP. The logarithm of the allocated bandwidth is the bandwidth utility of user i.

Based on the proportional fairness, when there is no limit to the allocated bandwidth, T is equally allocated to the associated users. However, with the consideration of bandwidth demand, the allocated bandwidth has an upper limit. We illustrate the combined effect of bandwidth demand and proportional fairness on the time scheduling as follows.

When the aggregated transmission time demand of the users associated to the AP is no more than T, each user's transmission time demand is satisfied. Otherwise, T is allocated to each user as equally as possible, while satisfying the constraints $\sum_{i=1}^{m} \widetilde{t}_i = T$ and $0 \le \widetilde{t}_i \le \widetilde{T}_i$, $1 \le i \le m$. The time allocation is as equal as possible means that the variance of allocated transmission time is minimized. We prove the conclusion with Theorem 1.

Based on the proportional fairness, the optimization objective is

$$\max \ \sum_{i=1}^{m} \log(\widetilde{b}_i) = \sum_{i=1}^{m} \log \frac{\widetilde{r}_i \widetilde{t}_i}{T}$$

Since $\widetilde{r}_i$ and T are constant, the objective can be transformed into max $\sum_{i=1}^{m} \log \widetilde{t}_i$.

We denote the average allocated transmission time among users as $\tilde{\tilde{t}}$. The conclusion that optimizing the above objective is equal to minimizing the variance of allocated transmission time can be expressed as the following theorem.

Theorem 1 *Suppose that $\{\widetilde{t}_i^*\}$ is the optimal solution of Eq. 5, then $\{\widetilde{t}_i^*\}$ is the optimal solution of Eq. 6.*

$$\begin{array}{ll} \max & \sum_{i=1}^{m} \log \widetilde{t}_i \\ \textit{subject to:} & \sum_{i=1}^{m} \widetilde{t}_i = T \\ & 0 \le \widetilde{t}_i \le \widetilde{T}_i, 1 \le i \le m. \end{array} \tag{5}$$

$$\begin{array}{ll} \min & \sum_{i=1}^{m} (\widetilde{t}_i - \tilde{\tilde{t}})^2 / m \\ \textit{subject to:} & \sum_{i=1}^{m} \widetilde{t}_i = T \\ & 0 \le \widetilde{t}_i \le \widetilde{T}_i, 1 \le i \le m. \end{array} \tag{6}$$

Proof Since $\widetilde{t}_i^* \le \widetilde{T}_i$, the indexes of $\{\widetilde{t}_i^*\}$ can be divided into two subsets: set $A : \{i | \widetilde{t}_i^* = \widetilde{T}_i\}$, set $B : \{i | \widetilde{t}_i^* < \widetilde{T}_i\}$. To prove the theorem, we first give the two following lemmas:

Lemma 1 *$\{\widetilde{t}_i^*\}$ has the following two properties:*

(1) *The value of each $\widetilde{t}_i^*$ with index in set B equals to each other, i.e., $\exists t_B, \widetilde{t}_i^* = t_B, \forall i \in B$.*
(2) *Based on (1), the value of each $\widetilde{t}_i^*$ is no greater than t_B, i.e., $\widetilde{t}_i^* \le t_B, 1 \le i \le m$.*

Proof We prove Lemma 1 by solving Eq. 5 with the KKT method. According to the KKT conditions [6], let

$$L(t, \lambda, \mu, \omega) = -\sum_{i=1}^{m} \log \widetilde{t}_i + \lambda \left(\sum_{i=1}^{m} \widetilde{t}_i - T \right) + \sum_{i=1}^{m} \mu_i (\widetilde{t}_i - \widetilde{T}_i) - \sum_{i=1}^{m} \omega_i \widetilde{t}_i.$$

The optimal solution $\{\widetilde{t_i^*}\}$, λ^*, $\{\mu_i^*\}$, $\{\omega_i^*\}$ must satisfy the following KKT conditions:

$$-\frac{1}{\widetilde{t_i^*}} + \mu_i^* - \omega_i^* + \lambda^* = 0, 1 \le i \le m \quad (7)$$

$$\mu_i^* \left(\widetilde{t_i^*} - \widetilde{T_i}\right) = 0, 1 \le i \le m \quad (8)$$

$$-\omega_i^* \widetilde{t_i^*} = 0, 1 \le i \le m \quad (9)$$

$$\lambda^* (\sum_{i=1}^{m} \widetilde{t_i^*} - T) = 0 \quad (10)$$

$$\mu_i^* \ge 0, \omega_i^* \ge 0, 1 \le i \le m. \quad (11)$$

Since $\left\{\widetilde{t_i^*}\right\}$ is the optimal solution, none of the $\widetilde{t_i^*}$s is zero. Thus $\omega_i^* = 0$ according to Eq. 9. We can then rewrite Eq. 7 as

$$-\frac{1}{\widetilde{t_i^*}} + \mu_i^* + \lambda^* = 0, 1 \le i \le m. \quad (12)$$

On the other hand, with i in subset B, $\widetilde{t_i^*} < \widetilde{T_i}$, thus $\widetilde{t_i^*} - \widetilde{T_i} < 0$ and $\mu_i^* = 0$ according to Eq. 8. Thus Eq. 12 can be rewritten as $-\frac{1}{\widetilde{t_i^*}} + \lambda^* = 0, i \in B$. Thus $\widetilde{t_i^*} = \frac{1}{\lambda^*}, \forall i \in B$. Since λ^* is a fixed value, property (1) is proved and t_B is fixed as $t_B = \frac{1}{\lambda^*}$, i.e., $\lambda^* = \frac{1}{t_B}$.

Using Eq. 12, we have $\mu_i^* = \frac{1}{\widetilde{t_i^*}} - \frac{1}{t_B}, 1 \le i \le m$. Since $\mu_i^* \ge 0$ in Eq. 11, we have $\widetilde{t_i^*} \le t_B, 1 \le i \le m$. Then property (2) is proved. This completes the proof. □

Lemma 2 *If some solution of Eq. 6, $\{\widetilde{t_i}\}$, satisfies the properties presented in Lemma 1, $\{\widetilde{t_i}\}$ is the optimal solution of Eq. 6.*

Proof We assume that another solution is $\{\widetilde{t_i'}\} = \{\widetilde{t_i} + \delta_i\}, 1 \le i \le m$. Since $\{\widetilde{t_i'}\}$ is different from $\{\widetilde{t_i}\}$, at least one δ_i is unequal to 0. Thus we have $\sum_{i=1}^{m} \delta_i^2 > 0$

We divide the indexes of δ_i to two subsets: $I^+ : \{i|\delta_i > 0\}$, $I^- : \{i|\delta_i \le 0\}$. We have

$$\sum_{i \in I^+} \delta_i + \sum_{i \in I^-} \delta_i = \sum_{i=1}^{m} \delta_i = 0. \quad (13)$$

Since t_i with i in set A reaches the upper bound, i in set I^+ must come from set B, i.e., $I^+ \subset B$. Due to property (1) in Lemma 1, $\widetilde{t_i} = t_B, \forall i \in I^+$. We denote $d_i = \widetilde{t_i} - T/m$, $d_B = t_B - T/m$. Thus $d_i = d_B, \forall i \in I^+$.

We denote the variances as V and V', the difference of the variances is

$$V' - V = \left[\sum_{i=1}^{m} (d_i + \delta_i)^2 - \sum_{i=1}^{m} d_i^2\right]/m = \left(\sum_{i=1}^{m} 2d_i\delta_i + \delta_i^2\right)/m.$$

Due to $\sum_{i=1}^{m} \delta_i^2 > 0$ and $d_i = d_B, \forall i \in I^+$, we have

$$V' - V > \left(\sum_{i=1}^{m} 2d_i\delta_i\right)/m = \left(\sum_{i \in I^+} 2d_i\delta_i + \sum_{i \in I^-} 2d_i\delta_i\right)/m$$
$$= (2d_B \sum_{i \in I^+} \delta_i + \sum_{i \in I^-} 2d_i\delta_i)/m.$$

According to Eq. 13, we have

$$\left(2d_B \sum_{i \in I^+} \delta_i + \sum_{i \in I^-} 2d_i\delta_i\right)/m = \left(\sum_{i \in I^-} -2d_B\delta_i + 2d_i\delta_i\right)/m$$
$$= \left[\sum_{i \in I^-} 2(d_i - d_B)\delta_i\right]/m.$$

Since $t_i \le t_B$ according to property (2) in Lemma 1, we have $d_i \le d_B$. Due to $\delta_i \le 0, i \in I^-$, we have $[\sum_{i \in I^-} 2(d_i - d_B)\delta_i]/m \ge 0$.

In summary, we have $V' - V > 0$. It illustrates that any other solution has a larger variance, i.e., $\{\widetilde{t_i}\}$ reaches the minimum variance. This completes the proof. □

According to Lemma 1, $\left\{\widetilde{t_i^*}\right\}$ has the two properties. According to Lemma 2, the solution with the two properties is the optimal solution of Eq. 6. Thus $\left\{\widetilde{t_i^*}\right\}$ is the optimal solution of Eq. 6. This completes the proof. □

We give an example to illustrate Theorem 1.

Example 1 The value of T is 100 ms and there are 3 users associated to the same AP with $\widetilde{T_1} = 10\ ms, \widetilde{T_2} = 70\ ms, \widetilde{T_3} = 120\ ms$. Solving Eq. 5, we have $\widetilde{t_1^*} = 10\ ms, \widetilde{t_2^*} = 45\ ms; \widetilde{t_3^*} = 45\ ms$. Since $\widetilde{t_1^*} = \widetilde{T_1}; \widetilde{t_2^*} < \widetilde{T_2}, \widetilde{t_3^*} < \widetilde{T_3}$, We have $A = \{i|i = 1\}$, $B = \{i|i = 2, 3\}$. $\left\{\widetilde{t_i^*}\right\}$ satisfies the two properties in Lemma 1: (1) $t_B = 45\ ms$, $\widetilde{t_2^*} = \widetilde{t_3^*} = t_B$. (2) $\widetilde{t_i^*} \le t_B, 1 \le i \le 3$. The variance of $\widetilde{t_i^*}$s is $272.74ms^2$. It is minimized compared with other solutions.

In summary, we can allocate the transmission time of AP to its associated users according to the optimal solution of Eq. 5. In order to calculate the optimal solution, we have the following theorem.

Theorem 2 *Suppose $\left\{\widetilde{t_i'}\right\}$ is one solution of Eq. 5. If it satisfies the two properties in Lemma 1, $\left\{\widetilde{t_i'}\right\}$ must be $\left\{\widetilde{t_i^*}\right\}$, i.e., the optimal solution of Eq. 5.*

Proof We prove the theorem by contradiction. We assume $\left\{\widetilde{t_i'}\right\}$ is different from $\left\{\widetilde{t_i^*}\right\}$, the variance of them are respectively V' and V^*. According to Lemma 1, $\left\{\widetilde{t_i^*}\right\}$ satisfies the properties. Since the properties are satisfied, $V^* < V'$ on the basis of Lemma 2's proof. Besides, $\left\{\widetilde{t_i'}\right\}$ satisfies the properties, thus $V' < V^*$. The above two conclusions contradict each other.

Therefore, if some solution satisfies the two properties of the optimal solution, it is equivalent to the optimal solution $\{\widetilde{t}_i^*\}$. This completes the proof. □

According to Theorem 2, the optimal solution can be calculated based on the properties of $\{\widetilde{t}_i^*\}$ in Lemma 1. The algorithm to construct $\{\widetilde{t}_i^*\}$ is proposed as Algorithm 1 in Section 4.

3.3 JAA problem

In this part, we focus on the JAA problem when multiple APs are considered. Taking the objective of maximizing the aggregated throughput based on the fairness criterion, we need to decide: (1) Which AP j $(1 \le j \le N)$ each user i $(1 \le i \le M)$ should be associated to, denoted by x_{ij}; (2) How much transmission time each user i should be allocated from AP j, denoted by t_{ij}.

Maximizing the aggregated bandwidth utility of the network-wide users achieves the proportional fairness of the network-wide users [23], as presented as follows:

$$\max \sum_{i=1}^{M} U_i = \sum_{i=1}^{M} \log \sum_{j=1}^{N} \frac{x_{ij} r_{ij} t_{ij}}{T}.$$

The proportional fairness encourages the user to be associated with AP that provides high transmission rate, but forbids that too many users gather on one AP. Associating the user to the AP with higher transmission rate increases the aggregated bandwidth utility. However, if too many users gather on one AP, each user may get a too small t_{ij}. The logarithm value of $\sum_{j=1}^{N} x_{ij} r_{ij} t_{ij}$ decreases dramatically when t_{ij} is small, which violates the object of maximizing the aggregated bandwidth utility. After the association is decided, the time scheduling of each AP is the same with that in Section 3.2.

We present the JAA problem's formulation as follows.

$$\max \quad \sum_{i=1}^{M} U_i = \sum_{i=1}^{M} \log \sum_{j=1}^{N} \frac{x_{ij} r_{ij} t_{ij}}{T} \tag{14}$$

subject to:

$$x_{ij} \in \{0, 1\}, 1 \le i \le M, 1 \le j \le N \tag{15}$$

$$\sum_{j=1}^{N} x_{ij} = 1, 1 \le i \le M \tag{16}$$

$$0 \le t_{ij} \le T_{ij}, 1 \le i \le M, 1 \le j \le N \tag{17}$$

$$0 \le \sum_{i=1}^{M} x_{ij} t_{ij} \le T, 1 \le j \le N. \tag{18}$$

Equation 15 means that the association coefficient is a binary variable. Equation 16 limits that the user can be associated to only one AP. Equation 17 means that the user's allocated transmission time must be no more than its transmission time demand. Equation 18 means that the sum of transmission time allocated from the AP must be no more than its allocable transmission time.

Theorem 3 *Computing an optimal solution for JAA in Eq. 14 is NP-hard.*

Proof We first introduce the NP-hard partition problem [14]. Based on the partition problem, we prove another Optimized Partition Problem (OPP) that we construct is NP-hard. At last, we reduce the OPP to JAA to prove JAA is NP-hard.

(1) Partition Problem: Suppose a given multiset A which contains M integers $A_1, A_2, ...A_M$. The partition problem is to determine whether A is able to be partitioned into two subsets, so that the sum of the numbers in the two subsets are equal. Let $C_j (1 \le j \le 2)$ be the sum of the numbers in each subset. The coefficient $y_{ij} (1 \le i \le M, 1 \le j \le 2)$ is used to indicate which subset the integer should be partitioned into. $y_{ij} = 1$, if A_i is in subset j; else, $y_{ij} = 0$. Thus $C_j = \sum_{i=1}^{M} y_{ij} A_i$.

Since $C_1 + C_2 = \sum_{i=1}^{M} A_i$ and $C_1 = C_2$, the problem is equivalent to determine the possibility of $C_1 = C_2 = (\sum_{i=1}^{M} A_i)/2$. We use S' to denote the value of $(\sum_{i=1}^{M} A_i)/2$.

(2) (OPP): In this part, we construct another problem OPP. If OPP can be solved in polynomial time, the partition problem can be solved in polynomial time. As the partition problem is NP-Hard, OPP is in fact an NP-Hard problem.

OPP is formulated as follows.

$$\max \quad \sum_{i=1}^{M} \log \sum_{j=1}^{2} y_{ij} a_i \tag{19}$$

subject to:

$$y_{ij} \in \{0, 1\}, 1 \le i \le M, 1 \le j \le 2$$

$$\sum_{j=1}^{2} y_{ij} = 1, 1 \le i \le M$$

$$0 \le a_i \le A_i, 1 \le i \le M$$

$$0 \le \sum_{i=1}^{M} y_{ij} a_i \le S', 1 \le j \le 2$$

In OPP, a_i is the value partitioned into one subset, which is limited by A_i. Each a_i can be partitioned into only one subset. The sum of the numbers in each subset is limited by S'. The variable to be solved is $\{y_{ij}\}$ and $\{a_i\}$. We then illustrate that the solution of OPP entails the solution of the partition problem.

We find that the largest possible value of the objective is $\sum_{i=1}^{M} \log A_i$, which is denoted by L'. Suppose the optimal solution is $\left\{y_{ij}^*\right\}$ and $\left\{a_i^*\right\}$. If the optimal objective value reaches L', we must have $a_i^* = A_i, 1 \le i \le M$. Thus $C_1^* + C_2^* = \sum_{i=1}^{M} a_i^* = \sum_{i=1}^{M} A_i = 2S'$, i.e., $C_1^* = C_2^* = S'$. At this point, the partition problem is determined to be possible. Otherwise, if the optimal value is less than L', there must be some i that satisfies $a_i^* < A_i$. The partition problem is determined to be impossible.

(3) The reduction of OPP to JAA: Now we can construct a special case of JAA to be mapped to OPP.

We assume that $N = 2$ and the transmission rate between different APs and different users is the same as $r = 1, 1 \le i \le M, 1 \le j \le N$. Then the user has the same transmission time demand when associated to different APs, given by $T_i = S_i$ according to Eq. 3. Correspondingly, We transform t_{ij} as t_i. The problem is to maximize $\sum_{i=1}^{M} \log \sum_{j=1}^{2} x_{ij} r t_i / T$. Since r and T are constants, the problem is equivalent to maximize $\sum_{i=1}^{M} \log \sum_{j=1}^{2} x_{ij} t_i$. Comparing Eq. 14 with Eq. 19, we find that x_{ij} is y_{ij}, t_i is a_i, T_i is A_i and T is S'. At this point, OPP is exactly reduced to JAA. Since OPP is NP-Hard, JAA is proved NP-Hard. This completes the proof.

□

4 MABU algorithm

In this section, we first propose the Fairness-based Bandwidth Allocation (FBA) algorithm to solve the bandwidth allocation problem in Section 3.2. Then we propose the 1/2-approximation MABU algorithm to solve the JAA problem in Section 3.3.

4.1 Bandwidth allocation

We propose FBA to allocate the bandwidth of single AP to its associated m users. As we illustrate in Section 3, we can construct the optimal time allocation $\left\{\widetilde{t}_i^*\right\}$ according to the two properties of the optimal solution in Lemma 1.

Recall the definition and the properties of set A, B.

$$A = \left\{i | \widetilde{t}_i^* = \widetilde{T}_i\right\}, B = \left\{i | \widetilde{t}_i^* < \widetilde{T}_i\right\}$$

According to the properties,

$$\exists t_B, \forall i \in A, \widetilde{T}_i = \widetilde{t}_i^* \le t_B; \forall j \in B, t_B = \widetilde{t}_j^* < \widetilde{T}_j.$$

Thus $\widetilde{T}_i \le t_B < \widetilde{T}_j (i \in A, j \in B)$, i.e., t_B divides $\{\widetilde{T}_i\}$ into the smaller part and the larger part.

Suppose $\left\{\widetilde{T}_i\right\}$ is in ascending order. The following property can be derived from the above illustrations.

$$\exists k, \widetilde{T}_k \le t_B < \widetilde{T}_{k+1}, A = \{i | 1 \le i \le k\}, B = \{i | k+1 \le i \le m\}.$$

Since $\sum_{i=1}^{m} \widetilde{T}_i = T$, $t_B = (T - \sum_{i=1}^{k} \widetilde{T}_i)/(m-k)$. Hence, the optimal solution can be found by guessing k iteratively.

The details of Algorithm 1 are as follows. First, we calculate the current average allocable transmission time as t_B, as presented in line 2. We guess k from 1 to m by comparing $\widetilde{T}_k$ with t_B. We allocate each user k with $\widetilde{T}_k$ when $\widetilde{T}_k \le t_B$. Then we update t_B with the rest allocable time and go into the next loop. We repeat the operation in lines 5-11 till $\widetilde{T}_k > t_B$. Then t_B is allocated to each of the rest users. The time complexity of Algorithm 1 is $\mathcal{O}(m)$.

Algorithm 1 FBA

Require: T, Ascending $\{\widetilde{T}_1, ..., \widetilde{T}_m\}$
Ensure: $\{\widetilde{t}_1^*, ..., \widetilde{t}_m^*\}$

```
1:  Initialize T_total = T, count = m, k = 1
2:  Guess the value t_B = T_total / m
3:  A = ∅
4:  while k ≤ m do
5:      if T̃_k > t_B then
6:          break
7:      else
8:          add k into set A
9:          t̃*_k = T̃_i, T_total −= T_i, count −= 1, k += 1
10:         guess t_B = T_total / count
11:     end if
12: end while
13: for each i ∉ A do
14:     t̃*_i = t_B, add i into set B
15: end for
```

4.2 AP association and bandwidth allocation

We then propose MABU to solve JAA. In MABU, we first decide the association relationship between the APs and users and then schedule the transmission time of each AP.

Considering the limited transmission time of each AP, the choice of the associated AP should follow the two principles.

(1) The user should be associated to the AP with higher transmission rate to reduce the transmission time.
(2) The user should be associated with the AP with the lighted load to balance the load.

As Algorithm 2 illustrates, it first sorts the users by the data size S_i in descending order. In terms of an arbitrary user i, it tries to associate to each AP j and estimates the allocated transmission time of AP j. The allocated transmission time of AP equals to the aggregated transmission time demand of the users associated to the AP, including user i. Based on the two principles above, the user chooses the AP with the least allocated transmission time, as presented in line 4. Then we use FBA to schedule the transmission time of each AP.

Algorithm 2 MABU

Require: $\{S_1, ..., S_M\}, \{T_{11}, ..., T_{MN}\}$
Ensure: $\{x_{11}, ..., x_{MN}\}$
1: Initialize $x_{ij} = 0, \forall i \in 1..M, j \in 1..N$
2: Sort the users by the size of data S_i in descending order
3: **for** each user i in the sorted users **do**
4: $\quad j_{min} = \underset{j}{\operatorname{argmin}}(\sum_{k=1}^{i-1} x_{kj} T_{kj} + T_{ij})$
5: $\quad x_{ij_{min}} = 1$
6: **end for**
7: **for** each AP j **do**
8: $\quad$ User algorithm 1 to schedule the transmission time
9: **end for**

The time complexity of association decision is $\mathcal{O}(MN + M \log M)$. The time complexity of scheduling transmission time for N APs is $\mathcal{O}(MN)$. Thus the time complexity of the MABU is $\mathcal{O}(MN + M \log M)$.

4.3 Bound analysis of MABU

In this part, we analyze the lower bound of the solution given by MABU. We assume the transmission rate of all the users associated to different APs are the same. Therefore, r_{ij} equals to a constant, denoted as r. Then T_{ij} is simplified as T_i, with $T_i = S_i/r$.

In order to analyze the bound, we first estimate the upper bound of one AP's load when MABU is adopted. Then we compare the actual load with the upper bound to calculate the gap between MABU's solution and optimal solution.

Suppose $\mathbb{T} = \sum_{k=1}^{M} T_k$, i.e., $\mathbb{T}$ is the sum of the transmission time demand. MABU is in fact allocating the load of users averagely to each AP. Estimating the maximum load is to consider the most unbalanced case. This condition may happen when there are exact $N+1$ users with the same T_k associated to N APs. Based on this idea, we generate the following lemma:

Lemma 3 *Suppose there are N APs and M users. Each user's transmission time demand is no more than $\frac{2\mathbb{T}}{N+1}$. With MABU algorithm, the maximum load of one AP will be no greater than $\frac{2\mathbb{T}}{N+1}$, i.e.,*

$$max \sum_{k=1}^{M} x_{kj} T_k \leq \frac{2\mathbb{T}}{N+1}, j = 1, 2, ..., N. \tag{20}$$

Proof We prove the lemma by contradiction. Suppose that when the i_{th} is associated to AP j, the aggregated transmission time demand of the users associated to AP j is more than $\frac{2\mathbb{T}}{N+1}$, i.e.

$$\sum_{k=1}^{i-1} x_{kj} T_k + T_i > \frac{2\mathbb{T}}{N+1}. \tag{21}$$

We discuss the contradiction under the two cases:

Case 1 $T_i > \frac{\mathbb{T}}{N+1}$. Since the users are sorted by the size of data in descending order and the transmission rate between the APs and users are the same, the transmission time demand is also in descending order according to Eq. 3). Thus $\sum_{k=1}^{i} T_k \geq i T_i$.

As we assume that each T_i is no more than $\frac{2\mathbb{T}}{N+1}$, it satisfies $i \geq N+1$. We have $iT_i \geq (N+1)T_i > (N+1)\frac{\mathbb{T}}{N+1} = \mathbb{T}$

Combining the two expressions above, we have

$$\sum_{k=1}^{i} T_k > \mathbb{T}. \tag{22}$$

It means the aggregated transmission time demand of the i users is more than $\mathbb{T}$, which is contradictory against the assumption that the aggregated transmission time demand of the M users equals to $\mathbb{T}$.

Case 2 $T_i \leq \frac{\mathbb{T}}{N+1}$. Considering Eq. 21, we have

$$\sum_{k=1}^{i-1} x_{kj} T_k > \frac{\mathbb{T}}{N+1}. \tag{23}$$

Since the aggregated transmission time demand of the users associated to AP j is the least among all the APs according to Algorithm 2, it satisfies that

$$\begin{aligned}\sum_{k=1}^{i-1} T_k + T_i &\geq N \sum_{k=1}^{i-1} x_{kj} T_k + T_i \\ &= (\sum_{k=1}^{i-1} x_{kj} T_k + T_i) + (N-1) \sum_{k=1}^{i-1} x_{kj} T_k.\end{aligned} \tag{24}$$

By applying Eqs. 21 and 23 to Eq. 24, we have

$$\sum_{k=1}^{i} T_k > \frac{2\mathbb{T}}{N+1} + (N-1)\frac{\mathbb{T}}{N+1} = \mathbb{T}.$$

Same with Eq. 22, it is a contradictory. This completes the proof. □

Based on Lemma 3, we give the following theorem:

Theorem 4 *Let U' be the aggregated bandwidth utility given by MABU and U^* be the aggregated bandwidth utility given by the optimal solution. We have*

$$U^* - U' \leq M \log \frac{2\mathbb{T}/(N+1)}{T}. \tag{25}$$

Proof We assume $T \leq \frac{2\mathbb{T}}{N+1}$. If user i is allocated to transmission time $\frac{T}{2\mathbb{T}/(N+1)} T_i$ from its associated AP, the aggregated allocated transmission time of each AP is no more than T by applying Lemma 3. Since the time scheduling proposed by Algorithm 1 is optimal with the same AP association, we have

$$U' \geq \sum_{i=1}^{M} \log \frac{r \frac{T}{2\mathbb{T}/(N+1)} T_i}{T}$$
$$= \sum_{i=1}^{M} \log \frac{rT_i}{T} - M \log \frac{2\mathbb{T}/(N+1)}{T}. \tag{26}$$

Since the allocated transmission time of user i is no more than its transmission time demand, we have

$$\sum_{i=1}^{M} \log \frac{rT_i}{T} \geq U^*. \tag{27}$$

By applying Eqs. 26 and 27, we have

$$U^* - U' \leq M \log \frac{2\mathbb{T}/(N+1)}{T}. \tag{28}$$

The above proof is based on the assumption that each T_i is no more than $\frac{2\mathbb{T}}{N+1}$. When there is some T_i that is larger than $\frac{2\mathbb{T}}{N+1}$, we can think it equals to $\frac{2\mathbb{T}}{N+1}$. This is because thinking it equals to $\frac{2\mathbb{T}}{N+1}$ will not influence the value of U^* and U'. When we think it equals to $\frac{2\mathbb{T}}{N+1}$, the assumption that each T_i is no more than $\frac{2\mathbb{T}}{N+1}$ is satisfied and we can use the proof above. This completes the proof. □

Corollary 1 *If $\overline{b}'$ is the geometric mean of $\{\widetilde{b}'_i\}$, $\overline{b}^*$ is the geometric mean of $\{\widetilde{b}^*_i\}$ and $\mathbb{T} \leq NT$, then we have $\overline{b}' > 1/2\overline{b}^*$. $\{\widetilde{b}'_i\}$ is the bandwidth assignment given by MABU and $\{\widetilde{b}^*_i\}$ is the bandwidth assignment given by the optimal solution.*

Proof According to the definition of U^* and U', we have

$$U^* - U' = \log \frac{\prod_{i=1}^{M} \widetilde{b}^*_i}{\prod_{i=1}^{M} \widetilde{b}'_i} = \log(\frac{\overline{b}^*}{\overline{b}'})^M = M \log \frac{\overline{b}^*}{\overline{b}'}. \tag{29}$$

Combing Eq. 28 with Eq. 29, we have

$$\frac{\overline{b}^*}{\overline{b}'} \leq \frac{2\mathbb{T}}{(N+1)T}. \tag{30}$$

When $\mathbb{T} \leq NT$, i.e., the aggregated transmission time demand is no larger than the allocable transmission time of all the APs, we have

$$\frac{2\mathbb{T}}{(N+1)T} \leq \frac{2NT}{(N+1)T} < 2. \tag{31}$$

Combing Eq. 30 with Eq. 31, we have $\overline{b}^*/\overline{b}' < 2$. Then we get $\overline{b}' > 1/2\overline{b}^*$. This completes the proof. □

5 Performance evaluation

5.1 Evaluation setup and methodologies

In this part, we report our simulation results for the scenario where the network contains static users. The simulation of the scenarios and the algorithms is implemented by Python. All the experiments were performed on a Win 7 PC with Intel Core i3-3220 3.30GHz and 4G RAM.

We compare the performance of MABU with those of the following ones:

- Strong Signal First (SSF): the default AP selection scheme in the 802.11 standard.
- NLAO-PF: a time-based fairness algorithm proposed in [25]. This algorithm is selected, because it targets the similar nonlinear problem formulation as that of MABU. But it ignores the demand of users.
- Norm Load-Based AP selection (NLB): a decentralized AP selection algorithm with the objective of minimizing the norm load of AP proposed in [30]. By comparing with this scheme, we demonstrate the influence of different optimizing objectives on the user throughput.

All the algorithms are examined carefully according to the following performance metrics:

- Normalized transmission time demand on AP, is computed by $\frac{\sum_{i=1}^{M} x_{ij} T_{ij}}{T}$;
- Average AP utilization, is computed by $\frac{\sum_{j=1}^{N}(\sum_{i=1}^{M} x_{ij} t_{ij})/T}{N}$;
- Per-user throughput in Mbps and the corresponding statistical information;

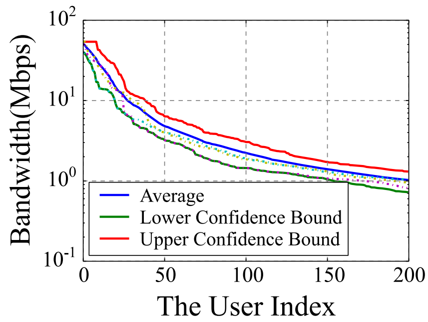


Fig. 1 User bandwidth demand in real traffic scenario

– Jain's Fairness Index [18], is defined as $J = \frac{(\sum_{i=1}^{M} y_i)^2}{M(\sum_{i=1}^{M} y_i{}^2)}$, where y_i is the allocated transmission time, the allocated bandwidth of user, or the demand on AP. Note that a larger value of $J \in [0, 1]$ indicates a better fairness.

For ease of comparison, we employ the same settings as those in [25], which are detailed as follows.

The network contains 20 APs placed on a 5 × 4 grid, with each on a grid point. The distance between two adjacent APs is 100m and the coverage area of AP is about 150m. There are users residing in the network with different bandwidth demands, resulting in different levels of network loads. Two types of user distribution are considered: (1) Uniform: the users are uniformly positioned within the coverage area of the network, with each on a grid point; (2) Hotspot: the users are randomly positioned in a circle-shaped hotspot area with a radius of 100m near the center of the 20-AP network.

Since the transmission rate of user is difficult to collect, we derive it based on the distance between the AP and user. The relationship between the distance and SINR is given by Eq. 1. We assume that all the APs have the same transmission power as 20dBm. We set the shadowing factor α as 4 and the noise power as -80dBm. The relationship between SINR and transmission rate is given in Table 2.

We capture the downlink data for 120 seconds in a campus network to simulate the bandwidth demand of user. We plot the bandwidth demand of the most active 200 users during each second in Fig. 1. The average value, lower and upper confidence bounds are highlighted in the fig. We can find that the bandwidth demand differs vastly between different users, no matter in which seconds. This motivates us to schedule the transmission with consideration of the bandwidth demand.

5.2 Evaluation results

In this part, we evaluate the performance of MABU and compare it to other algorithms. We mainly focus on the two aspects of performance: throughput increase and fairness guarantee. We first analyze the performance of MABU under different loads, which shows the average value of throughput increase and fairness index in each type of distribution. Then we investigate in detail why MABU can increase throughput while ensuring fairness. We find that

Fig. 2 Performance vs loads (uniform)

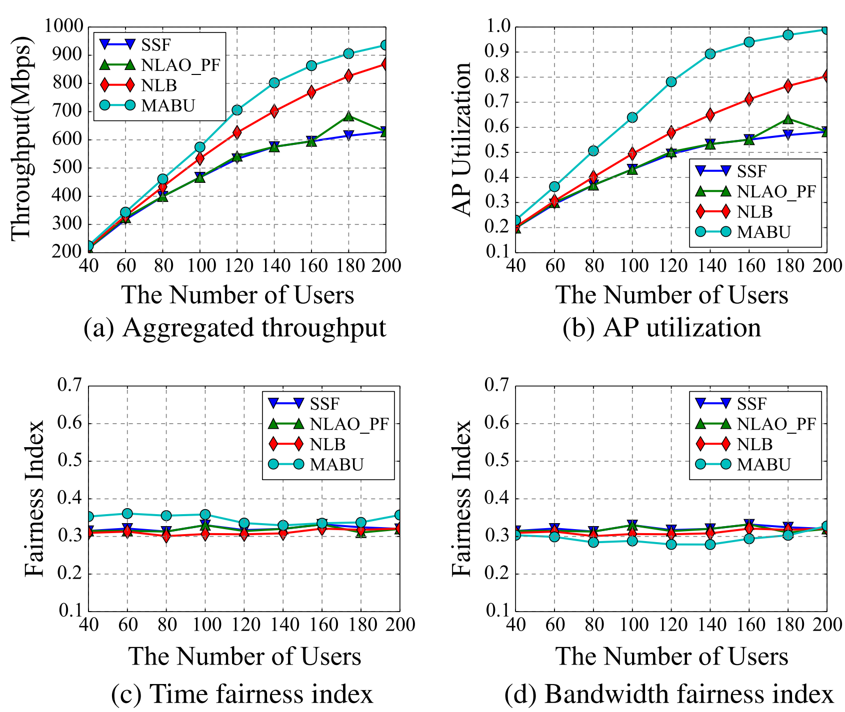
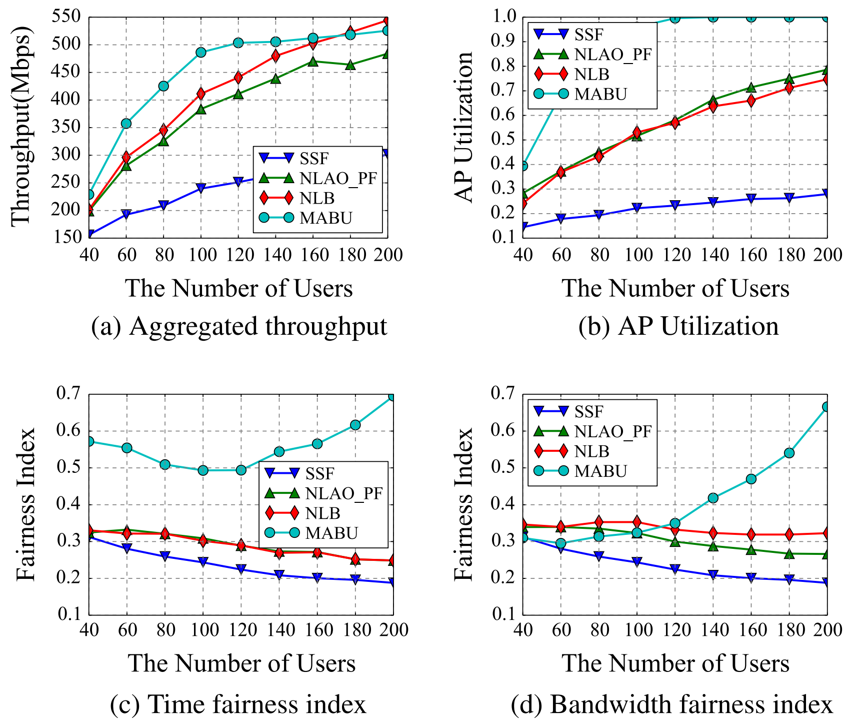


Fig. 3 Performance vs loads (hotspot)



MABU redistributes the load carefully and increases the average AP utilization, leading to high throughput and fairness guarantee.

5.2.1 Performance under different loads

In this part, we evaluate the performance of MABU under different loads in the two distributions. We fix the number of AP as 20 and change the scale of users to change the load. For each scale, the algorithm runs for 50 times to obtain the average effect. The aggregated throughput is the main metric to show the benefit. The AP utilization shows whether the load is proper. The two fairness indexes reflect the user fairness in time and throughput.

Comparing Fig. 2 with Fig. 3, the results in the two distributions show the similar trends. As Figs. 2a and 3a show, when the load is light, i.e., the number of users is about 40, the throughput of all the algorithms is very close. At this point, the AP utilization is lower than 50%, as Figs. 2b and 3b show. It illustrates that all the users' demands are satisfied, even though the load is not so balanced.

With the growth of user scale, the AP utilization increases. When the number of users is about 60 to 180 in uniform case and 60 to 120 in the hotspot case, the AP utilization of MABU increases larger than other algorithms, as Figs. 2b and 3b show. As a result, MABU gets a significant throughput increase, i.e., 14.4 % in the uniform and 23.1 % in hotspot distribution, compared with NLB. In the bound analysis of Section 4.3, we illustrate when the sum of user demands is close to the aggregated capacity of all the APs, the MABU algorithm has good performance. The simulation result also confirms the analysis. Since the fairness we use in our paper is time equality, the time fairness index of MABU is higher than other algorithms, as Figs. 2c and 3c show. The time equality leads to well bandwidth equality. Hence, the bandwidth fairness index of MABU is close to that of the other algorithms. as shown in Figs. 2d and 3d.

In the situation that the network is overloaded, the AP utilization of MABU reaches 1, as Figs. 2b and 3b show. This happens when there are about 200 users in the uniform distribution, while it is about 140 users in the hotspt distribution. The throughput can not be increased with the load balancing, thus the advantage of MABU becomes not obvious. The throughput of all the algorithms are close to each other.

Furthermore, the results in the two distributions have two main differences. One difference is that the network in the

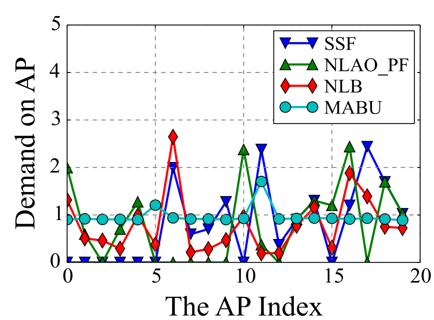


Fig. 4 Normalized transmission time demand

Table 3 Statistics of the fairness results

Type	Algorithm	Fairness of user's allocated time		Fairness of user's allocated bandwidth		Fairness of user's demand on APs	
		Mean	Jain's fairness index	Mean	Jain's fairness index	Mean	Jain's fairness index
Uniform	MABU	0.132	0.273	5.776	0.216	0.969	0.967
	NLAO_PF	0.076	0.232	4.078	0.232	0.794	0.484
	NLB	0.089	0.230	4.819	0.230	0.794	0.620
	SSF	0.076	0.232	4.078	0.232	0.794	0.484
Hotspot	MABU	0.199	0.448	5.277	0.217	1.155	0.956
	NLAO_PF	0.094	0.213	3.996	0.192	0.735	0.439
	NLB	0.087	0.222	3.703	0.241	0.724	0.336
	SSF	0.045	0.172	2.427	0.172	0.623	0.148

hotspot distribution is more easily to become overloaded. As Figs. 3b and 2b show, the AP utilization reaches 1 with smaller user scale in the hotspot distribution than the uniform distribution. It is because in the hotspot distribution, too many users concentrating in the center of the area, which makes few APs provide high transmission rate. While in the uniform distribution, each AP can provide high transmission rate for some users. The other difference is that the benefit of load balancing and fairness guarantee by MABU is more significant in the hotspot distribution. The reason is that the hotspot distribution is easier to become load imbalance and MABU performs effectively to solve the problem.

In the real world, that a network keeps too light load means the network capacity is wasted. That a network is seriously overloaded means the network capacity needs to be upgraded. Both the two conditions are not so valuable to balance the load. When the demand of users is close to the capacity of network, the load is proper to the network. At this condition, the result shows that MABU brings the great throughput increase and fairness guarantee. Then we give a detailed analysis on the reason why MABU can achieve the better performance.

5.2.2 Throughput increase

Since the results of uniform and hotspot distributions are similar, we take the uniform distribution as an example. Fig. 4 plots the sum of normalized transmission time demand of the users associated with each AP. It indicates the overall load of AP. The load is more balanced when the difference of normalized transmission time demand becomes smaller. We find that the load gap between the overloaded AP and light-loaded AP in MABU is much smaller than that in other algorithms. The fairness of user's demand on AP in Table 3 also shows the load difference. Since MABU has the highest Jain's fairness index, it achieves the best load balance.

As SSF only makes the user associate with the AP with the highest signal, many users may gather on the same AP. As a result, the load is unbalanced. NLAO-PF and NLB avoid the user gathering, but the user's demand is not considered by them. Due to load balance, MABU makes all the APs fully used, since the normalized transmission time demand on all the APs is close to 1. While in other algorithms, some APs are overloaded and the others are light-loaded. The transmission time of light-loaded APs is wasted. MABU redistributes the demand on the overloaded APs to the light-loaded APs. In a word, the balanced load of MABU is the root cause of throughput increase.

Figure 5 plots the achieved per-user bandwidth, with the users sorted by their bandwidth in the nondecreasing order. The per-user bandwidth of SSF and NLAO-PF is nearly the same, since the uniform distribution of users leads to the same maximum transmission rate for all the users. The throughput of MABU obviously outperforms that of other algorithms. It illustrates that MABU can better solve the problem of load imbalance and increase the throughput.

5.2.3 Fairness guarantee

In this part, the fairness of transmission time and bandwidth allocated to each user is investigated. Table 3 shows the time

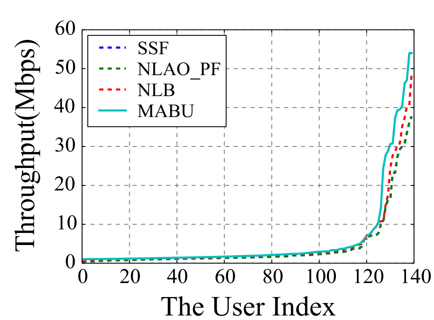


Fig. 5 Per-user throughput

fairness and bandwidth fairness of all the algorithms in the two distribution types.

In Table 3, the time fairness index of MABU is higher than that of the other algorithms in each distribution. That is because MABU focuses more on the fair time allocation. Furthermore, the mean of the allocated transmission time of MABU is higher than that of other algorithms, which is a result of the increased AP utilization. It results from the increase of AP utilization. As to the bandwidth fairness index in Table 3, the value of MABU is close to that of the other algorithms. It illustrates that the time fairness brings the bandwidth fairness. The mean of bandwidth of MABU is higher than that of other algorithms. The reason is that the increase of AP utilization leads to the throughput increase.

6 Conclusion

The wide spread of IEEE 802.11 based WLAN applications makes the network management more complex and challenging. Load balancing and fair bandwidth allocation are the two critical issues. In contrast to the existing work, we formulate the JAA problem with the consideration of the difference of the user's bandwidth demand. Then we propose the MABU algorithm as a solution to achieve the load balance and time-based fairness. As shown in the real-trace based evaluations, the conventional AP association strategies lead to an obvious load imbalance, especially in the uniform and hotspot cases. Our association strategy considers the bandwidth demand of the user, and effectively balances the load and brings a significant throughput increase. We have also shown that the bandwidth allocation algorithm provides better user fairness.

Acknowledgments Supported by National Natural Science Foundation of China (no. 61422206, 61120106008), National 863 project (no. 2015AA015701), and the Academy of Finland (no. 278207).

References

1. Abusubaih M, Wolisz A (2007) An optimal station association policy for multi-rate ieee 802.11 wireless lans. In: MSWiM. ACM, pp 117–123
2. Akella A, Judd G, Seshan S, Steenkiste P (2007) Self-management in chaotic wireless deployments. Wirel Netw 13(6):737–755
3. Ala-Laurila J, Mikkonen J, Rinnemaa J (2001) Wireless lan access network architecture for mobile operators. IEEE Commun Mag 39(11):82–89
4. Bejerano Y, Han S-J (2009) Cell breathing techniques for load balancing in wireless lans. IEEE Trans Mob Comput (TMC) 8(6):735–749
5. Bejerano Y, Han S-J, Li LE (2004) Fairness and load balancing in wireless lans using association control. In: MOBICOM. ACM, pp 315–329
6. Boyd S, Vandenberghe L (2009) Convex optimization. Cambridge University Press, Cambridge
7. Chen L (2010) A distributed access point selection algorithm based on no-regret learning for wireless access networks. In: VTC. IEEE, pp 1–5
8. Chen X, Yuan W, Cheng W, Liu W, Leung H (2013) Access point selection under qos requirements in variable channel-width wlans. IEEE Wireless Commun Lett 2(1):114–117
9. Chen X, Zhao Y, Peck B, Qiao D (2012) Sap: smart access point with seamless load balancing multiple interfaces. In: INFOCOM. IEEE, pp 1458–1466
10. Chen Z, Xiong Q, Liu Y, Huang C (2014) A strategy for differentiated access service selection based on application in wlans. In: INFOCOM WKSHPS. IEEE, pp 317–322
11. Cui Y, Li W, Cheng X (2011) Partially overlapping channel assignment based on node orthogonality for 802.11 wireless networks. In: INFOCOM. IEEE, pp 361–365
12. Dandapat SK, Mitra B, Choudhury RR, Ganguly N (2012) Smart association control in wireless mobile environment using max-flow. IEEE Trans Netw Serv Manag (TNSM) 9(1):73–86
13. Du L, Bai Y, Chen L (2007) Access point selection strategy for large-scale wireless local area networks. In: WCNC, pp 2161–2166
14. Garey MR, Johnson DS (1979), Computers and intractability: An introduction to the theory of np-completeness
15. Ge W, Ji H, Leung VC, Si P (2011) Access point selection for wlans with cognitive radio: a restless bandit approach. In: ICC. IEEE, pp 1–5
16. Gong D, Yang Y (2012) Ap association in 802.11 n wlans with heterogeneous clients. In: INFOCOM. IEEE, pp 1440–1448
17. Gong H, Kim J (2008) Dynamic load balancing through association control of mobile users in wifi networks. IEEE Trans Consum Electron 54(2):342–348
18. Jain R, Chiu D-M, Hawe WR (1984) A quantitative measure of fairness and discrimination for resource allocation in shared computer system. Eastern Research Laboratory, Digital Equipment Corporation, Hudson
19. Karimi OB, Liu J, Rexford J (2014) Optimal collaborative access point association in wireless networks. In: INFOCOM. IEEE, pp 1141–1149
20. Kauffmann B, Baccelli F, Chaintreau A, Mhatre V, Papagiannaki K, Diot C (2007) Self organization of interfering 802.11 wireless access networks. In: INFOCOM. ACM, p 2005
21. Kelly F (1997) Charging and rate control for elastic traffic. Eur Trans Telecommun 8:33–37
22. Keranidis S, Korakis T, Koutsopoulos I, Tassiulas L (2011) Contention and traffic load-aware association in ieee 802.11 wlans: Algorithms and implementation. In: WiOpt. IEEE, pp 334–341
23. Li L, Pal M, Yang YR (2008) Proportional fairness in multi-rate wireless lans. In: INFOCOM. IEEE
24. Li W, Cui Y, Cheng X, Al-Rodhaan MA, Al-Dhelaan A (2011) Achieving proportional fairness via ap power control in multi-rate wlans. IEEE Trans Wirel Commun (TWC) 10(11):3784–3792
25. Li W, Wang S, Cui Y, Cheng X, Xin R, Al-Rodhaan M, Al-Dhelaan A (2014) Ap association for proportional fairness in multirate wlans. IEEE/ACM Trans Networking (TON) 22(1):191–202
26. Lu M, Wu J (2011) Localized access point selection in infrastructure wireless lans with performance guarantee. Wirel Commun Mob Comput 11(8):994–1010
27. Viterbi AJ et al. (1995) CDMA: principles of spread spectrum communication, vol 129. Addison-Wesley, Reading
28. Xie L, Li Q, Mao W, Wu J, Chen D (2009) Achieving efficiency and fairness for association control in vehicular networks. In: ICNP. IEEE, pp 324–333

29. Xu F, Tan CC, Li Q, Yan G, Wu J (2010) Designing a practical access point association protocol. In: INFOCOM. IEEE, pp 1–9
30. Xu F, Zhu X, Tan CC, Li Q, Yan G, Wu J (2013) SmartAssoc: Decentralized access point selection algorithm to improve throughput. IEEE Trans Parallel Distrib Syst 24.12:2482–2491
31. Xue G, He Q, Zhu H, He T, Liu Y (2013) Sociality-aware access point selection in enterprise wireless lans. IEEE Trans Parallel Distrib Syst (TPDS) 24(10):2069–2078
32. Yen L-H, Li J-J, Lin C-M (2011) Stability and fairness of ap selection games in ieee 802.11 access networks. IEEE Trans Veh Technol 60(3):1150–1160