Fairness and Load Balancing in Wireless LANs Using Association Control

Yigal Bejerano, Seung-Jae Han and Li (Erran) Li
Bell Laboratories, Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974

ABSTRACT

Recent studies on operational wireless LANs (WLANs) have shown that user load is often unevenly distributed among wireless access points (APs). This unbalanced load results in unfair bandwidth allocation among users. We observe that the unbalanced load and unfair bandwidth allocation can be greatly alleviated by intelligently associating users to APs, termed *association control*, rather than having users greedily associate APs of best received signal strength.

In this study, we present an efficient algorithmic solution to determine the user-AP associations that ensure max-min fair bandwidth allocation. We provide a rigorous formulation of the association control problem that considers bandwidth constraints of both the wireless and backhaul links. Our formulation indicates the strong correlation between fairness and load balancing, which enables us to use load balancing techniques for obtaining near optimal max-min fair bandwidth allocation. Since this problem is NP-hard, we present algorithms that achieve a constant-factor approximate max-min fair bandwidth allocation. First, we calculate a fractional load balancing solution, where users can be associated with multiple APs simultaneously. This solution guarantees the fairest bandwidth allocation in terms of max-min fairness. Then, by utilizing a rounding method we obtain an efficient integral association. In particular, we provide a 2-approximation algorithm for unweighted greedy users and a 3-approximation algorithm for weighted and bounded-demand users. In addition to bandwidth fairness, we also consider time fairness and we show it can be solved optimally. We further extend our schemes for the on-line case where users may join and leave. Our simulations demonstrate that the proposed algorithms achieve close to optimal load balancing and max-min fairness and they outperform commonly used heuristic approaches.

Categories and Subject Descriptors

C.2.1. [**COMPUTER-COMMUNICATION NETWORKS**]: Network Architecture and Design – Wireless communication

General Terms

Algorithms, Design, Theory

Keywords

Wireless Local Area Networks (WLAN), IEEE 802.11, Max-Min Fairness, Load Balancing, Approximation Algorithms.

1. INTRODUCTION

In recent years, IEEE 802.11 wireless LANs (WLANs) have been rapidly deployed in enterprises, public areas and homes. Recent studies [1, 2, 3] on operational WLANs have shown that user load is often distributed unevenly among wireless access points (APs). In current WLANs, each user scans the wireless channel to detect its nearby APs and associate itself with the AP that has the strongest received signal strength indicator (RSSI), while ignoring its load. As users are, typically, not uniformly distributed, most of them may be associated with a few APs while adjacent APs may carry only light load or be idle. This load unbalance among APs is undesirable as it hampers the network from providing satisfactory service to its users. Initial studies [4, 5, 6] show that the problem can be reduced by balancing the load among the APs. This motivates the need for more efficient methods to select the user-AP association, termed *association control*. Obviously, association control can be used to achieve different objectives. For instance, it can be used to maximize the overall system throughput by shifting users to idle or lightly loaded APs. Although, this plausible objective can be obtained by each AP serving only its associated user with maximal data rate, clearly, this is not a desired system behavior. Consequently, a more desirable goal is to provide network-wide fair bandwidth allocation among users independent of their locations, while maximizing the fair share of each user. This type of fairness is known as *max-min fairness*. Informally, we say that an allocation of bandwidth through user-AP association is max-min fair if there is no way to give more bandwidth to any user without decreasing the allocation of a user with less or equal bandwidth. To achieve this objective, we present efficient user-AP association algorithms that ensure max-min fair bandwidth allocation among users and we show that this goal can be obtained by balancing the load on the APs.

1.1 Related Work

Association control has already been considered by both the research community and the industry. Various vendors of WLAN products have incorporated load-balancing features in their network device drivers, AP firmwares and

WLAN cards [7, 8]. In these proprietary solutions, the APs broadcast their load to users in their vicinity in the Beacon messages and each user chooses the least loaded AP. In [4, 5, 6], rather than using the RSSI as the association criteria, the proposed heuristics define different metrics and associate each user with the AP that optimizes these metrics. These metrics typically takes into account factors such as the number of users currently associated with an AP, the mean RSSI of users currently associated with an AP, the RSSI of the new user and the bandwidth a new user can get if it is associated with an AP. For example, Balachandran et al. [5] propose to associate new users with the AP that can provide a minimal bandwidth required by the user. If there are multiple such APs, the one with the strongest signal is selected. Most of these heuristics only determine the association of newly arrived users, except the one in [6]. Tsai and Lien [6] propose to reassociate users periodically each time some bandwidth thresholds are violated.

Load balancing has also been considered in cellular networks, both TDMA and CDMA networks. Usually, it is achieved via dynamic channel allocation (DCA) techniques [9]. These methods are not applicable in WLAN setting where each AP normally uses one channel and channel allocation is fixed. They are also not applicable to CDMA packet data networks [10]. Another approach is to use cell overlapping to reduce the blocking probability of calls and maximize the network utilization. In [11, 12], a newly arrived mobile station is associated with the base station with the greatest number of available channels. In [13], Lagrange and Jabbari address fairness issues in this approach by restricting the number of available channels for new calls that are made in overlapping areas. Tinnirello and Bianchi [14], propose to take into account the channel conditions of mobile stations associated with a base station. Recently, load balancing integrated with coordinated scheduling technique has been studied in [10] for CDMA networks. However, these techniques are not suitable to our goal, since they consider different objective functions, *e.g.*, blocking probability, or they do not provide any guarantee on the bandwidth allocated to each user.

Load balancing and max-min fairness have been extensively studied in the literature and we discuss here just the most relevant ones for our study. Most of the work on max-min fairness addresses the problem of allocating bandwidth to a set of pre-determined routes in a wired network [15, 16, 17]. The problem of selecting routes for providing max-min fair bandwidth allocation to a set of connections is much harder and has been studied in [18, 19]. Megiddo [18] addresses the problem in the setting of single-source fractional flow and presents a polynomial time algorithm that finds an optimal max-min fair solution. Extending this work, Kleinberg *et al.* [19], consider the problem where a connection is routed along a single path. In particular, their approach can be applied to the load balancing problem of parallel machine scheduling [20] where each job imposes the same load per unit time on the subset of machines in which it can be run, *i.e.,* a *load conserving* system. This problem is a special case of our problem where each user has the same bit rate to the APs it can associate with. They argue that a coordinate-wise constant-factor approximation cannot be found for this problem, and presented a prefix-sum 2-approximation algorithm to the allocation of fairest fractional solution. In other words, for every integer $k > 0$, the sum of the first k coordinates of the calculated allocation vector sorted in increasing order is at most twice the sum of the first k coordinates of the allocation vector of the fairest fractional assignment. They use Megiddo's algorithm [18] to compute a fractional solution and invoke the rounding scheme of Lenstra, Shmoys and Tardos [20] for obtaining an integral solution. However, their result cannot be directly applied to our problem since each user gets different rate from different APs, i.e. our jobs are *not load conserving.* In the context of online load balancing of unrelated parallel machines, Aspnes *et al.* [21] and Goel *et al.* [22] present an algorithm with a logarithmic competitive ratio when compared with the offline optimal allocation. We will apply these results to deal with the online case of our problem.

1.2 Our Contributions

In this paper, we present an algorithmic solution for determining use-AP association that ensures a network-wide max-min fair bandwidth allocation to the users. This goal is obtained by utilizing sophisticated algorithms that balance the load on the APs. Since some WLAN deployments utilize low capacity backhaul links for connecting the APs to a fixed backbone, *e.g.*, T1 lines, our solution considers capacity constraints of both the wireless channels and the backhaul connections. Previous studies in WLANs and cellular networks have not explicitly considered fairness in conjunction with load balancing. Actually, as we show in our simulations, if load-balancing is not done carefully, users may experience even poorer connections compared with the strongest received signal approach. To the best of our knowledge, we are the first that present a comprehensive association control scheme that provides guarantees on the quality of the bandwidth allocation against the optimal solution. Our solution can be used as the theoretical foundation of practical network management systems.

In our scheme, each mobile device is equipped with client software for monitoring the wireless channel quality that the user experiences from each one of its nearby APs. The client provides this information to a network control center (NOC) that determines the users' associations and updates the clients about its decisions. Accordingly, the users switch their associations. In this study, we do not address the issue of providing fair service for users associated with a given AP. We assume that such a mechanism is deployed at each AP, for instance, by using the emerging IEEE-802.11-e extension [23] or any fair scheduling mechanism, such as [24], [25] [26], and we build our association control solution on top of it.

To achieve our objectives we need a formal definition of the load of an AP. However, there is no such common notion in the literature. Several studies have already shown that naive definitions such as the number of users that are associated with an AP or the AP throughput do not reflect the AP load [1, 2, 3]. To this end, we introduce a rigorous definition of the load of an AP in WLANs. Generally speaking, the load that a user generates on its associated AP is inversely proportional to their effective bit rate. Our definition enables us to prove the strong correlation between balancing the load on the APs and providing fair service to the users. Moreover, it allows us to provide a rigorous formulation of the association control problem. Since the latter is NP-hard we develop several approximation algorithms for different settings of the fair service problem. Intuitively, we would like to guarantee to each user a bandwidth of at least $1/\rho$ of the bandwidth that it receives in the optimal (integral) solution, for a constant $\rho \geq 1$. However, due to the unbounded integrality gap, it is impossible to provide this type of approximation [19]. Accordingly, our guarantees are relative to an optimal fractional solution, where users can be associated with multiple APs simultaneously. First, we calculate a fractional load balancing solution that guarantees a max-min fair bandwidth allocation. It is the fairest among all allocations and we use it as the basis to compare with our integral solution. Then, we extend the rounding method of Shmoys and Tardos [27] to obtain an efficient solution where each user can only associate with one AP. In particular, we provide a 2-approximation algorithm for unweighted greedy users and a 3-approximation algorithm for weighted and bounded-demand users. In addition to bandwidth fairness, we also consider time fairness and we present an optimal algorithm. We further extend our schemes for the online case where users may join and leave. Our simulations demonstrate that the proposed algorithms achieve close to optimal load balancing and max-min fairness and they outperform popular heuristic approaches. Furthermore, our simulations show that in the presence of hot-spots our algorithms also provide higher network utilization than the one obtained by the strongest signal approach. Although, this work was done in the context of WLAN management, the methods developed in this work may be applicable to cellular networks as well.

2. THE NETWORK AND THE SYSTEM DESCRIPTION

2.1 The Network Model

We consider an IEEE 802.11 based wireless LAN (WLAN) that comprises a large number of *access points* (APs). We use A to denote the set of access points and let m denotes their number, i.e. $m = |A|$. All the APs are attached to a fixed infrastructure, which connects them to wired data networks such as the Internet. This infrastructure provides to each AP $a \in A$ a fixed transmission bit rate of R_a bits/second. Each AP has a limited transmission range and it can serve only users that reside in its range. We define the network coverage area to be the union of the area covered by each AP in A.

We use U to denote the set of mobile users that reside in the network coverage area and let $n = |U|$ denotes the total number of users in U. We assume that the users have a quasi-static mobility pattern. In other words, the users are free to move from place to place, but they tend to stay in the same physical locations for long time periods. This assumption is backed up by recent analysis of mobile user behavior [1, 2]. Each user is associated with a single AP to obtain connectivity service over a wireless channel. Note that the channel condition between an AP and a user is dynamic. However, since our goal is to achieve a long-term[1] fairness, our decisions are based on the long-term channel conditions observed by the users and the APs. The latter are mainly influenced by path loss and slow fading. For each user $u \in U$ and each AP $a \in A$, we use $r_{a,u}$ to denote the average *effective bite rate*[2] with which they can communicate.

Throughout this study, we first consider *greedy users* that consume all the bandwidth allocated to them by the network and always have traffic to send or receive. Furthermore, we assume that each user $u \in U$ has a weight w_u that specifies its priority. This weight is used to determine the bandwidth allocation, b_u, it entitles to have with respect to the other users. For instance, a user $u \in U$ entitles to have a bandwidth of $b_u = \frac{w_u}{w_v} \cdot b_v$ of any other user $v \in U$ in a nearby location. We then consider *bounded-demand users* that have specific maximal bandwidth demands, d_u, that upper bound their bandwidth needs. We assume that, each AP runs a scheduling algorithm that allocates bandwidth fairly to its associated users, *e.g.*, by using one of the mechanisms described in [23, 26].

[1]Long-term time scale is measured in terms of tens of seconds, which is still attractive for all practical purposes.

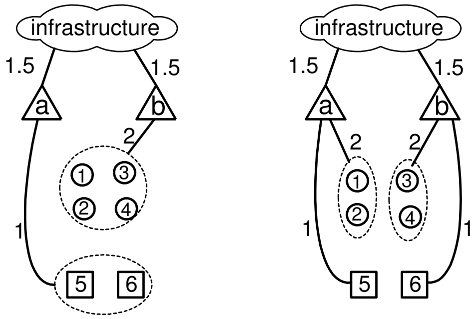
[2]The effective bit rate also takes into account the overhead of retransmissions due to reception errors.

2.2 The System Description

In this work we develop an algorithmic solution that determines the appropriate user-AP association for providing a long-term max-min fair service to all the mobile users. As such, our solution can be used as the theoretical foundations in the design of practical network management systems. It is well known that data flows have bursty characteristics and they generate dynamic load on the APs. Therefore, it is practically impossible to provide short-term fairness through association control to the users in multiple-AP networks without generating high communication overhead and disrupting ongoing sessions. Consequently, our scheme addresses the need to provide long-term fairness in both the worst and average case without interfering with the network operation. By using the greedy user model, our scheme maximizes the minimal bandwidth allocated to each user in worst case scenarios. Moreover, the bounded demand model can be used to maximize the average throughput that each user experiences in a fair system. In other words, our system takes into account the difference between users' average demands and balances their load among the APs. This makes sure that users with high average bandwidth requirements will be evenly distributed among the APs.

We now discuss the main implementation aspects of an association control system. First, the system requires relevant information on each user $u \in U$, such as its weight w_u, its average bandwidth demand d_u and the effective bit rate $r_{a,u}$ that it experiences from each AP $a \in A$. Second, it needs an algorithm to determine the appropriate user-AP association. Third, it needs a mechanism to enforce these association decisions.

We observe that the required information, mainly the effective bit rate $r_{a,u}$ between every user u and every AP a are not available from the existing 802.11 AP products, because an AP maintains the bit rate information only for the users who are currently associated with it. In fact, the effective bit rates can only be measured from the user side, by monitoring the signal strength of beacons from nearby APs. To this end, we assume that every user computer is equipped with a client software that periodically collects the bit rate information and evaluates its average bandwidth demand. The collected information is reported to a *network operation center* (NOC) which runs our algorithm to come up with the user-AP association decisions. Since the users are free to move, the NOC periodically recalculates the optimal user association by using one of the offline algorithms, described in Sections 4 and 5. Between two successive executions of the offline algorithm, the NOC uses an online method that maintains the APs' load as balanced as possible. We elaborate on the online algorithm in Section 6.



(a) An unfair association.

(b) The optimal association.

Figure 1: Examples of bottlenecks both over the wireless and the wired channels.

After determining a user association, the NOC notifies the user client software of his decision. The client changes the user association accordingly.

2.3 Wireless and Wired Bottlenecks

It is commonly believed that in wireless systems the wireless channels are the scarce resources that determine the bandwidth allocation. Although this is true in many cases, there are many WLANs where this assumption is not valid. For instance, consider an IEEE 802.11 network where the APs are connected to the infrastructure over $T1$ lines, with capacity around 1.5 Mbps, as illustrated in Example 1. Note that $T1$ lines are commonly used as the access link that connects small and medium companies to the Internet. Example 1 demonstrates the need to consider both the wireless and the wired channel when calculating load balanced associations.

Example 1. Consider a wireless system with 2 access points, **a** and **b**, and 6 users, enumerated from 1 to 6, as depicted in Figure 1. Users $1, 2, 3$ and 4 experience a bit rate of 2 Mbps from both APs, while users 5 and 6 have a bit rate of 1 Mbps from both APs. The APs are connected to a fixed network with $T1$ lines with capacity of 1.5 Mbps. In the following we consider two possible associations and we analyze the average bandwidth that they provide to the users.
Case I: A fair user association only from the wireless perspective - Consider the association depicted in Figure 1-(a). Here, the system can allocate a bandwidth of 0.5 Mbps to each user over the wireless links. However, while AP **a** can allocate a bandwidth of 0.5 Mbps to users 5 and 6 on its $T1$ line, AP **b** can only provide $\frac{3}{8}$ Mbps to its associated users over its $T1$ line. In this case, the wireless link of AP **a** is the bottleneck that affects the bandwidth allocation. Meanwhile, the wired link is the bottleneck of AP **b**.
Case II: A fair user association - Consider the association shown in Figure 1-(b). This association provides a bandwidth of 0.5 Mbps to each user over the wired and wireless channels. Observe that in this case different users may gain different service time on the wireless links and wired backhauls. For instance, user 5 captures $\frac{1}{3}$ of the service time of the $T1$ link of AP **a**, while, it is served $\frac{1}{2}$ of the time by its wireless channel. This ensures that user 5, indeed, receives a bandwidth of 0.5 Mbps. □

3. FAIRNESS AND LOAD BALANCING

In this section we provide formal definitions of fair bandwidth allocation and load balancing. Additionally, we describe some useful properties that we need for constructing our algorithmic tools. For the sake of simplicity, these definitions are given only for greedy users. We extend our definitions for bounded demand users in Section 5. In the following, we consider two association models. The first is a *single-association* model, so-called an *integral-association*, where each user is associated with a single AP at any given time. This is the association mode that is used in IEEE 802.11 networks. The second is a *multiple-association* model, also termed a *fractional-association*, that allows each user to be associated with several APs and to get communication services from them simultaneously. Accordingly, a user may receive several different traffic flows from different APs, and its bandwidth allocation is the aggregated bandwidth of all of them. This model is required to develop our algorithmic tools for the integral-association case. For both association models, we denote by U_a all the users that are associated with AP $a \in A$ and A_u denotes the set of APs that user $u \in U$ is associated with.

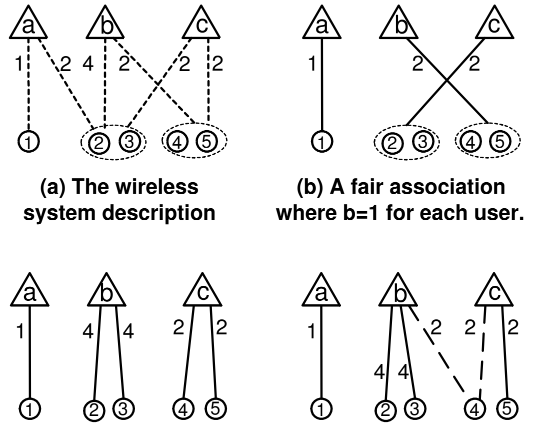
3.1 Max-Min Fairness

Consider a wireless network as described in Section 2.1. A *bandwidth allocation* is a matrix, $\mathcal{B} = \{b_{a,u} | u \in U,\ a \in A\}$, that specifies the average bandwidth, $b_{a,u}$, allocated to each user $u \in U$ by every AP $a \in A$. We denote by $b_u = \sum_{a \in A} b_{a,u}$ the *aggregated bandwidth* allocated to user u and let $\bar{b}_u = b_u / w_u$ be its *normalized bandwidth* (NB) allocation. On average, AP a is required to serve user u a period of $b_{a,u}/r_{a,u}$ over the wireless channel and a period of $b_{a,u}/R_a$ over the infrastructure link, at every time unit. Consequently, we say that a bandwidth allocation $\mathcal{B}$ is *feasible* if every AP $a \in A$ can provide the required bandwidth to all its associated users both in the wireless and the wired domains, that is, $\sum_{u \in U} b_{a,u}/r_{a,u} \leq 1$ and $\sum_{u \in U} b_{a,u}/R_a \leq 1$. In the case of an integral-association, we also require that each user is associated with a single AP.

Intuitively, a system provides a fair service if all users have the same allocated bandwidth[3]. Unfortunately, such a degree of fairness may cause significant reduction of the network throughput, since all users get the same bandwidth allocation as the bottleneck users, as we illustrate in Example 2 below. The common approach to address this issue of fair allocation that also maximizes the network throughput is to provide *max-min fairness* [17]. Informally, a bandwidth allocation of a weighted system is called *max-min fair* if there is no way to increase the bandwidth of a user without decreasing the bandwidth of another user with the same or less normalized bandwidth. Consider a bandwidth allocation $\mathcal{B}$ and let $\bar{b}_u$ be the normalized bandwidth allocated to user $u \in U$. We define the *normalized bandwidth vector* (NBV), $\vec{B} = \{\bar{b}_1, \cdots, \bar{b}_n\}$ as the users' normalized bandwidth allocations sorted in increasing order and users are renamed according to this order.

Definition 1 (*Max-Min Fairness*). *A feasible bandwidth allocation $\mathcal{B}$ is called* max-min fair *if its corresponding NBV $\vec{B} = \{\bar{b}_1, \cdots, \bar{b}_n\}$ has the same or higher lexicographical value than the NBV $\vec{B}' = \{\bar{b}'_1, \cdots, \bar{b}'_n\}$ of any other feasible bandwidth allocation $\mathcal{B}'$. In other words, if $\vec{B} \neq \vec{B}'$ then*

[3]The same normalized bandwidth in the case of weighted system.



(a) The wireless system description

(b) A fair association where b=1 for each user.

(c) The Max-Min fairness single association

(d) The Max-Min fairness multiple association

Figure 2: Examples of a wireless system with 3 APs and 5 users.

there is an index j such that $\bar{b}_j > \bar{b}'_j$ and for every index $i < j$, it follows that $\bar{b}_i = \bar{b}'_i$.

Example 2. Consider a wireless system with 3 APs, $A = \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$, and 5 users, $U = \{1, 2, 3, 4, 5\}$, as depicted in Figure 2-(a). In this figure, doted lines represent possible association and the number near each line represents the bit rate $r_{a,u}$ of the corresponding wireless link. All the users have weight 1 and we assume that all the APs are connected to a high bandwidth infrastructure. Figure 2-(b) presents a feasible fair association in which every user receives a bandwidth $b = 1$, where the solid lines represents the users' associations. Note that this is the maximal bandwidth that can be allocated to user 1. Thus, one can argue that this is the optimal bandwidth allocation. However, in Figures 2-(c) and (d), we describe two feasible associations, in which each user get at least 1 unit of bandwidth. Here, the solid lines indicates an integral association and the dashed line represents fractional association. Figure 2-(c) presents the integral max-min fair allocation with NBV $\vec{B} = \{1, 1, 1, 2, 2\}$. While, Figure 2-(d) introduces the fractional max-min fair allocation with NBV $\vec{B} = \{1, \frac{4}{3}, \frac{4}{3}, \frac{4}{3}, \frac{4}{3}\}$. □

Clearly, the NBV of a fractional max-min fairness allocation always has the same or higher lexicographical value than the NBV of the integral max-min fairness allocation. We will use this property to construct our solution for the integral-association case. Furthermore, consider a max-min bandwidth allocation $\mathcal{B}$ of either a fractional or an integral association. The users can be divided into *fairness groups*, such that each fairness group, $F_k \subseteq U$, consists of all users that experience the same normalized bandwidth allocation, denoted by $\bar{b}_k$.

Theorem 1. *Let $\mathcal{B}$ be a max-min fair bandwidth allocation and let $\{F_k\}$ be its corresponding fairness groups. Then all the users served by a given AP belongs to the same fairness group. Formally, for each fairness group F_k, $\bigcup_{u \in F_k} \bigcup_{a \in A_u} U_a = F_k$.*

Proof: Initially we prove that $\bigcup_{u \in F_k} \bigcup_{a \in A_u} U_a \supseteq F_k$. This is trivial since every user $u \in F_k$ is included in the set U_a for each AP a it is associated with. Now, we turn to prove that $\bigcup_{u \in F_k} \bigcup_{a \in A_u} U_a \subseteq F_k$. In the case of an integral association, this is satisfied since each user is associated with a single AP and this AP guarantees the same normalized bandwidth allocation to all its associated users. For fractional-association, lets suppose that this property is not valid. Thus, there is an AP a that serves users of two different fairness groups F_j and F_i. Suppose that $\bar{b}_j < \bar{b}_i$. Thus, AP a may increase the bandwidth of its associated users in F_j on behalf of its associated users in F_i. This results in a NBV with a higher lexicographical value. However, this contradicts the assumption that the given allocation is max-min fair. □

3.2 Min-Max Load Balancing

It is widely accepted that the prime approach for obtaining a fair service is balancing the load on the access points. However, for WLANs the notion of load is not well defined. Several recent studies [1, 2, 3] have shown that neither the number of users associated with an AP nor its throughput reflect the AP's "load". This motivates the need for an appropriate definition. *Intuitively, the load of an AP needs to reflect its inability to satisfy the requirements of its associated users and as such it should be inversely proportional to the average bandwidth that they experience.* Our load definition captures this intuition and it is also aligned with the standard load definition that are used in the computer science literature, *e.g.*, scheduling of unrelated parallel machines [28]. Consequently, we are able to *extend* existing load balancing techniques to balance the AP loads and obtain a fair service.

For our needs, we define the notion of fractional association. A *fractional association* is a matrix $\mathcal{X} = \{x_{a,u} | a \in A \wedge u \in U\}$, such that for each user $u \in U$, Equation $\sum_{a \in A} x_{a,u} = 1$ holds. Each parameter $x_{a,u} \in [0, 1]$ specifies the *fractional association of user u with AP a*. Generally speaking, $x_{a,u}$ reflects the fraction of user u's total flow that it expects to get from AP a. A fractional association $\mathcal{X}$ is termed *feasible* if the users are associated only with APs that can serve them, *i.e.*, for each pair $a \in A$ and $u \in U$, it follows that $x_{a,u} > 0$ only if $r_{a,u} > 0$. Moreover, a feasible association matrix that consists of just 0 and 1 is termed an *integral association.*

Consider a feasible association $\mathcal{X}$, either integral or fractional. We define the *load induced by user u on AP a* to be the time that is required of AP a to provide user u a traffic volume of size $x_{a,u} \cdot w_u$. Thus, user u produces a load of $x_{a,u} \cdot w_u / r_{u,a}$ on the wireless channel of AP a and a load of $x_{a,u} \cdot w_u / R_a$ on its backhaul link. Consequently, we define the *load*, y_a, on AP a to be the period of time that takes AP a to provide a traffic volume of size $x_{a,u} \cdot w_u$ to all its associated users $u \in U_a$. Formally,

Definition 2 (*Access-Point Load). The* load *on an AP $a \in A$, denoted by y_a, is the maximum of its aggregated loads on both its wireless and infrastructure links produced by all the users. Thus,*

$$y_a = \max \left\{ \sum_{u \in U} \frac{x_{a,u} \cdot w_u}{r_{u,a}}, \sum_{u \in U} \frac{x_{a,u} \cdot w_u}{R_a} \right\}$$

Therefore, the load of an AP is given in terms of the time it takes to complete the transmission of certain traffic volume from each associated user. This is not surprising, since the load should be inversely proportional to the bandwidth that the AP provides to its users. Furthermore, the bandwidth that AP a provides to user u is

$$b_{a,u} = x_{a,u} \cdot w_u / y_a \quad (1)$$

We define the *load vector* $\vec{Y} = \{y_1, \cdots, y_m\}$ of an association matrix $\mathcal{X}$ to be the n-tuple consisting of the load of each AP sorted in decreasing order.

Definition 3 (*Min-Max Load Balanced Association*). *A feasible association $\mathcal{X}$ is termed* min-max load balanced *if its corresponding load vector $\vec{Y} = \{y_1, \cdots, y_m\}$ has the same or lower lexicographical value than any other load vector $\vec{Y}' = \{y'_1, \cdots, y'_m\}$ of any other feasible assignment $\mathcal{X}'$. In other words, if $\vec{Y} \neq \vec{Y}'$, then there is an index j such that $y_j < y'_j$ and for every index $i < j$, it follows that $y_i = y'_i$.*

Example 3. Consider the wireless system described in Example 2. Figure 2-(c) presents the min-max load balanced association for the single-association case and its load vector is $\vec{Y} = \{1, 1, \frac{1}{2}\}$. While, Figure 2-(d) introduces the min-max load balanced association for the multiple-association case and its load vector is $\vec{Y} = \{1, \frac{3}{4}, \frac{3}{4}\}$. Recall that in this case the association of user 4 is $x_{\mathbf{b},4} = x_{\mathbf{c},4} = \frac{1}{2}$, thus the load that it induces on each one of these APs is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. □

Consider the min-max balanced association $\mathcal{X}$ and its corresponding load vector $\vec{Y}$. Recall that users can be partitioned into fairness groups. Similarly, APs can be partitioned into *load groups*. Each load group, $L_k \subseteq A$ contains all the APs with the same load, denoted by y_k. Furthermore, lets assume that the indices of the load groups are assigned in decreasing order according to their corresponding loads.

Theorem 2. *Consider a min-max load balanced association $\mathcal{X}$ and let $\{L_k\}$ be its APs partitioned into load groups, then each user is associated with APs with the same load,* i.e., *, for each load group L_k we have $\bigcup_{a \in L_k} \bigcup_{u \in U_a} A_u = L_k$.*

The proof of Theorem 2 is along the same line of the proof of Theorem 1.

Theorem 3. *Consider a min-max load balanced association $\mathcal{X}$ and consider any user $u \in U$ and any one of its associated APs $a \in A_u$. Then, the bandwidth allocation for user u determined by $\mathcal{X}$ is $b_u = w_u / y_a$.*

Proof: Since $\mathcal{X}$ is a min-max load balanced association, it follows that $\sum_{q \in A_u} x_{q,u} = 1$ and all the APs $q \in A_u$ has the same load y_a as the selected AP a. By Equation 1, we have,

$$b_u = \sum_{q \in A_u} b_{q,u} = \sum_{q \in A_u} x_{q,u} \cdot w_u / y_q = w_u / y_a$$

□

From Theorems 2 and 3, we have Corollary 1.

Corollary 1. *Consider a min-max load balanced association $\mathcal{X}$. $\mathcal{X}$ partitions the APs into load groups $\{L_k\}$, where the load on each AP in a group L_k is y_k. It also divides the users into fairness groups $\{F_{k'}\}$ such that all the users in the same group experience the same normalized bandwidth $\bar{b}_{k'}$. Furthermore, the APs of a given load group L_k serve only users from a corresponding fairness group $F_{k'}$ and the normalized bandwidth that each user in $F_{k'}$ experiences is $1/y_k$.*

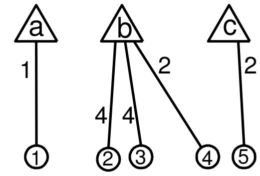


Figure 3: Examples of a single association that is min-max load balanced but is not max-main fair.

In the following we refer to the load group of the most loaded APs and the corresponding fairness group as the bottleneck groups. We now turn to prove the strong relationship between fairness and load balancing in the case of fractional-association. A sketch of Theorem 4's proof can be found in Appendix A.

Theorem 4 (The Main Theorem). *In the fractional-association case, a min-max load balanced association $\mathcal{X}$ defines a max-min fair bandwidth allocation and vise versa.*

Unfortunately, Theorem 4 is not satisfied in the case of a single association, as we illustrate in Example 4. However, by using approximation algorithm we can provide an approximated solution to these NP-hard problems by rounding the calculated fractional solutions, as described in Section 4.

Example 4. Consider the wireless system described in Example 2. As mentioned above, Figure 2-(c) presents the min-max load balanced association $\mathcal{X}$. Its load vector is $\vec{Y} = \{1, 1, \frac{1}{2}\}$ and its corresponding NBV is $\vec{B} = \{1, 1, 1, 2, 2\}$. However, the association $\mathcal{X}'$ presented in Figure 3 has the same load vector while its NBV vector is $\vec{B}' = \{1, 1, 1, 1, 2\}$. Observe that in both associations $\mathcal{X}$ and $\mathcal{X}'$, one of the two APs **b**,**c** has a load 1 and the other has $\frac{1}{2}$. However, in association $\mathcal{X}$ only two users are associated with each one of these two APs, while in association $\mathcal{X}'$ three users are associated with AP **b** whose load is 1 and only one user is associated with AP **c** whose load is $\frac{1}{2}$. This disparity leads to the sub-optimality of association $\mathcal{X}'$. □

4. ASSOCIATION CONTROL OF GREEDY USERS

In this section we present our algorithms that give approximate solutions to the integral max-min fair bandwidth allocation for greedy users. This is a challenging problem, as even identifying the users in the bottleneck fairness group and finding their normalized bandwidth is NP-hard. From Definition 2 and Equation 1, it follows that the minimal normalized bandwidth allocation is maximized when the maximal load on the APs is minimized, *i.e.,* when the load on the APs is balanced. Our load balancing problem is actually an extension of the *scheduling unrelated parallel machines* problem [20, 27]. For this problem, Lenstra, Shmoys and Tardos, in [20], proved that for any positive $\epsilon < \frac{1}{2}$ there is no polynomial-time $(1+\epsilon)$-approximation algorithm exists, unless $P = NP$. Moreover, in [20] and [27], they gave a polynomial-time 2-approximation algorithms, which is currently the best known approximation ratio achieved in polynomial time. However, unlike the solutions given in [20, 27] that balance the load on the most loaded machines, our solution seeks for a complete min-max load balanced association. We consider three different settings. We provide a 2-approximation algorithm for unweighted users, a

3-approximation algorithm for weighted users and an optimal solution for fair time allocation.

4.1 ρ^*-Approximation with Threshold

Intuitively, we would like to guarantee to each user a bandwidth of at least $1/\rho$ of the bandwidth that it receives in the optimal integral solution, for a constant $\rho \geq 1$. However, due to the unbounded integrality gap, it is impossible to provide this type of approximation [19]. Let y_a^{int} and y_a^{frac} be the load on a given AP $a \in A$ in the optimal integral and fractional solutions, respectively. We show that there is neither upper nor lower constant bounds for the ratio y_a^{int}/y_a^{frac}.

EXAMPLE 5. Consider a wireless network with 2 APs $\{a, b\}$ and 2 users $\{1, 2\}$, where $r_{a,1} = r_{b,1} = c$ and $r_{a,2} = r_{b,2} = c/(2 \cdot c - 1)$ for a given constant $c > 1$. In the optimal fractional solution, the load on each AP is $y_a^{frac} = y_b^{frac} = 1/2 \cdot (1/c + (2c-1)/c) = 1$. However, in any integral solution, one AP, let say a, experiences a load of $y_a^{int} = 1/c$ while the other has a load of $y_b^{int} = (2c-1)/c$. Consequently, the ratio $y_a^{int}/y_a^{frac} = 1/c$ and it cannot be lower bounded by any constant. □

Example 5 demonstrates the difficulty to provide guarantees that are comparable with the integral solution. Accordingly, our guarantees are relative to an optimal fractional solution. Recall that the NBV of the latter has the same or higher lexicographical value than the NBV of the optimal integral solution. Thus, the fractional solution is at least as fair as an integral one. In fact, the optimal fractional solution is the fairest among all feasible allocations.

EXAMPLE 6 (FROM [28]). Consider a wireless network with m APs, denoted by A, and a single user u, and let $r_{a,u} = 1$ for each $a \in A$. Clearly, in the fractional solution the load of u is equally divided among all the APs and thus for each $a \in A$, it follows that $y_a^{frac} = 1/m$. However, in the integral solution user u is associated with a single AP, lets say a, and the load of this AP is $y_a^{int} = 1$. Thus, the ratio between y_a^{int} and y_a^{frac} is m and it cannot be upper bounded by any constant. □

This obstacle occurs since the fractional load is smaller than the load induced by a single user on any AP. Since, our practical goal is to reduce the load of highly-loaded APs, there is no need to balance the load of APs with load below a certain threshold T. To this end, we select T to be the maximal load that a user may generate on an AP as formulated in Equation 2.

$$T = \max_{\{u,a|u \in U \wedge a \in A \wedge r_{a,u} > 0\}} \max\{\frac{w_u}{r_{a,u}}, \frac{w_u}{R_a}\} \qquad (2)$$

Recall that T is indeed a very small value and in practical 802.11 networks $T \leq 1$ sec/Mb. In light of these difficulties, we now formulate load and bandwidth guarantees that we provide in our solutions.

DEFINITION 4. *Let $\mathcal{X}^*$ be a fractional min-max load balances association and let y_a^* be the load of each AP $a \in A$. Then, a ρ^** min-max load balanced approximation with threshold T *is an integral association $\mathcal{X}$ such that the load y_a of each AP $a \in A$ satisfies $y_a \leq \rho \cdot \max\{y_a^*, T\}$.*

DEFINITION 5. *Let $\mathcal{X}^*$ be a fractional max-min fair association, and let $\bar{b}_u^*$ be its normalized bandwidth allocation to user $u \in U$. Then, a ρ^** max-min fairness approximation with threshold T *is an integral association $\mathcal{X}$ such that the normalized bandwidth $\bar{b}_u$ of each user $u \in U$ satisfies $\bar{b}_u \geq \frac{1}{\rho} \cdot \min\{\bar{b}_u^*, \frac{1}{T}\}$.*

```
Alg Integral_Load_Balancing(A, U)
    X^frac ← Fractional_Load_Balancing(A, U)
    X^int ← Rounding(X^frac)
    return X^int
end
```

Figure 4: A formal description of the integral load balancing algorithm

4.2 The Scheme Overview

We now present our *integral load balancing algorithm*. The algorithm comprises two steps. Initially, it calculates the optimal fractional association *i.e.,* the min-max load balanced fractional association. From Theorem 4, it follows that this association is also a min-max fair fractional allocation. Then, the algorithm utilizes the rounding method of Shmoys and Tardos [27] to obtain an approximate max-min fair integral association. A formal description of the algorithm is provided in Figure 4.

4.2.1 The Fractional Load balancing Algorithm

Our algorithm results from the observations made in Section 3. More specific, let $\mathcal{X}$ be a max-min load balanced fractional association. According to Corollary 1, $\mathcal{X}$ partitions the APs and the users into load groups $\{L_k\}$ and corresponding fairness groups $\{F_k\}$, such that the APs in a load group L_k are associated only with the users in a fairness group F_k and vise versa. Moreover, all APs in a given load group L_k have the same load y_k and the corresponding users in the fairness group F_k experience a normalized bandwidth allocation of $1/y_k$.

Based on these observations, we obtain an iterative algorithm that calculates the load groups and their corresponding load values. We refer to this algorithm as the *fractional load balancing algorithm*. To ease our presentation, lets assume that the load groups are enumerated in decreasing order according to their loads y_k. Thus, the APs in the group L_1 are the ones with the maximal load according to the association $\mathcal{X}$. We refer to the group L_1 as the *bottleneck load group* and the set F_1 of their associated users as the *bottleneck fairness group*. Moreover, load y_1 on the APs in L_1 is termed as the *bottleneck load* and it is denoted by $\tilde{Y}$.

The iterative algorithm detects the load groups according to their indices until all the users are associated with APs. At each iteration, the algorithm invokes the *bottleneck-group detection routine* to calculate the bottleneck load and fairness groups and updates its current fractional solution accordingly. Before proceeding to the next iteration, the algorithm removes the bottleneck load and fairness group from the system. Note that in the new iteration the load group with the succeeding index becomes the bottleneck group. A formal description of the algorithm is given in Figure 5.

Now, we turn to present the bottleneck-group detection routine. In this routine, we denote by $\tilde{L}$ and $\tilde{F}$ the load and fairness bottleneck group respectively. This routine consists of three steps. In the *first step*, we calculate the optimal bottleneck load value $\tilde{Y}$, that upper bounds the load y_a of every AP $a \in A$ in any min-max load balancing association.

```
Alg Fractional_Load_Balancing(A, U)
    Initialize 𝒳
    k ← 1
    while (U ≠ ∅) do
        {L_k, F_k 𝒳_k} ← bottleneck_detection(A, U)
        Update 𝒳 with the association 𝒳_k.
        A ← A − L_k
        U ← U − F_k
        k ← k + 1
    end of while
    return 𝒳
end
```

Figure 5: A formal description of the fractional load balancing algorithm

To infer its value, we utilize a linear program, denoted as **LP1**, that calculates a feasible association $\mathcal{X}$, which also minimizes the maximal load on all the APs over both their wireless and wired channels.

$$\begin{array}{ll} \textbf{LP1}: & \min\ \tilde{Y} \\ \textit{subject to}: & \\ \forall a \in A: & \sum_{u\in U}(w_u \cdot x_{a,u})/r_{a,u} \leq \tilde{Y} \\ \forall a \in A: & \sum_{u\in U}(w_u \cdot x_{a,u})/R_a \leq \tilde{Y} \\ \forall u \in U: & \sum_{a\in A} x_{a,u} = 1 \\ \forall u \in U,\ \forall a \in A: & x_{a,u} \in [0,1] \end{array}$$

Note that **LP1** minimizes the maximal load on all the APs. Consequently, the calculated association $\mathcal{X}$ ensures that the load on each AP in the bottleneck load group $\tilde{L}$ is exactly $\tilde{Y}$ and it also specifies the association of the APs in $\tilde{L}$ with the corresponding users in $\tilde{F}$. However, $\mathcal{X}$ does not optimize the load on the other APs, which may be as high as $\tilde{Y}$. We observe that, in the worst case, **LP1** may calculate a bad association such that the load on *all* the APs is $\tilde{Y}$ although the optimal association contains several load groups with lower loads, as illustrated in Example 7.

Example 7. Consider the wireless system described in Example 2 and the association presented in Figure 2-(b). This association induces a load of $\tilde{Y} = 1$ on all the APs. However, from Example 3 we know that a min-max fair allocation generates a load of $\frac{3}{4}$ on AP **b** and **c** and accordingly the allocated bandwidth to each of the associated user $2, 3, 4, 5$ is $\frac{4}{3}$. □

Such association is very deceptive, since it gives the impression that all the APs are included in the bottleneck load group. Therefore, we have developed a method to separate the APs in the bottleneck load group $\tilde{L}$ from the rest of the APs. In the *second step*, we use an auxiliary linear program, **LP2**, which enables us to identify whether some APs are not in $\tilde{L}$ or whether $\tilde{L}$ comprises all the APs. **LP2** is based on Property 1, proved in Appendix B

Property 1. *The bottleneck load group $\tilde{L}$ contains all the APs if there is* no *feasible association such that*
(1) Every AP has a load at most $\tilde{Y}$ and
(2) Some APs have load strictly less than $\tilde{Y}$.

LP2 looks for an association $\mathcal{X}$ that minimizes the overall load on all the APs subject to the constraint that the load on each AP is no higher than $\tilde{Y}$.

$$\begin{array}{ll} \textbf{LP2}: & \min\ \sum_{a\in A} y_a \\ \textit{subject to}: & \\ \forall a \in A: & y_a \leq \tilde{Y} \\ \forall a \in A: & \sum_{u\in U}(w_u \cdot x_{a,u})/r_{a,u} \leq y_a \\ \forall a \in A: & \sum_{u\in U}(w_u \cdot x_{a,u})/R_a \leq y_a \\ \forall u \in U: & \sum_{a\in A} x_{a,u} = 1 \\ \forall u \in U,\ \forall a \in A: & x_{a,u} \in [0,1] \end{array}$$

Clearly, if the bottleneck load groups do not comprise all the APs then **LP2** should find an association where some APs have load strictly less than $\tilde{Y}$ and these APs are not included in $\tilde{L}$. However, **LP2** does not specify the APs that are included in $\tilde{L}$, as APs with loads equal to $\tilde{Y}$ are not necessarily included in $\tilde{L}$, as we illustrate in Example 8 bellow. Consequently, in the *third step*, we introduce a method to separate $\tilde{L}$ from the other APs based on the results given in Definition 3; The load of each AP $a \notin \tilde{L}$, $y_a = \tilde{Y}$, can be reduced by shifting the association of some of its associated users to less loaded APs.

Consider the association $\mathcal{X}$ determined by **LP2**. Initially, we build a directed graph $G = (V, E)$ that each node $a \in V$ represents an AP in A, and there is an edge $(a, b) \in E$ if AP a can shift some load to AP b. In other words, there exists a user $u \in U$ such that $x_{a,u} > 0$ and $r_{b,u} > 0$. Note that the graph $G = (V, E)$ represents paths in which loads may be shifted. The method colors each node either white or black, where *white* represents APs not in $\tilde{L}$ and *black* indicates APs that *may* be included in the bottleneck group. Thus, the initial color of each node with load $\tilde{Y}$ is black, while the other nodes are colored white. Now, as long as there is an edge $(a, b) \in E$ such that node a is black and node b is white, we color node a white. At the end of this iterative process, the bottleneck load group $\tilde{L}$ comprises all the APs that are colored black and their associated users $\tilde{F}$ are determined by the association $\mathcal{X}$ calculated by **LP1** (or **LP2**). Finally, the bottleneck-group detection routine returns the sets $\tilde{L}$, $\tilde{F}$ and their corresponding user-AP association $\tilde{\mathcal{X}}$. A formal description of this routine is given in Figure 6 and an example of its execution is provided in Example 8.

Example 8. Consider the wireless system described in Example 2. In this case, a possible association $\mathcal{X}$ calculated by **LP2** is the one depicted in Figure 7-(a). Figure 7-(b) represents the calculated graph $G = (V, E)$ and the nodes' initial colors. Recall that $y_{\mathbf{a}} = y_{\mathbf{c}} = 1$ and $y_{\mathbf{b}} = \frac{1}{2}$. Moreover, some load of user 2 or 3 can be shift from AP **b** to APs **c** or **a**, which is indicated by the edges $(\mathbf{b}, \mathbf{c})$ and $(\mathbf{b}, \mathbf{a})$, and some load of user 4 or 5 can be shift from AP **c** to AP **b**, which is indicated by the edge $(\mathbf{c}, \mathbf{b})$. In the following, our routine colors AP **c** with white and ends the coloring iterations. Consequently, the computed groups are $\tilde{L} = \{\mathbf{a}\}$ and $\tilde{F} = \{1\}$, which are indeed the bottleneck groups. □

Theorem 5. *The load balancing algorithm calculates a min-max load balanced association in the case that users are allowed to have fractional associations with APs.*

Routine bottleneck_detection(A, U)
Use **LP1** to calculate $\tilde{Y}$.
Use **LP2** to calculate an association $\mathcal{X}$.
Construct a graph $G = (V, E)$.
Color each AP a black if $y_a = \tilde{Y}$.
Color each AP a white if $y_a < \tilde{Y}$.
while exist $(a, b) \in E$ and a is black and b is white **do**
Color AP a white.
end while
$\tilde{L} \leftarrow \{a | a \text{ is colored black}\}$
$\tilde{F} \leftarrow \{u | \exists x_{a,u} > 0 \ \wedge \ a \in \tilde{L}\}$
$\tilde{\mathcal{X}} \leftarrow$ the association of $\tilde{F}$ and $\tilde{L}$.
Return $\{\tilde{L}, \tilde{F}, \tilde{\mathcal{X}}\}$
end

Figure 6: A formal description of the bottleneck-group detection routine.

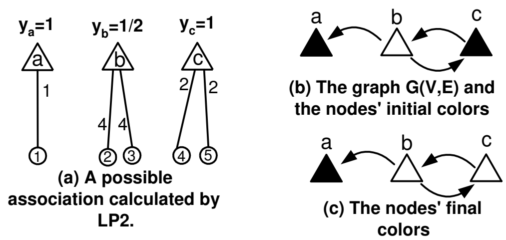


Figure 7: Examples of an execution of the bottleneck-groups detection routine.

Theorem 5 is proven in Appendix B.

4.2.2 The Rounding Method

For the sake of completeness, we provide a short description of the rounding method of Shmoys and Tardos [27]. This description is tailored for unweighted greedy users but with minor modifications it can address other user characteristics, as we explain in the following sections. Consider a fractional association $\mathcal{X}$ and for each AP $a \in A$ let $S_a = \lceil \sum_{u \in U} x_{a,u} \rceil$. Initially, the rounding method constructs a bipartite graph $G'(\mathcal{X}) = (U, V, E)$. Each node u in the set U of the bipartite graph represents a user u in U. The set V contains S_a nodes for each AP $a \in A$ denoted by $\{v_{a,1}, v_{a,2}, \cdots, v_{a,S_a}\}$. The graph edges are determined by the following process. For each AP $a \in A$, the users U_a are sorted according to a given *sorting criterion*. In the case of unweighted greedy users, the users in U_a are sorted in non-decreasing wireless bit rate $r_{a,u}$ and they are renamed according to this order, $\{u_1, u_2, \cdots, u_{|U_a|}\}$. Moreover, let $C(a, u_j) = \sum_{i=1}^{j} x_{a,u_i}$. For each AP a, we divide the users in U_a into S_a groups, denoted by $Q_{a,s}$ where $1 \le s \le S_a$, according to their $C(a, u_j)$ values. Each group $Q_{a,s}$ contains all the users u_j such that $s - 1 < C(a, u_j) \le s$ or $s - 1 \le C(a, u_{j-1}) < s$. A user that is included in two groups is referred as *border node*. The edges E of the graph represent user-AP association. Thus, for each AP a and every integer $s \in S_a$ node $v_{a,s}$ is connected to each user u_j in $Q_{a,s}$. Such bipartite graph is given in Example 9. After constructing the graph G', the rounding method looks for a maximal matching [29] from each user to one of the nodes $v_{a,s} \in V$. Since the association $\mathcal{X}$ specifies a fractional matching such

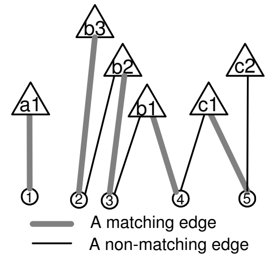


Figure 8: Examples of the graph G' and a matching.

maximal matching exists (more details are provided in [27]) and it determines the integral association of the users.

EXAMPLE 9. Consider the wireless system described in Example 2 and the fractional max-min fair association depicted in Figure 2-(d). In this association $x_{\mathbf{b},4} = x_{\mathbf{c},4} = \frac{1}{2}$ and its NBV is $\vec{B} = \{1, \frac{4}{3}, \frac{4}{3}, \frac{4}{3}, \frac{4}{3}\}$. Figure 8 presents the graph G' calculated by the rounding method and a corresponding matching. Consequently, the obtained load vector $\vec{Y} = \{1, 1, \frac{1}{2}\}$ and the corresponding NBV is $\vec{B} = \{1, 1, 1, 1, 2\}$. The latter is not the optimal max-min fair association. However, the bandwidth of each user u is at least half of its bandwidth in the fraction association. $\square$

4.3 Analysis of the Unweighted Case

We now prove the approximation ratio of our algorithm for the case of unweighted greedy users. We start with a useful property of the rounding method. We assign to each edge e of G' a weight, $x'(e)$, termed the *association weight*, that represents the fractional association of the corresponding user and AP. More specifically, consider an edge $e = (v_{a,s}, u) \in E$ indicating that user u is associated with AP a. If user u is a non-border node then it is included only in the set $Q_{a,s}$ and we assign $x'(v_{a,s}, u) = x_{a,u}$. Otherwise, user u is included in the sets $Q_{a,s-1}$ and $Q_{a,s}$ and we partition the association $x_{a,u}$ with the two edges $(v_{a,s-1}, u)$ and $(v_{a,s}, u)$, such that $x'(v_{a,s}, u) = C(a, u) - s + 1$ and $x'(v_{a,s-1}, u) = x_{a,u} - x'(v_{a,s}, u)$. This assignment ensures the following property.

PROPERTY 2. *Consider an AP $a \in A$ and a set $Q_{a,s}$, where s is an integer between 1 and S_a. Then, for any $s < S_a$, it follows that $\sum_{u \in Q_{a,s}} x'(v_{a,s}, u) = 1$ and $\sum_{u \in Q_{a,S_a}} x'(v_{a,S_a}, u) \le 1$.*

Consider a node $v_{a,s} \in V$. We define its fractional wireless load as $y^{frac,w}(v_{a,s}) = \sum_{u \in Q_{a,s}} x'(v_{a,s}, u)/r_{a,u}$. Moreover, suppose that node $v_{a,s}$ is associated to user $u \in Q_{a,s}$ in the calculated matching. We define its integral wireless load as $y^{int,w}(v_{a,s}) = 1/r_{a,u}$. Similarly, we define the fractional and integral infrastructure load of node $v_{a,s}$ as $y^{frac,i}(v_{a,s}) = \sum_{u \in Q_{a,s}} x'(v_{a,s}, u)/R_a$ and $y^{int,i}(v_{a,s}) = 1/R_{a,u}$. Consequently,

LEMMA 1. *Consider a node $v_{a,s} \in V$ such that $s > 1$. Then, $y^{int,w}(v_{a,s}) \le y^{frac,w}(v_{a,s-1})$ and $y^{int,i}(v_{a,s}) \le y^{frac,i}(v_{a,s-1})$.*

Proof: This lemma results directly from the selected sorting criterion and we first prove it for wireless channel. For each user $u \in Q_{a,s}$, $s > 1$ satisfied that $r_{a,u} \ge r_{a,u'}$ for every user

$u' \in Q_{a,s-1}$. This is also true for the user $u^* \in Q_{a,s}$ that is matched with node $v_{a,s}$. Thus,

$$y^{frac,w}(v_{a,s-1}) = \sum_{u' \in Q_{a,s-1}} \frac{x'(v_{a,s}, u')}{r_{a,u'}} \geq$$

$$\geq \sum_{u' \in Q_{a,s-1}} \frac{x'(v_{a,s}, u')}{r_{a,u^*}} = \frac{1}{r_{a,u^*}} = y^{int,w}(v_{a,s})$$

We now consider the backhaul link. Recall that all the users pose the same load, $1/R_a$, on the backhaul link. Therefore, independent of the user order, for each node $v_{a,s} \in V$ such that $s < S_a$, it follows that $y^{frac,i}(v_{a,s}) = 1/R_a$ and for any node $v_{a,S_a} \in V$, it follows that $y^{frac,i}(v_{a,S_a}) \leq 1/R_a$. Consequently, $y^{int,i}(v_{a,s}) \leq y^{frac,i}(v_{a,s-1})$. □

THEOREM 6. *The association $\mathcal{X}$ calculated by integral load balancing algorithm ensures 2^* max-min fairness approximation with threshold T, defined by Equation 2.*

Proof: First, we prove for each AP $a \in A$ that $y_a^{int} \leq y_a^{frac} + T$. We prove this property for the wireless link. The proof for the backhaul link is similar. From Lemma 1 and the definition of T follows,

$$y_a^{int,w} = \sum_{s \in [1..S_a]} y^{int,w}(v_{a,s}) \leq$$

$$\leq T + \sum_{s \in [1..(S_a-1)]} y^{frac,w}(v_{a,s}) \leq T + y_a^{frac,w}$$

Consequently, $y_a^{int} \leq T + y_a^{frac}$. In the sequel we consider two cases:
Case I: suppose that $y_a^{frac} \geq T$. Thus $y_a^{int} \leq 2 \cdot y_a^{frac}$. From Theorems 3 and 5, it results that bandwidth allocation of each user u associated with AP a in the integral solution is $b_u^{int} = \frac{1}{y_a^{int}} \geq \frac{1}{2 \cdot y_a^{frac}} = \frac{b_u^{frac}}{2}$.
Case II: Suppose that $y_a^{frac} < T$. Thus $y_a^{int} \leq 2 \cdot T$. Accordingly, each user u that is associated with AP a in the integral solution experiences a bandwidth $b_u^{int} = \frac{1}{y_a^{int}} > \frac{1}{2 \cdot T}$, and this complete our proof. □

4.4 Weighted Greedy Users

We turn to describe our integral load balancing algorithm for weighted users. This algorithm is similar to the one described in Section 4.2 with different sorting criterion. We observed that in weighted instances, the calculated fractional solution $\mathcal{X}^{frac}$ does not satisfy Lemma 1. This prevents from us to providing 2^* max-main fairness approximation. However, by using a different sorting criterion, our algorithm ensures 3^* approximation. For our needs, we define the *joined load* of user u on AP a as,

$$J_{a,u} = \frac{x_{a,u} \cdot w_u}{r_{a,u}} + \frac{x_{a,u} \cdot w_u}{R_{a,u}}$$

The joined load may be either fractional or integral. For a given AP a, the algorithm sorts the users U_a in decreasing order of their joined loads, $J_{a,u}$. This order determines the manner in which the users U_a are divided into groups $\{Q_{a,s}\}$. The rest of the rounding method remains the same.

We turn to calculate the approximation ratio of the algorithm with same threshold T defined in Equation 2. Consider a node $v_{a,s} \in V$ we define its fractional joined load $J^{frac}(v_{a,s}) = \sum_{u \in Q_{a,s}} x'(v_{a,s}, u) \cdot J_{a,u}$. Now, suppose that node $v_{a,s}$ is associated to user $u \in Q_{a,s}$ in the integral solution. Thus, its integral joined load is $J^{int}(v_{a,s}) = J_{a,u}$. Note that the fractional and integral joined loads of AP $a \in A$ satisfy,

$$J_a^{frac} = y_a^{frac,w} + y_a^{frac,i} = \sum_{u \in U_a} J_{a,u}^{frac} = \sum_{s=1}^{S_a} J^{frac}(v_{a,s})$$

Similarly,

$$J_a^{int} = y_a^{int,w} + y_a^{int,i} = \sum_{u \in U_a} J_{a,u}^{int} = \sum_{s=1}^{S_a} J^{int}(v_{a,s})$$

LEMMA 2. *Consider a node $v_{a,s} \in V$ such that $s > 1$. Then, $J^{int}(v_{a,s}) \leq J^{frac}(v_{a,s-1})$.*

Proof: This proof is similar to the proof of Lemma 1 and it is direct result from the definition of joined load. □

LEMMA 3. *Consider an AP $a \in A$ then $J_a^{frac} \leq 2 \cdot y_a^{frac}$*

Proof: By definition, $J_a^{frac} = y_a^{frac,w} + y_a^{frac,i} \leq$
$\leq 2 \cdot \max\{y_a^{frac,w}, y_a^{frac,i}\} = 2 \cdot y_a^{frac}$ □

THEOREM 7. *The association $\mathcal{X}$ calculated by integral load balancing algorithm ensures 3^* max-min fairness approximation with threshold T, defined by Equation 2.*

Proof: First, we prove that for each AP $a \in A$ follows that $y_a^{int} \leq 2 \cdot y_a^{frac} + T$. From Lemma 2 and the definition of T, it follows,

$$y_a^{int} = \max\left\{\sum_{s=1}^{S_a} y^{int,w}(v_{a,s}), \sum_{s=1}^{S_a} y^{int,i}(v_{a,s})\right\} \leq$$

$$\leq \sum_{s=1}^{S_a} J^{int}(v_{a,s}) \leq T + \sum_{s=1}^{S_a-1} J^{frac}(v_{a,s}) \leq T + J_a^{frac}$$

From Lemma 3 results that $y_a^{int} \leq T + 2 \cdot y_a^{frac}$. In the sequel we consider two cases:
Case I: Suppose that $y_a^{frac} \geq T$. Thus, $y_a^{int} \leq 3 \cdot y_a^{frac}$. From Theorems 4 and 5, it results that the normalized bandwidth $\bar{b}_u^{int}$ allocated to user u associated with AP a in the integral solution is $\bar{b}_u^{int} = \frac{1}{y_a^{int}} \geq \frac{1}{3 \cdot y_a^{frac}} = \bar{b}_u^{frac}/3$.
Case II: Suppose that $y_a^{frac} < T$. Thus $y_a^{int} \leq 3 \cdot T$. Accordingly, each user u that is associated with AP a in the integral solution experiences a normalized bandwidth $\bar{b}_u^{int} = \frac{1}{y_a^{int}} \geq \frac{1}{3 \cdot T}$, and this complete our proof. □

4.5 Time Fairness

Finally, we show that our scheme finds the optimal integral solution for max-min time fairness. Time fairness attempts to provide a fair service time to the users regardless of the effective bit rates, $r_{a,u}$ and R_a, that they experience. Such fairness is considered, for instance, when the system bottlenecks are the backhaul links and all these links have the same bit rate, R. In such instance, a max-min time fairness solution also guarantees max-min bandwidth fairness. To achieve this goal, we use the scheme presented in Section 4.2 with the following modifications. First, for each user $u \in U$ and AP $a \in A$, we set their effective bit rates $r_{a,u}$ and R_a to 1 and we utilize the unweighted greedy variant for obtaining a fractional solution. Then, after calculating the bipartite graph $G'(\mathcal{X}) = (U, V, E)$, we assigned a cost $c(v_{a,s}, u) = s$ to each edge $(v_{a,s}, u) \in E$. Finally, the integral association is determined by the minimal cost maximal matching [29] of the graph G'.

THEOREM 8. *The time fairness algorithm calculates the optimal max-min time fairness association.*

Proof: From Theorem 5, it follows that our scheme finds the optimal fractional solution. Thus, to complete the proof it is sufficient to prove that the algorithm finds the optimal integral association for every fairness group $F_k \subseteq U$ and its corresponding load group $L_k \subseteq A$ with load y_k of the fractional solution. Clearly, in this case the load of each AP $a \in L_k$ is $y_k = y_a = \sum_{u \in U_a} x_{a,u}$. Thus, from the definition of S_a in Section *4.2.2*, it results that $S_a - 1 < y_a \leq S_a$ for every AP $a \in L_k$. Since all APs in L_k have the same S_a we denote it by S_k and the number of users that are associated with any AP $a \in L_k$ is at most S_k. We consider two cases.
Case I: $y_k = S_k$. Thus, each AP in L_k is associated with exactly S_k users and this guarantees the required time fairness.
Case II: $y_k < S_k$. Consequently, some APs are associated with fewer than S_k users. Note that we are addressing now a load conserving system, *i.e.,* in any possible association of the user in F_k associated with the APs in L_k, the total load on all the APs is $y_k \cdot |L_k| = |F_k|$. Since, our algorithm seeks for minimal cost matching no AP is associated with fewer than $S_k - 1$ users. From this, it results that exactly $(S_k - y_k) \cdot |L_k|$ APs are associated with $S_k - 1$ users and others are associated with S_k users. This is a max-min time fair association and this completes our proof. □

5. ASSOCIATION CONTROL OF BOUNDED DEMAND USERS

We now consider users with bounded bandwidth demands. For our calculations, we first modify the load definition. Then, we use the new definition to construct a new algorithm, termed the *adjusted load balancing algorithm*, that guarantees a 3^* max-min fairness approximation. In the following we denote by d_u the bounded demand of user $u \in U$ and we define its *normalized demand* to be $\bar{d}_u = d_u/w_u$.

5.1 The Adjusted Load Definition

Let us start with an example that illustrates the need for a new load definition.

EXAMPLE 10. Consider the wireless system described in Example 2. We assume that users $1, 2, 3, 4$ are greedy (have very high demands), while the demand of user 5 is $d_5 = 2$. Here, the max-min fair fractional-association is provided in Figure 2-(d) and the allocated bandwidth to user 5 is $\frac{4}{3}$. In this case the system does not satisfy the user bandwidth requirement. Therefore, user 5 behaves as a greedy user and the load that it poses on AP $\mathbf{c}$ is $x_{\mathbf{c},5}/r_{\mathbf{c},5} = \frac{1}{2}$. Now, suppose that $d_5 = \frac{1}{2}$. Clearly, the system satisfies the demand of user 5. However, user 5 consumes only a bandwidth of $\frac{1}{2}$. As a result, the bandwidth that is allocated to user 4 by AP $\mathbf{c}$ is 1.5 and so its overall bandwidth allocation is $2\frac{1}{6}$. The new NBV is $\vec{B} = \{\frac{1}{2}, 1, \frac{4}{3}, \frac{4}{3}, 2\frac{1}{6}\}$, which is obviously not a min-max fair allocation. □

We turn to provide informal description of the new load definition. As an example, consider a bounded demand user u that is associated with a single AP a and let $w_u = 1$. From the previous load definition in Section 3.2, it follows that the load induced by user u on AP a is the time that AP a takes to provide user u one unit of traffic, and the load y_a on AP a is the aggregated load of all its associated users. Thus, in a fair system, the bandwidth allocated to user u is $1/y_a$. However, if $d_u < 1/y_a$ user u receives higher bandwidth than its demand. Conceptually, in a period of y_a time units, user u consumes at most a traffic volume of $d_u \cdot y_a$, and the load of user u on AP a is just the time that takes AP a to provide this volume to user u. So, for the wireless channel, the load that user u induces on AP a, denoted by $y^w_{a,u}$, is $y^w_{a,u} = d_u \cdot y_a / r_{a,u}$ if $d_u > 1/y_a$ and it is $y^w_{a,u} = 1/r_{a,u}$ otherwise. We now provide a formal definition of the *adjusted load* for weighted system and fractional association $\mathcal{X}$. By using similar arguments we conclude that,

$$y^w_{a,u} = \begin{cases} \frac{x_{a,u} \cdot w_u}{r_{u,a}} & : \quad \bar{d}_u \geq \frac{1}{y_a} \\ \frac{x_{a,u} \cdot w_u \cdot \bar{d}_u \cdot y_a}{r_{u,a}} & : \quad \bar{d}_u < \frac{1}{y_a} \end{cases} \quad (3)$$

Similarly, the load posed by user u on the backhaul link of AP a, denoted as $y^i_{a,u}$, is defined as follows:

$$y^i_{a,u} = \begin{cases} \frac{x_{a,u} \cdot w_u}{R_a} & : \quad \bar{d}_u \geq \frac{1}{y_a} \\ \frac{x_{a,u} \cdot w_u \cdot \bar{d}_u \cdot y_a}{R_a} & : \quad \bar{d}_u < \frac{1}{y_a} \end{cases} \quad (4)$$

Consequently, the load of an AP $a \in A$ is,

$$y_a = \max \left\{ \sum_{u \in U} y^w_{a,u}, \sum_{u \in U} y^i_{a,u} \right\} \quad (5)$$

From the adjusted load definition, it results that as long as a user demand is satisfied the load that it poses on its associated AP decreases as the total load of this AP decreases. In Example 11, we illustrate that this definition can be used to obtain max-min fair bandwidth allocation.

EXAMPLE 11. Consider the same settings as described in Example 10, Where $d_5 = \frac{1}{2}$. Now, assume that the load on APs $\mathbf{b}$ and $\mathbf{c}$ is $y_{\mathbf{b}} = y_{\mathbf{c}} = \frac{4}{7}$ and let the association of user 4 be $x_{\mathbf{b},4} = \frac{1}{7}$ and $x_{\mathbf{c},4} = \frac{6}{7}$. We show that these assignments satisfy Equations 3-5. Let us start with AP $\mathbf{b}$: $y_{\mathbf{b}} = \frac{1}{4} + \frac{1}{4} + \frac{1}{7} \cdot \frac{1}{2} = \frac{8}{14} = \frac{4}{7}$. The load of AP $\mathbf{c}$ is: $y_{\mathbf{c}} = \frac{6}{7} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{4}{7} \cdot \frac{1}{2} = \frac{4}{7}$. From this, we know that the bandwidth allocation to users $2, 3$ and 4 is 1.75 and the corresponding NBV is $\vec{B} = \{\frac{1}{2}, 1, \frac{7}{4}, \frac{7}{4}, \frac{7}{4}\}$. Since the load on APs $\mathbf{b}$ and $\mathbf{c}$ is balanced, this is a min-max load balanced association. □

Note that, with the adjusted load definition, Theorems 2 and 4 still hold[4]. This means that any fractional min-max load balanced association $\mathcal{X}$ for bounded demand users divides the APs into load groups $\{L_k\}$, such that the load on the APs in each group L_k is y_k and each group L_k is uniquely associated with a set of users F_k, termed a fairness group. This association also obtains a max-min fair bandwidth allocation. However, unlike the greedy user model, a fairness group F_k may include users with allocated normalized bandwidth less than $1/y_k$. These users, of course, are bounded-demand users with normalized demands $\bar{d}_u < 1/y_k$. Consequently, we define the users in a given fairness group F_k as the users served by the APs of a given load group L_k, regardless of their allocated bandwidth.

[4]The proofs of these theorems for the bounded demand user model are similar to the case of greedy user and their details are omitted from this document.

5.2 The Adjusted Load Balancing Algorithm

We now present our method to calculate a max min fair association $\mathcal{X}$ in the presence of bounded-demand users. Since Theorems 2 and 4 are satisfied in this case, we would like to extend our scheme to support also bounded demand users. The main challenge that we are facing is calculating the load of the load groups $\{L_k\}$ induced by the optimal fractional solution, in particular, how to detect the bottleneck load value $\tilde{Y}$ in the bottleneck-group detection routine in Section 4. This task cannot be easily resolved, since the adjusted load definitions yield a complicated set of non-linear equations with feedback. To simplify the proof, we would like all the users to behave as if they are greedy ones. This goal can be achieved by assigning virtual channel conditions to bounded demand users appropriately, such that for each user u with $d_u/w_u < 1/\tilde{Y}$ its demand will be satisfied only when the load of its associated APs is exactly $\tilde{Y}$. For that purpose, we estimate the bottleneck load value and we denote this estimation by Y. We use Y to define adjusted bit rates $r'_{a,u}$ and $R'_{a,u}$ of both the wireless and the wired links for each pair of AP a and user u. Consequently, for user u with maximal bandwidth requirement[5], d_u, its *adjusted bit rates* are:

$$r'_{a,u} = \begin{cases} r_{u,a} & : \ \bar{d}_u \geq \frac{1}{Y} \\ \frac{r_{u,a}}{\bar{d}_u \cdot Y} & : \ \bar{d}_u < \frac{1}{Y} \end{cases} \tag{6}$$

$$R'_{a,u} = \begin{cases} R_a & : \ \bar{d}_u \geq \frac{1}{Y} \\ \frac{R_a}{\bar{d}_u \cdot Y} & : \ \bar{d}_u < \frac{1}{Y} \end{cases} \tag{7}$$

Now, we employ **LP1** to verify our estimation whether $Y = \tilde{Y}$, using the adjusted bit rates $r'_{a,u}$ and $R'_{a,u}$. Thus, by performing a binary search, we accurately calculate the value of $\tilde{Y}$. After $\tilde{Y}$ is detected, the second and third steps of the bottleneck-group detection routine can be used to infer the bottleneck groups $\tilde{L}$ and $\tilde{F}$ and iteratively calculate the entire optimal fractional solution. Finally, we invoke the rounding algorithm of the weighted greedy Users in Section 4.4 to obtain an integral solution. In our calculation we utilize the adjusted bit rates of the users, as calculated by the fractional solution, to construct the sets $Q_{a,s}$.

5.3 The Algorithm Analysis

Consider a calculated fractional solution $\mathcal{X}$. We initially calculate the actual bandwidth allocated to each user in this solution.

THEOREM 9. *Let $\mathcal{X}$ be the calculated fractional association, when using the adjusted bit rates, and let $\{L_k\}$ and $\{F_k\}$ be the corresponding load and fairness groups. Consider a user $u \in F_k$ with bounded demand d_u. Then, the bandwidth allocated for user u is $b_u = \frac{w_u}{y_k}$ if $\frac{d_u}{w_u} \geq \frac{1}{y_k}$. Otherwise, its allocated bandwidth is $b_u = d_u$*

Proof: Recall that all the APs in F_k has the same load y_k. Thus, if $\frac{d_u}{w_u} \geq \frac{1}{y_k}$, it follows that $r'_{a,u} = r_{a,u}$ and $R'_{a,u} = R_a$ for each $a \in F_k$. Since we treat the users as greedy we can utilize Theorem 3 and conclude that $b_u = w_u/y_k$.
Otherwise, if $\frac{d_u}{w_u} < \frac{1}{y_k}$, it follows that for each $a \in F_k$, $r'_{a,u} = \frac{r_{u,a}}{d_u \cdot y_k}$ and $R'_{a,u} = \frac{R_a}{d_u \cdot y_k}$. For simplicity, we calculate b_u over wireless domain, the calculation for the wired domain is very

Alg Online_Load_Balancing(A, U, u)
  if (elapsed time from last offline optimization $> \tau$) **then**
    Integral_Load_Balancing(A, $U \cup u$)
    $y_{offline} \leftarrow \max_{a \in A} y_a$
  else
    $a \leftarrow$ AlgorithmByAAFPW(A, u)
    $y_{online} \leftarrow \max_{a \in A} y_a$
    if ($y_{online} - y_{offline} > \Delta$) **then**
      Integral_Load_Balancing(A, $U \cup u$)
    else
      assign u to AP a
end

Figure 9: A formal description of the online load balancing algorithm

similar. Let $t_{a,u}$ be the time that user u is served by AP a during a single time unit. Thus, $t_{a,u} = \frac{x_{a,u} \cdot w_u}{y_k \cdot r'_{a,u}}$. As a result,

$$b_u = \sum_{a \in A_u} t_{a,u} \cdot r_{a,u} = \sum_{a \in A_u} \frac{x_{a,u} \cdot w_u}{y_k \cdot r'_{a,u}} \cdot r_{a,u} =$$

$$= \sum_{a \in A_u} \frac{x_{a,u} \cdot w_u}{y_k \cdot \frac{r_{u,a} w_u}{d_u \cdot y_k}} \cdot r_{a,u} = \sum_{a \in A_u} x_{a,u} \cdot d_u = d_u$$

This completes our proof. □

By similar arguments of the proof of Theorem 5 we conclude,

THEOREM 10. *The adjusted load balancing algorithm calculates a min-max load balanced association for the fractional-association model with bounded demand users. This association also ensures max-min fair bandwidth allocation.*

In our analysis, we provide guarantees on the quality of the integral solution as long as the bounded demand users behave as greedy users. Therefore, we define our threshold T' to be,

$$T' = \max_{\{u,a | u \in U \wedge a \in A \wedge r_{a,u} > 0\}} \max\{\frac{w_u}{r_{a,u}}, \frac{w_u}{R_a}, \frac{1}{\bar{d}_u}\} \tag{8}$$

For this threshold we can use similar proof as the one of Theorem 7 to prove Theorem 11.

THEOREM 11. *The association $\mathcal{X}$ calculated by adjusted load balancing algorithm ensures 3^* max-min fairness approximation with threshold T', defined by Equation 8.*

6. ONLINE INTEGRAL-ASSOCIATION

In this section, we present an algorithm that deals with dynamic user arrivals and departures. Clearly, a repeated execution of the offline algorithm each time a user arrives or departs may cause frequent association changes that disrupt existing sessions. To avoid this, we propose a strategy that enables us to strike a balance between the frequency of the association changes and the optimality of the network operation in terms of load balancing. For this propose we use two configuration parameters; *time threshold*, τ, and *load threshold* Δ. We rerun our offline algorithm if either of the following two conditions hold.
(1) The time elapsed since our last offline optimization is more than the time threshold τ.
(2) The current bottleneck load, *i.e.,* , the maximal load among all APs, is Δ more than the bottleneck load obtained by the last execution of the offline algorithm.

[5]For greedy users let us denote $d_u = \infty$.

After rerunning the algorithm, each user who needs to change association can be done between its session arrivals to avoid disruption of its ongoing sessions. Our algorithm is illustrated in Figure 9.

Between two offline optimization occurrences, we need to associate users to APs as they arrive. We adapt the online algorithm of Aspnes *et al.* , in [21], to achieve a $O(logn)$ approximation factor as compared to the offline optimal, where n is number of users in the system. We refer their algorithm as AlgorithmByAAFPW. All we need to change is to substitute the load in their algorithm by the integral load of the APs, y_a^{int}. In online user association, we need to address two conflicting factors. Intuitively, a user should be assigned to the less loaded APs that are within its transmission range. However, the data rate from the user to these APs can be very low which adds very high additional load to them. Therefore, a user should be assigned to an AP where it causes small additional load. To capture these two trade-offs, Aspnes et al. [21] define a potential function that is exponential in the load of an AP. When a new user arrives, all possible user-AP association are evaluated. After the evaluation, the assignment that minimizes the increase of the potential function is selected. They show that, using certain potential functions, the highest load among all APs of the online algorithm is within $O(logn)$ factor of the highest load among all APs of the offline algorithm.

7. SIMULATION RESULTS

In our simulation compared the performance (in the context of max-min fairness) of our scheme with two popular heuristics, namely the Strongest-Signal-First(SSF) method and the Least-Loaded-First(LLF) method. The SSF method is the default user-AP association method in the 802.11 standard. The LLF method is a widely-used load-balancing heuristic, in which a user chooses the least-loaded AP that he can reach. For a fair comparison, we assume the same scheduling mechanism at the APs for all three methods, such that the only difference is the assignment decisions between users and APs. The simulation setting is as follows. We use a simple wireless channel model in which the user bit rate depends only on the distance to the AP. Adopting the values commonly advertised by 802.11b vendors, we assume that the bit rate of users within 50 meters from AP is 11 Mbps, 5.5 Mbps within 80 meters, 2 Mbps within 120 meters, and 1 Mbps within 150 meters, respectively. The maximum transmission range of an AP is 150 meters. The backhaul capacity is set to 10 Mbps to emulate the Ethernet infrastructure. A total of 20 APs are located on a 5 by 4 grid, where the distance between two adjacent APs is set to 100 meters and we assume that an appropriate frequency planning was made. The number of users is either 100 to simulate a moderately loaded network or 250 to simulate a heavily loaded network. We assume all users are greedy.

Due to space limitation we present our results only for the case hot-spots that more common in practical WLANs. We locate all users in a circle-shape hot spot at the center of the network. The radius of the hot spot is set to 150 meters. Even if the size of the hot spot is the same as that of one 802.11 cell, the users still can reach several cells because of the overlap between cells. Figures 10 and 11 show the results with 100 and 250 users, respectively. The Y axis represents the per-user bandwidth and the X axis represents the user index. Note that the users are sorted by their bandwidth in increasing order. The user locations are different at each

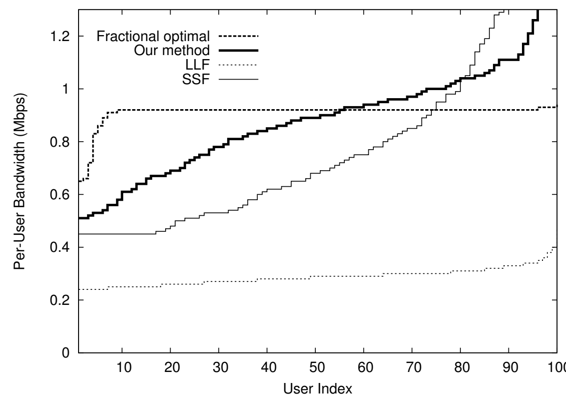


Figure 10: Per-user bandwidth of 100 users.

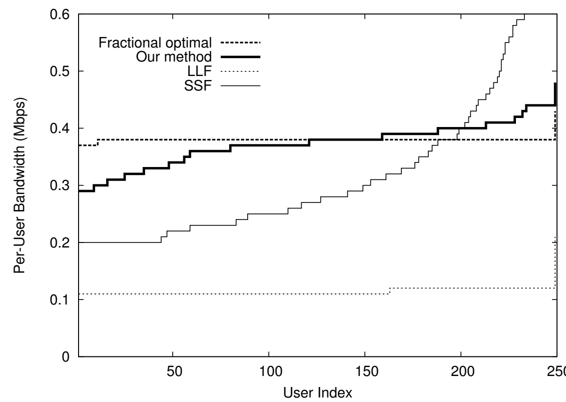


Figure 11: Per-user bandwidth of 250 users.

run, and therefore the bandwidth of the user with the same x index actually indicates the average bandwidth of x-th lowest bandwidth user. Somewhat surprisingly, our method outperforms the two heuristics not only in terms of fairness but also in terms of total system throughput. For instance, in Figure 10, the median per-user bandwidth value of our method is over 20% higher than that of the SSF method. The bandwidth values are obtained by averaging the results of 100 simulation runs. We also noticed that the SSF approach outperforms the LLF method in terms of both max-min fairness and overall network throughput. This supports our claim above that a naive load-balancing algorithm may yield very poor results. By comparing Figure 10 and 11, we also conclude the gap between our method and the fractional optimal solution narrows as the number of users increases. It can be explained by the fact that the impact of each user in the integral association scheme decreases as the number of users increases. Thus, with an infinite number of users, the results of integral association and fractional association will converge.

We also simulated the online algorithm. To simulate the dynamic user departure/arrival (or the user mobility), at each time slot a certain portion of users are taken out of the system and the same number of new users are injected into the system. The result of the case that we replace 20% of users at every time slot is shown in Figure 12. Unlike other plots the Y axis represents the lowest user bandwidth

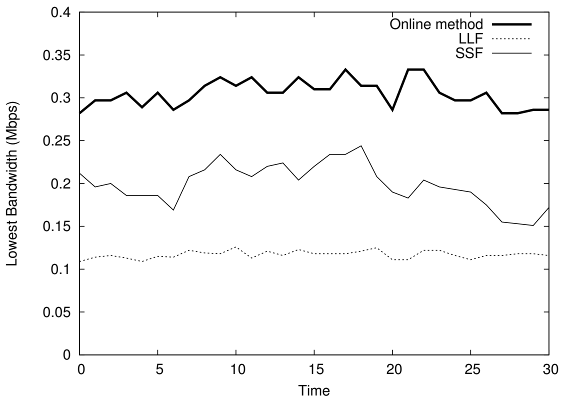


Figure 12: Simulation result of the online case with 250 users.

and the X axis represents the time. The offline algorithm is periodically invoked at every 15 time slots or when the bottleneck difference exceeds 25%. Note that the result is episodic, since it depicts the evolution of the system for one simulation run. Nevertheless, the presented result is very typical. In the presented case, the offline algorithm was invoked 5 times including 20th time slot.

8. CONCLUSION

As wireless LANs are deployed to cover larger areas and are increasingly more and more relied on to carry important tasks, it is essential that they be managed in order to achieve desired system performance objectives. In this paper, we study one of the problems—providing fair service to users and balancing the load among APs. These goals are obtained by intelligently determining the user-AP association, termed association control. We first rigorously formulate this association control problem in the context of wireless LANs and we present approximation algorithms that provide guarantees on the quality of the solution. Our simulations confirm that the proposed methods, indeed, achieve close to optimal load balancing and max-min fair bandwidth allocation, and significantly outperform popular heuristics. Moreover, we show that in some cases, by balancing the load on the APs the overall network throughput is increased. In the future, we intend to develop a practical management system based on the theoretical foundation presented in this study.

9. REFERENCES

[1] A. Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan. Characterizing user behavior and network performance in a public wireless LAN. In *Proc. of ACM SIGMETRICS*, pages 195–205, 2002.

[2] D. Kotz and K. Essien. Analysis of a campus-wide wireless network. In *Proc. ACM MobiCom*, pages 107–118, 2002.

[3] M. Balazinska and P. Castro. Characterizing mobility and network usage in a corporate wireless local-area network. In *Proc. USENIX MobiSys*, 2003.

[4] I. Papanikos and M. Logothetis. A study on dynamic load balance for IEEE 802.11b wireless LAN. In *Proc. COMCON*, 2001.

[5] A. Balachandran, P. Bahl, and G. M. Voelker. Hot-spot congestion relief and service guarantees in public-area wireless networks. *SIGCOMM Comput. Commun. Rev.*, 32(1):59–59, 2002.

[6] T-C. Tsai and C-F. Lien. IEEE 802.11 hot spot load balance and QoS-maintained seamless roaming. In *Proc. National Computer Symposium (NCS)*, 2003.

[7] Proxim Wireless Networks. ORINOCO AP-600 data sheet, 2004.

[8] Cisco Systems Inc. Data sheet for cisco aironet 1200 series, 2004.

[9] I. Katzela and M. Nagshineh. Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey. *IEEE Personal Communications*, pages 10–31, 1996.

[10] S. Das, H. Viswanathan, and G. Rittenhouse. Dynamic load balancing through coordinated scheduling in packet data systems. In *Proc. IEEE INFOCOM*, 2003.

[11] B. Eklundh. Channel utilization and blocking probability in a cellular mobile telephone system with directed retry. *IEEE Trans. on Communications*, 34(4):329–337, 1986.

[12] T. P. Chu and S. R. Rappaport. Overlapping coverage with reuse partitioning in cellular communication systems. *IEEE Trans. on Vehicular Technology*, 46(1):41–54, 1997.

[13] X. Lagrange and B. Jabbari. Fairness in wireless microcellular networks. *IEEE Trans. on Vehicular Technology*, 47(2):472–479, 1998.

[14] I. Tinnirello and G. Bianchi. A simulation study of load balancing algorithms in cellular packet networks. In *Proc. ACM/IEEE MSWiM*, pages 73–78, 2001.

[15] J. M. Jaffe. Bottleneck flow control. *IEEE Trans. on Communications*, 29:954–962, 1981.

[16] Y. Afek, Y. Mansour, and Z. Ostfeld. Convergence complexity of optimistic rate based flow control algorithms. In *Proc. ACM STOC*, pages 89–98, 1996.

[17] Dimitri P. Bertsekas and Robert Gallager. *Data Networks (2nd Edition)*. Prentice Hall, 1991.

[18] N. Megiddo. Optimal flows in networks with multiple sources and sinks. *Mathematical Programming*, 7:97–107, 1974.

[19] J. M. Kleinberg, Y. Rabani, and E. Tardos. Fairness in routing and load balancing. In *Proc. IEEE FOCS*, pages 568–578, 1999.

[20] J. K. Lenstra, D. B. Shmoys, and E. Tardos. Approximation algorithms for scheduling unrelated parallel machines. *Mathematical Programming*, 46:259–271, 1990.

[21] J. Aspnes, Y. Azar, A. Fiat, S. Plotkin, and O. Waarts. On-line load balancing with applications to machine scheduling and virtual circuit routing. In *Proc. ACM STOC*, pages 623–631, 1993.

[22] A. Goel, A. Meyerson, and S. Plotkin. Approximate majorization and fair online load balancing. In *Proc. SODA*, pages 384–390. Society for Industrial and Applied Mathematics, 2001.

[23] Q. Ni, L. Romdhani, T. Turletti, and I. Aad. QoS issues and enhancements for IEEE 802.11 wireless LAN. Technical Report RR-461, INRIA, France, November 2002. URL: http://www.inria.fr/rrrt/rr-4612.html.

[24] P. Ramanathan and P. Agrawal. Adapting packet fair queueing algorithms to wireless networks. In *Proc. ACM MobiCom*, pages 1–9, October 1998.

[25] S. Lu, T. Nandagopal, and V. Bharghavan. A wireless fair service algorithm for packet cellular networks. In *Proc. ACM MobiCom*, pages 10–20, October 1998.

[26] M. Buddhikot, G. Chandranmenon, S-J. Han, Y-W. Lee, S. Miller, and L. Salgarelli. Integration of 802.11 and third-generation wireless data networks. In *Proc. IEEE INFOCOM*, March 2003.

[27] David B. Shmoys and Eva Tardos. An approximation algorithm for the generalized assignment problem. *Math. Program.*, 62(3):461–474, 1993.

[28] V. Vazirani. *Approximation Algorithms*. Springer-Verlag New York, Incorporated, 1999.

[29] L. Lovazs and M. D. Plummer. *Matching Theory*. North Holland - Amsterdam, 1986.

APPENDIX

A. PROOF SKETCH OF THEOREM 4

In the following we only prove that the min-max load balanced association determines a max-min fair bandwidth allocation. By similar arguments the other direction can be proven as well. Let $\mathcal{X}$ be a min-max load balanced association and let $\vec{B}$ be its normalized bandwidth vector. Lets assume, that $\mathcal{X}$ does not produce a max-min fair bandwidth allocation. Thus, there is an association $\mathcal{X}'$ that its normalized bandwidth vector $\vec{B}'$ has higher lexicographical value than $\vec{B}$. Let $\{F_k\}$, $\{F'_k\}$, $\{L_k\}$ and $\{L'_k\}$ be the fairness and the load groups of the associations $\mathcal{X}$ and $\mathcal{X}'$, respectively. We define an additional association, $\tilde{\mathcal{X}} = (\mathcal{X}+\mathcal{X}')/2$, *i.e.,* for each AP a and user u, it follows $\tilde{x}_{a,u} = (x_{a,u} + x'_{a,u})/2$, and let $\{\tilde{F}_k\}$ and $\{\tilde{L}_k\}$ be its fairness and load groups, respectively. Let j be the lowest index such that $F_j \neq F'_j$ or $L_j \neq L'_j$. Recall, that for every index $i < j$ follows that $\tilde{F}_i = F'_i$ and $\tilde{\bar{b}}_i = \bar{b}'_i$. Since, $\mathcal{X}$ is min-max load balanced association, it follows that $y_j \leq y'_j$. Similarly, $\mathcal{X}'$ is max-min fair bandwidth association, thus, $\bar{b}_j \leq \bar{b}'_j$. As $y_j = \frac{1}{\bar{b}_j}$ and $y'_j = \frac{1}{\bar{b}'_j}$ we have $y_j = y'_j$ and $\bar{b}_j = \bar{b}'_j$. In the following we assume, without lost of generality, that $F_j \neq F'_j$, the case where $L_j \neq L'_j$ can be proven in similar way. We consider three cases:

case I: $F_j \subset F'_j$: However, this contradicts the assumption $\mathcal{X}'$ is a max-min fair bandwidth association.

case II: $F'_j \subset F_j$: Now suppose that $L_j \subset L'_j$, but in this case the set of APs L_j is sufficient to provide the bandwidth $\bar{b}'_j$ to all the users in the set F'_j. While, APs in the sets $L'_j - L_j$ can be used to increase the bandwidth allocation of other users with the same or higher bandwidth, which contradicts the assumption that $\mathcal{X}'$ is max-min fair bandwidth association. Consequently, it follows that $L_j \not\subset L'_j$, which implies that $L_j - L'_j \neq \emptyset$. Thus, the association $\tilde{\mathcal{X}}$, obviously, reduces the load from every AP $a \in L_j - L'_j$, without increasing the load of any AP with load y_j or more. This contradicts the assumption that $\mathcal{X}$ is a min-max load balanced association.

case III: $F'_j - F_j \neq \emptyset$: In this case, the association $\tilde{\mathcal{X}}$ guarantees to each user $u \in F'_j - F_j$ a bandwidth $\tilde{\bar{b}}_u > \bar{b}_j$ without decreasing the bandwidth of any other user that has normalized bandwidth of $\bar{b}_j$ or less in $\mathcal{X}'$. This contradicts the assumption that $\mathcal{X}'$ is a max-min fair bandwidth association.

Consequently, we conclude that for every j, $L_j = L'_j$ and $F_j = F'_j$ and this complete our proof. □

B. THE CORRECTNESS OF THEOREM 5

We start with some properties of the bottleneck-group detection routine. We then prove the correctness of the load balancing algorithm.

LEMMA 4. **LP1** *infers the value of the bottleneck load $\tilde{Y}$ of any min-max load balanced association. Moreover, it calculates an association such that $\tilde{Y}$ upper bounds the load of each AP.*

Proof: **LP1** seeks for an association $\mathcal{X}$ that minimizes $\tilde{Y}$. The first and second conditions verify that $\tilde{Y}$ upper bounds the load of each AP both over the wireless and wired domain. While, the third and fourth condition ensure the $\mathcal{X}$ is a feasible association. □

LEMMA 5. *Let $\mathcal{X}$ be the association calculated by* **LP2** *for a giving bottleneck load value $\tilde{Y}$ as determined by* **LP1**. *The bottleneck load group comprises all the APs if and only if the load on each AP is $\tilde{Y}$.*

Proof: From Lemma 4, it follows that the bottleneck load value is $\tilde{Y}$. Recall that **LP2** finds a feasible association $\mathcal{X}$ that minimizes the overall load with the constraint that the load of each AP is at most $\tilde{Y}$ (the latter is termed as the upper bound constraint). Consequently, if all the APs are included in $\tilde{L}$, then, by definition, the overall load of any such association calculated by **LP2** is $|A| \cdot \tilde{Y}$. Thus, there is no feasible association that satisfies the upper bound constraint and some APs have load strictly less then $\tilde{Y}$. On the other hand, if not all the APs are included in $\tilde{L}$, then there is an association whose overall load is strictly less than $|A| \cdot \tilde{Y}$. In such cases, **LP2** finds a feasible association such that the load of some APs is strictly less than $\tilde{Y}$. □

LEMMA 6. *Let $G = (V, E)$ be the graph that results from the association $\mathcal{X}$ calculated by* **LP2** *and consider the initial node colors. A given AP is included in $\tilde{L}$ if and only if its corresponding node in G, denoted by b, is colored black and there is no directed path in G from b to any white colored node.*

Proof: consider a black node b that is included in a directed path of black nodes $P = \{b = v_1, v_2, \cdots, v_k = a\}$ ended with a white node a. This means that the corresponding AP of node v_{k-1} can shift some load to AP represented by node a. Therefore, it can reduce its load without increasing the load of any AP with load $\tilde{Y}$. In an iterative manner, this process can be done for any node $v_i \in P$. Thus, the AP represented by node b will not be included in $\tilde{L}$.

We now prove the other direction. From Corollary 1, it follows that all the APs in $\tilde{F}$ have load $\tilde{Y}$, hence their corresponding nodes are colored black. In addition, the load of any AP $b \in \tilde{F}$ cannot be reduced by shifting some load to a non-bottleneck AP. Thus, there is no directed link in G between a node representing a bottleneck AP to a node representing a non-bottleneck AP. Consequently, nodes that represent APs in $\tilde{F}$ are not included in any directed path ending with a white node. □

LEMMA 7. *The bottleneck-group detection routine determines the load and the fairness bottleneck groups, $\tilde{L}$ and $\tilde{F}$, and their corresponding user-AP association in the fractional-association model.*

Proof: From Lemma 4, it follows that **LP1** determines the bottleneck load value $\tilde{Y}$ and also calculates a feasible association that satisfies the upper bound constraint. From Lemmas 5 and 6, it follows that the routine separates the APs in $\tilde{L}$ from the other APs. Finally, from Corollary 1 the APs in $\mathcal{X}$ are associated only with the users in $\tilde{F}$. □

Proof of Theorem 5: From Lemma 7 and Corollary 1 results that at each iteration the load balancing algorithm detects the current load and fairness bottleneck groups, denoted as L_k and F_k, and their user-AP association. Thus, at each iteration, the algorithm reduces the size of the AP and user sets until a complete min-max load association is detected. □