EUROPEAN TRANSACTIONS ON TELECOMMUNICATIONS
Eur. Trans. Telecomms. 2009; **20**:494–507
Published online 7 October 2008 in Wiley InterScience
(www.interscience.wiley.com). DOI: 10.1002/ett.1335

Mobile Networks

Client-driven load balancing through association control in IEEE 802.11 WLANs

Eduard Garcia*, Josep L. Ferrer, Elena Lopez-Aguilera, Rafael Vidal and Josep Paradells

Wireless Networks Group, Entel Department, Technical University of Catalonia (UPC), Avda. del Canal Olímpic, 15, 08860 Castelldefels, Spain

SUMMARY

The growth of IEEE 802.11 wireless local area networks (WLANs) (Wi-Fi) brings new possibilities of getting connected in public spaces, known as Hot Spots. Current client-access point associations are an interesting research topic because in these scenarios, users tend to be ‘gregarious’ and essentially static. Under IEEE 802.11 standards, association and roaming decisions are made by client devices and most implementations are based on signal strength measurements; i.e. a client station selects the access point (AP) that provides the strongest signal, which leads to an uneven distribution of clients and load between neighbouring APs. As it can be observed in practical scenarios, the default AP-client association scheme followed in IEEE WLANs, produces unfair situations. This work provides means to effectively alleviate this performance issue and also gives details for a feasible implementation. In this paper we analyse how new IEEE 802.11 standards will allow new radio measurements to provide more efficient association decisions. We propose a new load metric that will produce client-driven associations that ensure greater fairness and throughput. Copyright © 2008 John Wiley & Sons, Ltd.

1. INTRODUCTION

Wireless local area networks (WLANs), especially those that are based on IEEE 802.11 standards operating in infrastructure mode using the distributed coordination function (DCF), are the most popular technologies used to provide broadband radio access to IP networks, whether in order to extend home networks LANs or to provide Internet access in public places (such as airports, hotels and libraries).

Various studies have shown that user service demands are highly dynamic and vary for each type of network, although certain long-term patterns can be identified [1, 2]. These studies conclude that users tend to be concentrated both temporally and spatially, creating highly congested areas known as Hot Spots. For example, the volume of wireless Internet access in a convention centre increases during coffee-breaks at the access points (APs) located close to the area where the coffee-break takes place. Therefore, the load is unevenly distributed across a small number of APs in the WLAN. Moreover, although mobility is increasing as users get into the habit of using wireless access, the mobility pattern can still be considered quasi-static [3] in the sense that users tend to remain in the same location for long periods.

This situation is compounded by the fact that the associations are determined by the client devices on the basis of signal level measurements, which means that users are generally associated with the closest AP. In other words, although a Hot Spot is served by several APs, most of the users will be connected through the AP that provides the strongest signal. Note that the maximum traffic carried by an IEEE 802.11 AP is limited and that this limit decreases as the number of active users increases due to the higher collision probability.

Following these considerations, we argue that by implementing different load balancing policies it will be possi-

* Correspondence to: Eduard Garcia, Wireless Networks Group, Entel Department, Technical University of Catalonia (UPC), Avda. del Canal Olímpic, 15, 08860 Castelldefels, Spain. E-mail: eduardg@entel.upc.edu

Received 12 June 2007
Revised 30 May 2008
Accepted 8 September 2008

ble to improve network efficiency and increase network ability to satisfy quality of service (QoS) requirements. Our proposal has previously been applied to cellular networks and makes use of the overlapping areas between neighbouring cells, i.e. areas under the coverage of more than one base station (BS). Mobile stations in overlapping areas are able to reach several BSs and when a service is required the system can decide to offer it through the BS with more available resources.

In this paper we propose a new client-driven load balancing scheme based on smart AP selection. The method uses the information available to client stations through the mechanisms and measurements provided by the new IEEE standards: 802.11e, 802.11h and, principally, 802.11k. To the best of our knowledge this is the first practical application based on these new 802.11 functionalities and the first time that some of the new radio measurements have been evaluated. We define an algorithm that can be implemented in APs to provide a new load metric that is sent to requesting stations: the available admission capacity (AAC), which is based on a more precise definition of an existing 11e field. We present an analytical study of the proposed metric and use the simulation results to compare it with other solutions.

The rest of the paper is structured as follows: Section 2 provides an overview of related work, which includes different load balancing techniques applied in IEEE 802.11 WLANs and different metrics used to measure load; Section 3 provides a more in-depth analysis of new load metrics available in IEEE 802.11e and 802.11k enabled networks; in Section 4 we propose a new load metric based on the available capacity calculated in the APs; Section 5 gives the details of the implementation; Section 6 contains an evaluation of our proposal in comparison with other schemes; finally, conclusions are given in Section 7.

2. RELATED WORK

2.1. Load balancing techniques

Different approaches have been proposed in the literature that try to change the client-driven nature of IEEE 802.11 association and roaming decisions. The authors of References [4] and [5] propose network-controlled schemes in which client stations send the required information to a central unit, which also has access to the load information for each cell. The scheme proposed in Reference [4] provides the best AP for association and the network also suggests roaming to APs located further away if nearby APs are considered unable to cover the station's requirements. In order to implement these solutions it is necessary to modify the client devices: firstly, they have to send new management frames before they are actually associated; secondly, they will no longer be responsible for association or roaming decisions. The first issue can be solved by using new radio measurements (future IEEE 802.11k). There is no standardised procedure for solving the second issue as yet, but it is expected to be revised by the IEEE 802.21 group, which will provide new mechanisms intended to assist handovers, and by IEEE 802.11v, which is in the development stage and will include management capabilities to allow network-directed roaming.

It is not vital to solve the second of these issues, since it is also possible to perform implicit admission control/association management. This involves actions taken on the network side that induce the desired client behaviour and therefore leave the roaming and association decisions to client stations so that hardware/software modifications are not required. In Reference [6] the APs accept or deny new association requests depending on the respective load. When the first choice is rejected, the stations will send association requests to the next AP in the signal strength-arranged list, until they are admitted. The algorithm proposed in Reference [7] is more sophisticated but follows similar logic. There are three possible AP states: under-loaded (will accept any request), balanced (will not accept extra load) and over-loaded (will expel the station on the assumption that it will automatically request a less loaded neighbouring AP). While these techniques provide good admission control and load balancing, they do not guarantee that all users with network access rights will be fairly served.

Cell breathing techniques consist in dynamically modifying cell dimensions by increasing or reducing transmit power. Cell breathing is a side effect in CDMA networks that reduces the cell coverage when more users are supported, but this could be advantageous in load balancing techniques if optimal strategies are applied. The concept of cell breathing for load balancing in WLANs is explained in Reference [8]: a highly congested AP reduces its coverage radius so that the furthest stations lose connectivity and try to roam to less loaded APs. An under-utilised AP may increase its transmit power in order to expand its coverage. Consequently, new users will roam to this AP and the load on neighbouring APs will decrease. Reference [9] provides an in-depth analysis of cell breathing in IEEE 802.11 WLANs. However, as stated in Reference [10], the furthest stations may sometimes be expelled arbitrarily as they may contribute an insignificant load depending on the applications they run.

DOI:10.1002/ett

Reference [10] also gives an overview of different load balancing techniques that can be applied by using future IEEE 802.11k measurements and statistics.

Hereafter we will focus on client-driven load balancing (based on client decisions). If the stations were able to gather more information they would be able to perform smarter associations and therefore maximise the network efficiency. In Reference [11], clients perform different tests on all APs within range in order to determine the most suitable association. A preferable and less intrusive solution would be to introduce a trade-off between the received signal strength and cell utilisation. Client devices must choose an AP that is close enough to allow frame exchange with the minimum received signal quality that guarantees correct transmission at the highest possible bit rate. However, it is advisable to avoid cell saturation by selecting the AP with the lowest load. In Reference [12], the Probe Response messages sent by the APs provide the number of associated stations and information about the received signal from the requesting station. Clients use this information to calculate a weight for all reachable APs and then associate with the AP with lowest weight. Different commercial products (e.g. References [13] and [14]) use similar systems in which the APs announce their utilisation in beacon frames. However, as with all proprietary solutions, there are interoperability problems that could be solved by the forthcoming IEEE 802.11k standard.

2.2. Load metrics

Load balancing in overlapping areas has traditionally been used in circuit-switched cellular networks. Since each user in these types of networks represents an identical utilisation of available resources, load balancing could be applied by using call level information, i.e. a load balancing scheme will try to balance the number of active calls between neighbouring BSs.

Nevertheless, call level information is not sufficient for modelling the actual load that is carried by a BS in current wireless packet networks, given that users may have different traffic profiles. This assertion is valid for IEEE 802.11 WLANs and is corroborated by different empirical studies [1–3] which state that a new metric based on packet level information is required. However, the number of active users still provides valuable information in networks that use CSMA-based access: more collisions occur as the number of active users increases, which leads to decreased performance. The number of competing stations can be calculated from any station by using the formulation given in Reference [15], but this parameter can only be used to estimate the saturation throughput of a cell and does not provide information about the actual load.

Different load metrics based on packet level information have been proposed. The authors of Reference [16] used the number of retransmission attempts needed to successfully transmit a single packet, which can be derived if all hidden pairs are known. The same concept was also used in Reference [17] to derive the Gross Load metric using a different formulation. Reference [17] also suggests using the packet loss estimation as a new load metric. Traffic (in bytes/s) was used as a load metric in References [4] and [7]. However, we should bear in mind that the IEEE 802.11 standards define several modulations with different physical bit rates (e.g. 1, 2, 5.5 and 11 Mbps for 802.11b); in this case an AP could be congested when carrying traffic of 1 Mbps if there are associated stations transmitting at the slowest bit rate. On the other hand, the same AP could also be considered under-utilised with a load of 3 Mbps if all of its clients use faster modulations. Therefore, carried traffic is not a valid representation of the load on an AP in a multi-rate scenario. Instead, in References [5] and [6] the measure of busy time is proposed as the representative load metric. More precisely, in Reference [6] the network congestion level is estimated using channel occupation time and by monitoring the occupation of the AP's buffer queue. Note that the AP queue size does not represent the overall network congestion level in the case of asymmetric traffic load since it does not depend on the uplink traffic, which also contributes to the resource saturation. However, downlink traffic generally overrides uplink traffic in many Internet applications.

3. LOAD INFORMATION IN NEW IEEE 802.11 STANDARDS

The IEEE 802.11k group (Radio Resource Measurement) is currently developing a standard which is intended to improve the provision of data from the physical and medium access layers by defining a series of measurement requests and reports that can be used in the upper layers to carry different radio resource management mechanisms. The current draft version is 9.0 [18], although the final standard is expected to be released soon (at the time of writing).

The current IEEE 802.11 standard [19] and the future 11k define a set of load metrics that are either broadcast by APs or measured directly by client stations:

Channel Load Report: any station accepting a Channel Load request shall respond with a Channel Load Report that is defined as the proportion of the time during which either the physical carrier sense, the virtual carrier sense (network allocation vector or NAV) or both indicate that the channel is busy. This measurement is similar, although not identical, to the CCA report in 802.11h.

Beacon Frames: these management frames are extended with three new elements that provide information about the load of an AP.

- *BSS Average Access Delay*: average medium access delay (MAD) for any transmitted frame measured from the time the frame is ready for transmission (i.e. begins CSMA/CA access) until the actual frame transmission start time.
- *BSS AC Access Delay*: in QAPs (QoS enabled APs), average MAD for each of the indicated Access Categories defined by the IEEE 802.11e (best effort, background, video and audio).
- *BSS Load*: contains information on the current STA population and traffic levels in the BSS. Includes the following fields:
    - *Station Count*: the number of stations currently associated with the AP.
    - *Channel Utilisation*: the percentage of time that the AP senses the medium is busy, which is indicated by either the physical or NAV mechanism.
    - *Available Admission Capacity* (AAC): the remaining amount of medium time available via explicit admission control.

3.1. Service time and medium access delay

Service Time is usually defined as the time interval from the moment at which a frame is at the head of its MAC queue ready for transmission until it is successfully received at the destination. This definition includes lost time caused by collisions or transmission errors. The value can be calculated by using analytical models [20–22]. Nevertheless, the simplest definition can be derived from Bianchi's formulation [23] for saturation conditions: Service Time is the interval between two consecutive, successful transmissions performed by the same station in a cell with n active nodes:

$$T_{\text{serv}} = n \cdot \frac{E_s}{P_{\text{tr}} P_s} \tag{1}$$

where E_s is the average length of a renewal interval, defined as the time between two consecutive transmissions or the time between two consecutive backoff decrements, taking into account collisions, transmission errors, idle time and data transmission time. P_{tr} is the conditional probability that at least one transmission occurs in a randomly chosen slot time and P_s the conditional probability that this transmission is successful. See Reference [23] for further details. From Equation (1) it is clear that T_{serv} is proportional to the number of active users, but note that E_s and the product $P_{\text{tr}}P_s$ also depend on n. The average service time is equal for all stations in the same cell due to the long-term fairness provided by the CSMA/CA access scheme, so any given station can estimate the average service time from local measurements.

Similar to the above parameter, 11k defines MAD as the average time a frame is held in the transmission buffer queue. MAD is measured from the time at which the DCF frame is ready for transmission (i.e. CSMA/CA access begins) until the actual frame transmission start time. If a frame needs to be retransmitted (because it has not been acknowledged) the value of MAD is calculated by averaging the waiting times of the r retransmission attempts:

$$\text{MAD} = \frac{\sum_{i=0}^{r} \text{MAD}_i}{r+1} \tag{2}$$

Figure 1 shows the difference between Service Time and MAD as the number of users is increased. The first parameter comprises the time needed to successfully send a frame and MAD is the average time a frame is held in transmission buffers due to contention. The figure was obtained from the analytical model based on stations in saturation state (1). MAD was also derived from Equation (1) with the average number of retransmissions

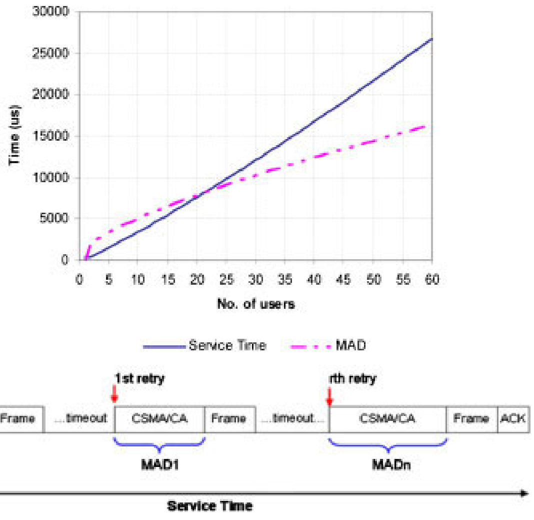

Figure 1. Service time and medium access delay (MAD).

DOI:10.1002/ett

[17]. The required formulation is omitted for the sake of brevity.

3.2. Channel busy time

Channel busy time is the percentage of time during which the medium is sensed as busy. Busy time can be modelled using the formulation defined in Reference [23] for saturation conditions, as in the previous section:

$$T_{\text{busy}} = 1 - \frac{(1 - P_{\text{tr}})\sigma}{E_s} \tag{3}$$

E_s and P_{tr} have already been defined; σ is the slot time unit defined by IEEE 802.11 standards (20 μs for.11b and 9 μs for.11a/g).

The medium is under-utilised when there are a small number of competing stations due to the backoff mechanism used in the IEEE 802.11 CSMA/CA implementation. On the other hand, a larger number of competing stations will decrease the probability of finding an idle slot. Figure 2 illustrates these efficiency issues and shows how busy time is affected by the number of active users, the average frame size and the offered traffic. The figure is obtained from simulations using IEEE 802.11g standard values at 48 Mbps in a single-cell scenario. Section 6.1 offers further explanation of the simulation working procedure.

Although channel busy time provides a good representation of the cell load even in a multi-rate scenario, it is not a valuable metric in the presence of greedy applications (e.g. FTP). For example, a channel busy time of 85% is achieved with a single greedy station (Figure 2a), but also with 20 users, each of which is loading the cell with 1 Mbps (Figure 2b). However, a new station will get much more bandwidth if it only has to compete with one user than if it has to share the medium with 20 other stations.

3.3. Available admission capacity

The IEEE 802.11e standard defines AAC as the remaining amount of medium time available in units of 32 μs, although it does not specify how it should be calculated. We propose a new load metric based on a more precise definition of available capacity. We expand the current definition of AAC to be the proportion of time a new station can take up if it is associated with the AP at a given physical rate. This new metric provides a vision of cell load that takes into account the effect of multi-rate stations, the presence of greedy users, the average frame size and the number of active users. We provide a more detailed explanation in the next section.

4. THROUGHPUT ESTIMATION FOR A NEW STATION

Straightforward measurements can be used to derive the throughput that is currently devoted to a station with ongoing data transmissions. However, more detailed study is required to determine how to predict the throughput that a new station will obtain before it actually starts to transmit or the maximum bandwidth that can be allocated to a station if it increases its offered traffic in a multi-rate cell. In a previous work [24], we proposed an algorithm that can accurately calculate the throughput available to a new station, which is obtained from AAC. This algorithm is based on the assumption that the IEEE 802.11 MAC maintains fairness in terms of access probability independently of the rate and bandwidth requirements of each station. It also takes into account the inherent 'performance anomaly' [25] in multi-rate CSMA/CA networks. Based on these statements, we define T_{cycle} as the average time

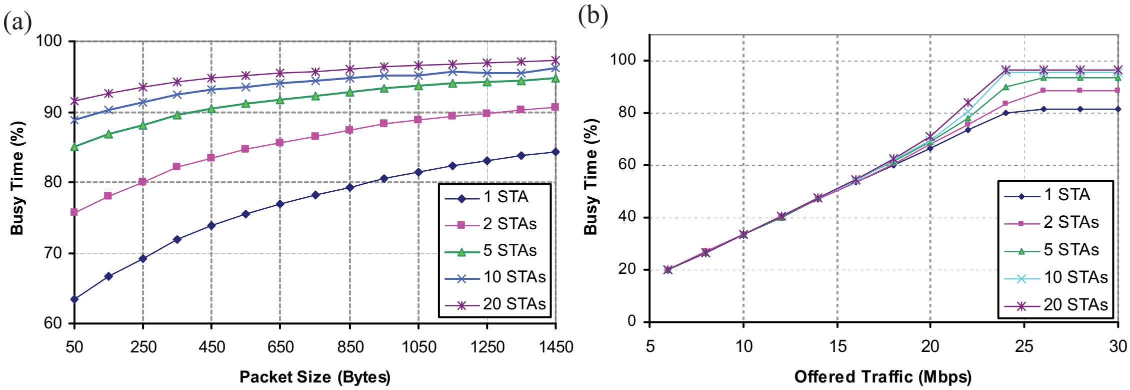

Figure 2. Busy time depending on (a) frame size and (b) number of saturated users (frame size = 1024 bytes).

required to send one frame from each of the competing stations:

$$t_i = \text{DIFS} + T_{\text{data}}(i) + \text{SIFS} + T_{\text{ACK}} \tag{4}$$

$$T_{\text{cycle}} = \max(\text{TBO}_i) + \sum_{i=1}^{n} t_i \tag{5}$$

Note that t_i is defined for a basic CSMA/CA access; if RTS/CTS handshake is used, 2·SIFS, T_{RTS} and T_{CTS} must be added to t_i. The duration of the data frame is $T_{\text{data}}(i)$. TBO_i represents the average time during which a station i waits for the backoff timer to expire before attempting to transmit. Note that under saturation conditions, backoff slots are shared. Therefore, the idle time spent on backoff within a T_{cycle} is equivalent to the largest average backoff in the cell. Without saturation conditions, stations are not 'synchronised' and the probability that two stations share their backoff periods (or a part of them) is lower. Consequently, without saturation, our approximation introduces some error. Nevertheless, in Reference [26] we show that the error introduced is small (typically below 6%).

TBO_i depends on the number of previous transmission attempts. The average value of the backoff interval, $T_{\text{BO}}(j)$ after j consecutive transmissions is given by

$$T_{\text{BO}}(j) = \begin{cases} \frac{2^j(CW_{\min}+1)-1}{2} T_{\text{slot}} & 0 \leqslant j \leqslant 6 \\ \frac{CW_{\max}}{2} T_{\text{slot}} & j \geqslant 6 \end{cases} \tag{6}$$

As in Reference [17], we consider that the number of packet retransmissions required to successfully transmit a single packet is a geometrically distributed random variable. If P_i is defined as the probability that a frame sent by i has to be retransmitted (including the effects of collisions and channel errors), the average backoff interval TBO_i for station i is

$$\text{TBO}_i = \sum_{j=0}^{\infty} (1 - P_i) P_i^j \cdot T_{\text{BO}}(j) \tag{7}$$

Note that from Equation (7) we have to compute an infinite series. However, its operation can be stopped when the required precision is met. For example, for typical P_i values ranging from 0.1 to 0.2 and a required precision of 1 μs, 6 to 10 iterations are enough.

T_{data} can be further decomposed into[†]

$$T_{\text{data}}(i) = T_{\text{preamble}} + \frac{8 \cdot (H_{\text{MAC}} + \text{MSDU}_i)}{r_i} \tag{8}$$

where r_i is the physical bit rate at which node i sends data frames with a payload of MSDU bytes. The remaining values vary according to the standard and the modulation used.

We define the overhead produced in layers 1 and 2 of a node i as $\text{OH}_i = r_i \cdot t_i / (8 \cdot \text{MSDU}_i)$. In order to calculate the actual TO (traffic offered) it is necessary to consider this overhead and the possible retransmissions given the traffic offered by the upper layers (TO_{app}):

$$\text{TO}_i = \frac{\text{OH}_i}{1 - P_i} \text{TO}_{\text{app}}(i) \tag{9}$$

Resource distribution in WLAN provides a Max-min fairness in which small flows receive the volume they demand and larger flows share the remaining capacity equally. In other words, all stations whose traffic (TO_i) is equal to or smaller than the bandwidth that should be allocated under saturation conditions will be able to carry all of their offered traffic (TO_{app}). Saturated stations will share their corresponding bandwidth and the excess time that is not used by the non-saturated nodes. The algorithm is defined as follows:

Algorithm 1: AAC Algorithm

```
$T_{cycle}' \leftarrow T_{cycle}$
OrderIncr(N, $\delta_i$)
for all $i \in N$ do
  if $\delta_i \leqslant 1$
    $L_i \leftarrow TO_i/r_i$
    $T_{exc} \leftarrow (1 - \delta_i)/r_i$
    $T_{cycle}' \leftarrow T_{cycle}' - 1/r_i$
    for all $j \in N$ and $\delta_j > 1$ do
      $L_j \leftarrow L_j(1 + T_{exc}/T_{cycle}')$
      $S_j \leftarrow L_j r_j$
      $\delta_j \leftarrow TO_j/S_j$
    end for
  end if
end for
```

where S_j is the throughput that station j would obtain in saturation conditions. The parameter δ_i is defined as the proportion of the maximum throughput that can be achieved by station i which is actually used: $\delta_i = \text{TO}_i / S_i$. Before the first execution, stations are ordered according to increasing value of the parameter δ. Stations with $\delta \leqslant 1$ use fewer resources (or an equal number) than those that would be allocated under saturation conditions, and

[†] case of DSSS-CCK modulations, for more details on OFDM see Reference [27].

DOI:10.1002/ett

their $\mathrm{TO}_{\mathrm{app}}$ will therefore be carried. The proportion of time that is not used by non-saturated stations (T_{exc}) will be divided fairly between the remaining stations according to a new time T_{cycle}' in which the stations that have already been served are not considered. Note that greedy applications are modelled with $\delta_j > 1$ regardless of the value of the S_j parameter. The TO of a new station is not known in advance and the maximum throughput it can potentially achieve is therefore calculated in saturation. The value of L_i represents the individual load contributed by node i defined as the proportion of T_{cycle} that is used by the ith node.

By using this formulation it is possible to perform the capacity estimation in real time, assuming that the required statistics are updated regularly by the firmware/driver of the wireless interface. Consequently, this estimation can handle varying traffic demands and varying channel conditions, since it also takes into account the effect of collisions and errors produced by noise and interference. Due to space constraints, the accuracy of the capacity estimations described in this section is evaluated in the following technical report [26]. In this regard, we could ascertain that increasing the accuracy of the model, entails more complexity but does not involve a better load balancing.

5. IMPLEMENTATION

It is essential to acquire the necessary parameters when implementing the proposed algorithm. The AP is the only node that is able to calculate the available bandwidth for a given user in real time since it maintains statistics from which all of the required parameters can be derived, including the physical rate used by each station, the MSDU size and P_f (details are given in Reference [24]).

The AP needs to know the amount of traffic contributed by each of its clients (TO_i) or the percentage of time devoted to each of its clients (L_i), in addition to their association rate. All of these parameters can be estimated by examining the AP's MIB counters and statistics. This association management has been implemented by using out-of-band signalling (through a cellular interface) which is described in the following, although we also discuss two alternatives for a more general case. The AAC calculation requires the candidate station's rate, which is not known until the association process has been completed. In the absence of out-of-band signalling we distinguish two implementation approaches depending on whether the new station's rate is known by the AP or not, following an active or a passive scan process. In the out-of-band approach, the association control can be managed from the network side, whereas the active and passive scan solutions remain purely client-driven.

5.1. Active scanning

The stations can perform an active scan in order to find the APs within range. For each channel, the station willing to associate with a new AP broadcasts a Probe Request frame. Any AP shall respond with a Probe Response to the address of the station (STA) that generated the Probe Request. Since the AP is able to measure the SNR of the received Probe Request, it can predict the most suitable rate for the requesting station. The new Probe Response frames include the AAC field; if the AP knows the best rate that the client is likely to use in subsequent transmissions it can calculate a specific AAC for each request. The requesting station also knows the best physical rate once it knows the signal quality perceived by the AP, which can be transmitted using the RCPI (received channel power indicator) information element present in the Probe Response. The AAC is then a 2-byte codification of L_i. The requesting station i gathers all L_i values received from the APs within range and associates with the cell that will provide the best throughput ($L_i \cdot r_i$).

5.2. Passive scanning

Stations that use passive scanning listen to each channel and wait for beacon frames to identify all APs within range. Unlike in active scanning, STAs do not generate any frame during a passive scan. As a result, APs are unable to calculate specific AACs for all of the possible candidate stations. Conversely, APs can set the AAC value broadcast through beacon frames with the available capacity that can be allocated to a new STA that uses the fastest possible rate. These beacons produce two important changes: STAs can select the most suitable rate by estimating the up-link margin from the TPC (Transmit Power Control) Report included in 11k beacons; and stations can determine the best AP based on the value of AAC (normalised to 1) and the number of associated stations (NSTA).

$$r_{\mathrm{avg}} = \frac{\mathrm{AAC}}{\frac{1-\mathrm{AAC}}{\mathrm{NSTA}}} r_{\max} \qquad (10)$$

With Equation (10), STAs can estimate the average rate used in the cell (r_{avg}) and are able to find the specific AAC for their particular rate by applying the previous algorithm.

Note that Equation (10) is only valid under saturation conditions and provides the same values as those obtained through the active scan; otherwise, the results are inconsistent with the current available capacity. If the associated STAs do not saturate the cell or, more precisely, when the channel utilisation field is approximately 1-AAC, the available capacity can be directly calculated by multiplying the broadcast AAC by the rate, although in some cases the assumptions made and the error produced affect the capacity estimation and there is no guarantee that the best AP will be selected.

This implementation provides approximate capacity estimations so active scanning is preferred. Nevertheless, the same results could be obtained by changing the definition of beacon frames if the available capacity were calculated and broadcast for each of the rates supported by the AP.

5.3. *Out-of-band signalling implementation*

The number of dual-mode devices (predominantly handsets) with WLAN/cellular network interfaces is growing exponentially. Future devices will be released with several network interfaces due to VoIP services and convergence driven by network operators. In this scenario, a feasible solution would be to use the cellular interface as an out-of-band signalling tool which would enable stations to transmit information prior to WLAN association.

In Reference [28], we presented the design and implementation of a novel management architecture for a wireless mesh network. Client stations and network nodes (i.e. APs and gateways) use the cellular interface for signalling purposes only. All signalling data is transmitted to a central management entity which performs several tasks including AAA (authentication, authorisation and accounting), mobility management, security key distributions and dynamic frequency allocation algorithms. Note that this approach uses the cellular interface as a highly available Internet connection that allows the communication with the central unit; i.e. the system is independent of the cellular network operator and in consequence it can be implemented by any kind of provider. Nevertheless, it would also be possible to take advantage of the operator's AAA mechanisms through additional agreements. This architecture also allows the implementation of our proposed association management, as detailed in the following paragraph.

In the case of stations, the WLAN interface is used to perform a passive scan of the medium prior to association which gathers information on the power level, channel and BSSID of all APs within range. This information is transmitted to the management entity which estimates the most suitable association rates that the client is likely to use with the different candidate APs, according to the power levels reported by the client. The candidate APs will compute the AAC according to the algorithm in Section 4 upon request by the central manager and send the value to the central entity. The central management entity then determines the best association for the requesting station and sends a response via the cellular interface which includes additional details that are required in order to establish a successful association: the BSSID of the selected AP, channel, ESSID and encryption key. As a result, the station will be able to associate with the best AP within the client's transmission range.

The client software for association control was implemented in a Linux-based laptop (Debian 4.0 distribution), with Intel® ipw3945 802.11a/b/g chipset [29] and HSDPA/UMTS PCMCIA card. The AAC module was installed on commercial APs (AccessCube [30]) with an embedded Linux distribution running a 2.4.27-r11 kernel. The APs' wireless interface is an 802.11b Prism 2.5 based card, supported by HostAP driver version 0.3.7 [31], which uses the */proc* filesystem to store a large number of statistics. The central entity is based on a set of software modules that run on a PC with an Intel® P4 3GHz dual core processor with 2 GB of RAM and a Linux OS Debian 4.0 distribution with kernel version 2.6.18.

6. EVALUATION

The new IEEE 802.11k APs will enable WLAN clients to select the best association based on signal strength measurements, MAD values, Channel Busy time, number of associated STAs and available admission capacity. In our simulations we compare four different association control schemes:

RSSI: current default association scheme based on the received signal strength indicator, i.e. stations select the AP that provides the strongest signal.

NSTA: STAs select the AP within transmission range that serves the lowest number of associated stations. Ties between APs with an equal number of associated STAs are broken by RSSI decisions.

MAD: STAs select the AP that provides the lowest MAD. Note that MAD is coded with 1 byte, the values of which will be a logarithmically-scaled representation of the current MAD. MAD is therefore more sensitive to low delay measurements. For example, $\text{MAD} = 1$

DOI:10.1002/ett

represents an access delay of between 50 and 51 µs, but if MAD = 253 the delay varies between 4396 and 5498 µs. Ties between APs broadcasting identical MAD values are broken by RSSI measurements.

AAC: STAs select the AP that provides the highest AAC value following an active scan.

6.1. *Scenario*

Note that current multi-cell WLANs are designed with three or four cell clusters due to spatial reuse restrictions and only three non-overlapping channels in the 2.4 GHz band. That is to say, in a well-planned ESS most users are able to obtain services from up to four different APs. Therefore, the evaluation process we designed is based on extensive simulations in a 60 × 60 m^2 indoor scenario with four IEEE 802.11g APs, without losing generality. Thus, it is easy to scale the conclusions derived from our results to larger scenarios. We performed the evaluation by using a simulation tool that closely adheres to all IEEE 802.11 protocol details [32]. Using this tool allowed customised simulations with a short time-to-deploy in which it was easy to define the new parameters to be measured. We ran a large number of independent simulations and obtained small confidence intervals, which are therefore not shown in the figures.

We assume that APs use non-interfering channels. As was stated in References [1–3], users are static and tend to be spatially concentrated. We simulate these characteristics by placing users at random, but forcing that 50% of users are concentrated in an area of 30 × 30 m^2 around one of the APs (AP0). This ensures that it is possible to identify the benefits of association management beyond signal strength measurements, but also a realistic scenario is met. Figure 3 shows the possible user distribution and association when the RSSI scheme is used.

The physical rate used for transmissions depends on the distance between an STA and its selected AP. Figure 4 shows the correlation between rate and distance using the propagation model for semi-open offices proposed in Reference [33].

6.2. *Saturation*

The first set of simulations are intended to recreate the worst-case scenario, in which all stations are running greedy applications and there is always a frame ready to be sent in every station's tx buffer. If a fixed rate is set for all the stations, there will be unnoticeable differences between the throughput carried by the APs, but if we use MAD, Service Time or AAC to measure the load, an unbalanced situation is revealed: 50% of the users are associated with AP0 when the traditional RSSI approach is used, which is clearly the worst solution. In this specific case (saturation and fixed rate), NSTA associations are slightly better than those using MAD and AAC since all stations represent the same load, i.e. balancing the number of stations implies balancing load. The effect of load balancing can be seen when multiple rates are added, even if we measure carried throughput (CT). Figure 5 shows the throughput balancing fairness between the APs. Our

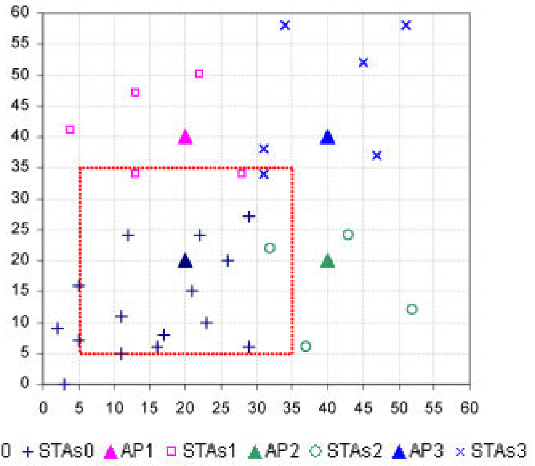

Figure 3. Random distribution of 30 users following an RSSI-based association scheme.

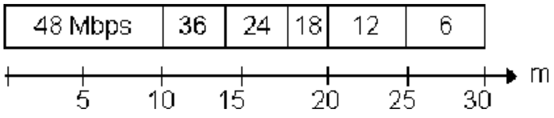

Figure 4. Coverage radius of an AP for different physical rates.

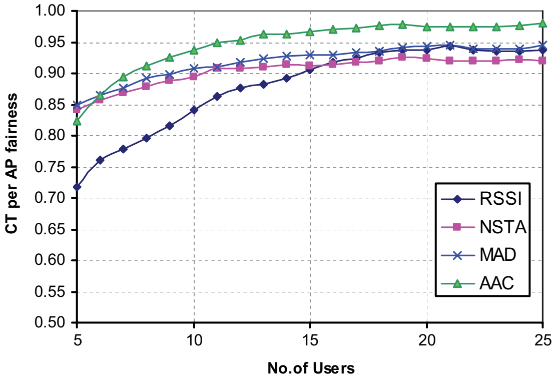

Figure 5. Throughput per AP fairness under saturation conditions.

approach improves the traditional RSSI scheme by 15%. Fairness is measured using the known Jain's Index [34]: β is a value between 0 (unfair) and 1 (fair); if only k of n flows receive equal resources and others receive none, the index is k/n. S_i is the throughput achieved by STA i.

$$\beta = \frac{\left(\sum_{i}^{n} S_i\right)^2}{n \sum_{i}^{n} S_i^2}; 0 \leqslant \beta \leqslant 1 \tag{11}$$

However, it is easier to identify the benefits of our approach if we analyse the load from the user perspective. Figure 6(a) shows the aggregate throughput of the whole network as we increase the number of associated stations. It can be seen that the RSSI-based scheme has the best performance since it is the only approach that can guarantee that all clients will always use the highest possible physical rate. In contrast, RSSI associations give an unfair share of resources among users. Figure 6(b) shows that the RSSI-based association scheme provides an imbalanced distribution of available throughput and is by far the most unfair solution (AAC scheme improves RSSI by more than 45% in the worst case); conversely, MAD provides a high degree of fairness but low throughput.

Two useful parameters for measuring performance and fairness are the maximum service time and the minimum throughput obtained by a given station (see Figures 6(c) and (d), respectively). By monitoring the evolution of these parameters it can be seen that our AAC proposal provides a clear improvement over the other solutions (up to 25 and 35%, respectively over RSSI). After reviewing this set of figures we can conclude that AAC has both a high fairness index and a good aggregate throughput under saturation conditions.

6.3. *Degree of saturation*

The next set of simulations was run under the same conditions as described in the previous subsection, with the

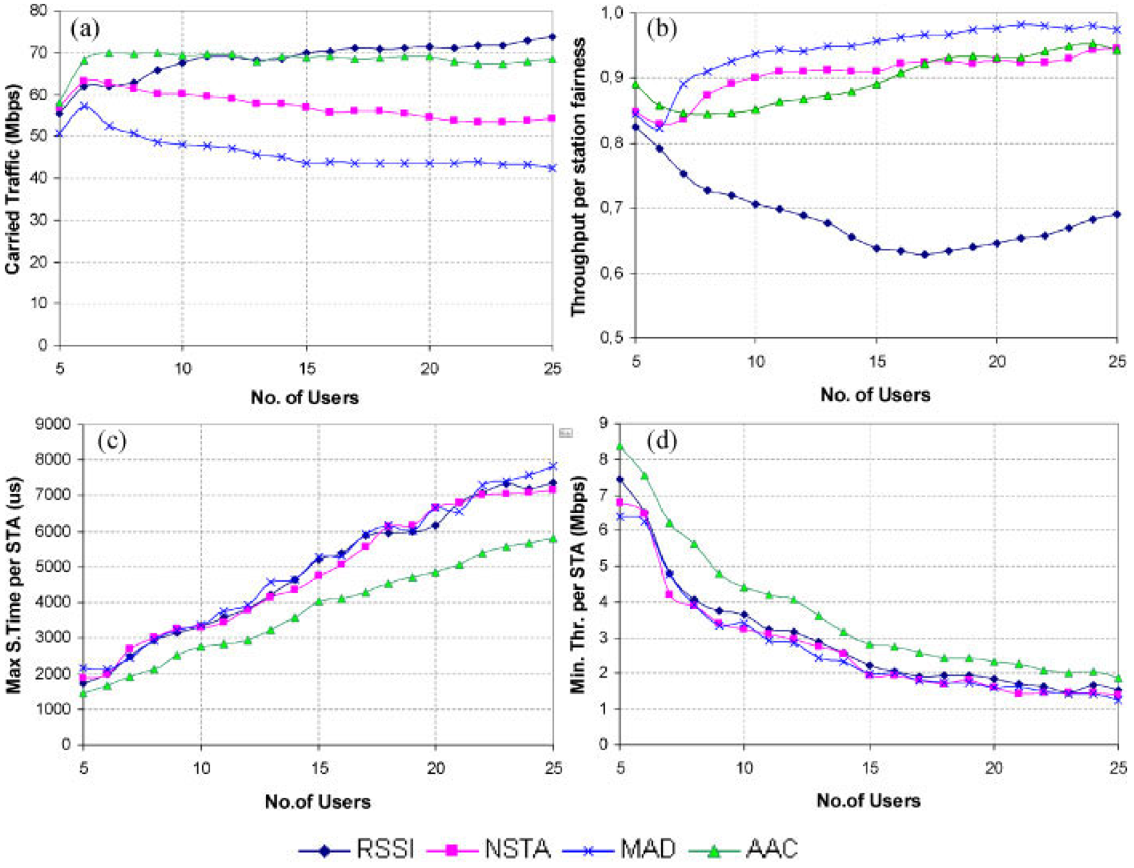

Figure 6. (a) Aggregate throughput versus number of users, (b) throughput fairness index, (c) max. service time (μs) for a STA and (d) min. throughput for a STA.

DOI:10.1002/ett

exception of user traffic load profiles. In this case we define three different user types according to level 2 traffic demands:

- *Saturated*: greedy users. They always have frames ready to be sent (frame size: 1450 bytes).
- *Medium*: bursty traffic of up to 2.5 Mbps with an average frame size of 1024 bytes. This user type accounts for 70% of non-saturated users.
- *Low*: constant bit rate of up to 1 Mbps with a frame size of 500 bytes. This user type accounts for 30% of non-saturated users.

Since the differences in carried throughput between the four schemes are minimal, it is interesting to show the level of balancing achieved with AAC (Figure 7a) and MAD (Figure 7b). Both graphs show the minMAX ratio (the lowest value found on an AP divided by the highest). It can be seen that RSSI provides the worst level of balancing in all cases. The load (understood as the proportion of time during which the cell is busy) which is translated into available capacity is better balanced when AAC is used, but it is outperformed by MAD if we measure T_{serv}. However, the best balance measured in terms of T_{serv} does not compensate for the poor performance of MAD if we also consider the network capacity. Figures 8(a) and (b) show the aggregate throughput and the fairness index as the number of saturated stations is increased (from 0 to 100%). The Fairness Index has been redefined to include non-saturated stations. Note that the presence of Low and Medium stations with Equation (11) would produce low fairness values even though the stations are able to carry all of their offered traffic. The new fairness index is calculated as follows:

$$\beta = \gamma \frac{\left(\sum_{i}^{i\in sat} S_i\right)^2}{N_s \sum_{i}^{i\in sat} S_i^2} + (1-\gamma)\frac{\left(\sum_{i}^{i\notin sat}\frac{S_i}{\mathrm{OT}_i}\right)^2}{N'\sum_{i}^{i\notin sat}\left(\frac{S_i}{\mathrm{OT}_i}\right)^2} \qquad (12)$$

where γ is the proportion of saturated stations, S_i is the throughput obtained by station i and OT_i is the traffic offered by station i.

Although the differences between the various association schemes become clearer as the number of saturated stations increases, we can draw almost the same conclusions as under saturation conditions: RSSI scheme provides the best aggregate throughput at the cost of providing the poorest fairness. AAC still provides good aggregate throughput (the best for a small number of saturated STAs) while, at the same time, fairness is assured.

We argue that if the conclusions are the same as those for saturation conditions, there is no need to apply more complex traffic generators. The use of more realistic traffic patterns mainly affects the way in which new radio measurements are taken: the key parameters that must be taken into account are the duration of the measurement, the frequency with which they are taken, the number of repetitions and the time during which the results remain valid, depending on the precision required. The authors of Reference [35] suggest using confidence intervals to optimise these parameters.

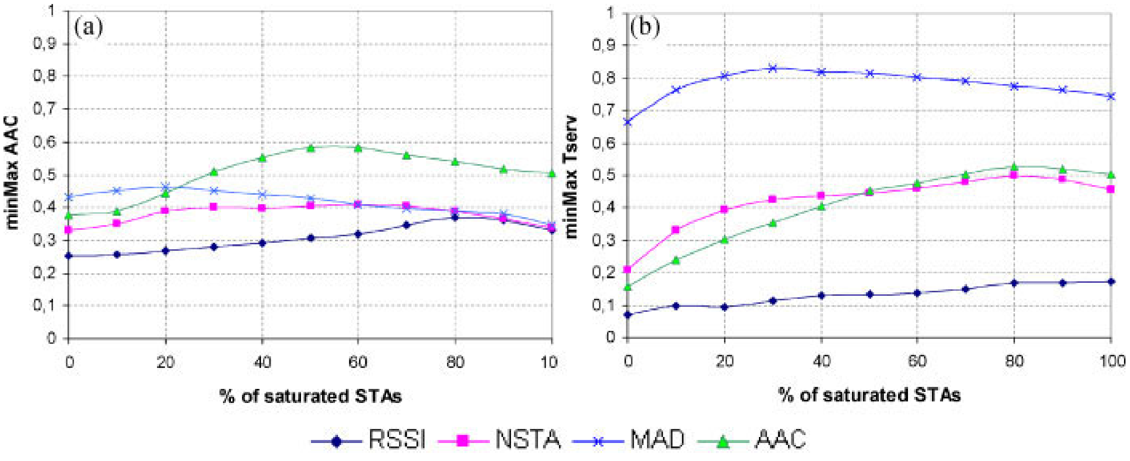

Figure 7. (a) Ratio minMAX(AAC) versus % of saturated users and (b) ratio minMAX(Tserv).

DOI:10.1002/ett

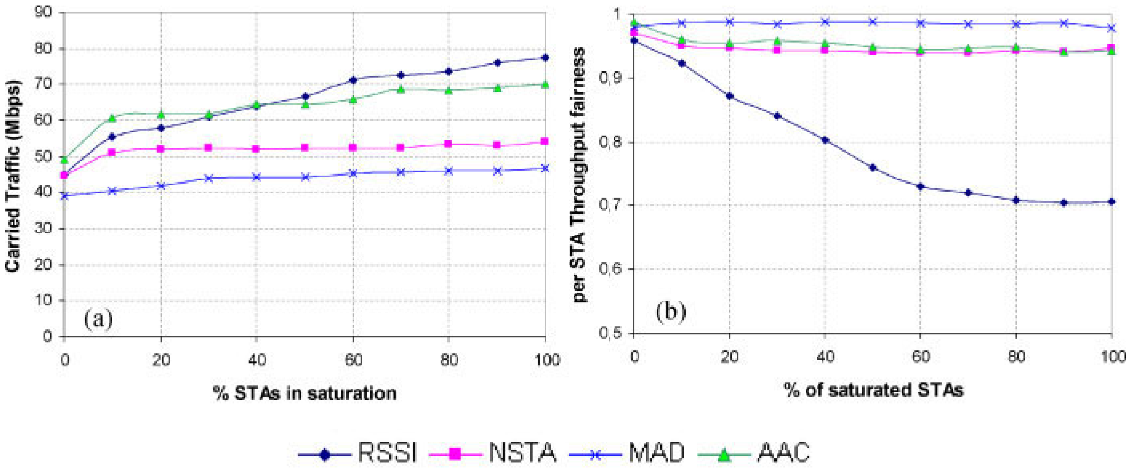

Figure 8. (a) Aggregate throughput versus % of saturated users and (b) throughput fairness index.

6.4. *Fairness versus throughput*

The AAC proposal could be improved in terms of fairness. We propose a weighted combination of MAD and AAC metrics in order to resolve the throughput/fairness trade-off: a STA compares AAC and MAD values received from all APs in range; if $\mathrm{AAC}_j > \alpha \mathrm{AAC}_{j-1}$ and $\mathrm{MAD}_j < \mathrm{MAD}_{j-1}$, then the STA selects AP_j for association. Where $\alpha \in [0, 1]$, a value of 0 produces a pure MAD selection and a value of 1 is based only on AAC. Figure 9 shows that the aggregate throughput decreases as the fairness index is increased in the same scenario described in Section 6.3, with 25 users and 40% of saturated STAs. Fairness decreases rapidly when $\alpha > 0.7$, but a slight loss of fairness in this region leads to a dramatic improvement in throughput: for example, a 3% decrease in fairness is compensated by a 10% improvement in throughput.

6.5. *Testbed*

Finally, we present practical measurements taken in a small testbed that we can use as a proof of concept. These practical results agree with the conclusions derived from the more comprehensive simulation study previously detailed. The scenario consists of two IEEE 802.11b APs, A_1 and A_2. A_1 serves one STA at 1 Mbps; A_2 serves two STAs, both using 11 Mbps. All STAs are in saturation, sending 1500Byte UDP frames. A forth STA is activated, which receives signal from A_1 with SNR = 20 dB and from A_2 with SNR = 11 dB; the new STA can be associated with A_1 at 11 Mbps, while it should use 5.5 Mbps if associated with A_2. It is clear that following an association based on RSSI, the number of associated stations, or either based on carried traffic, the new station chooses A_1. In saturation, competing against one STA at 1 Mbps, the new client gets 766 ± 30 kbps. However, if the new station associates with A_2, the throughput is increased to 1672 ± 54 kbps, even though it has halved its physical rate. Those numbers correspond with the values of AAC: 766 and 1672 is the available throughput for a new station in A_1 and A_2, respectively. Therefore, according to our proposed scheme, the new STA chooses A_2 obtaining more resources. Only when the three previously associated STAs offer less than 400 kbps each (i.e. not in saturation), all four schemes coincide in selecting A_1.

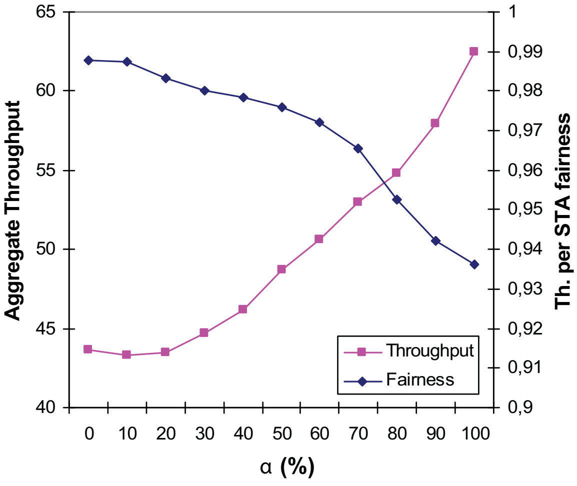

Figure 9. Throughput versus fairness with different values of α.

DOI:10.1002/ett

7. CONCLUSIONS

The radio measurements and mechanisms introduced by the new IEEE 802.11 standards provide valuable information that can be used by stations when they have to select the most suitable AP for association. We have introduced a new load metric that can be provided by APs to ensure that stations are still able to perform client-driven associations and roaming in accordance with the operations defined in the IEEE standards. The new metric proposed is derived from a more precise definition of the ACC field and takes into account the influence of the main factors that affect the load of a WLAN: the number of users in a cell, the user rate (modulation), the signal quality and the offered traffic.

In the worst-case scenario, which is in fact a realistic situation, a large number of users are concentrated in a small area and traditional association schemes based on RSSI measurements guarantee that the fastest physical rate is reached. However, this leads to an uneven distribution of load, since most of the users will be associated with a small number of APs and will therefore receive a poor service. The new AP selection method not only provides good network performance but also ensures an even share of bandwidth among clients and a balanced load among APs. In other words, an association scheme based on available capacity ensures load balancing among the APs of a given WLAN, increases the overall capacity and maintains fair resource sharing from the user perspective. Our mechanisms are of great help in the provision of QoS, but additional features are still required to guarantee any quality.

ACKNOWLEDGMENTS

This work has been supported in part by DURSI, the ERDF and the Spanish Government through project TEC2006-04504 and CDTI, Ministerio de Industria, through Segur@ project.

We wish to thank ETT's editorial board and the reviewers for their valuable and insightful comments that indubitably helped us to improve the quality of this document.

REFERENCES

1. Balachandran A, Voelker GM, Bahl P, Rangan PV. Characterizing user behavior and network performance in a public wireless lan. In *Proceedings of ACM SIGMETRICS 2002*, Vol. 30, June 2002; 195–205.
2. Balazinska M, Castro P. Characterizing mobility and network usage in a corporate wireless local-area network. In *Proceedings of First International Conference on Mobile Systems, Applications, and Services, MobiSys'03*, May 2003.
3. Henderson T, Kotz D, Abyzov I. The changing usage of a mature campus-wide wireless network. In *Proceedings of the 10th annual international conference on Mobile computing and networking*, September 2004; 187–201.
4. Balachandran A, Bahl P, Voelker GM. Hot-spot congestion relief in public-area wireless networks. In *Proceedings of 4th IEEE Workshop on Mobile Computing Systems and Applications*, June 2002; 70–80.
5. Bejerano Y, Han S-J, Li LE. Fairness and load balancing in wireless LANs using association control. In *Proceedings of the 10th international conference on Mobile computing and networking, MobiCom'04*, 2004; 315–329.
6. Bazzi A, Diolaiti M, Pasolini G. Measurement based call admission control strategies in infrastructured IEEE 802.11. In *The 16th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2005*, September 2005.
7. Velayos H, Aleo V, Karlsson G. Load balancing in overlapping wireless LAN cells. In *IEEE International Conference on Communications, ICC'04*, Vol. 7, June 2004; 3833–3836.
8. Brickley O, Rea S, Pesch D. Load balancing for QoS optimisation in wireless LANs utilising advanced cell breathing techniques. In *IEEE 61st Vehicular Technology Conference*, VTC 2005-Spring, May 2005.
9. Bejerano Y, Han S-J. Cell breathing techniques for load balancing in wireless LANs. In *Proceedings of the 25th IEEE Annual Conference INFOCOM'06*, April 2006.
10. Garcia E, Vidal R, Paradells J. Load balancing in WLANs through IEEE 802.11k mechanisms. In *Proceedings of the IEEE Symposium on Computers and Communications (ISCC'06)*, June 2006; 844–850.
11. Nicholson A, Chawathe Y, Chen M, Noble B, Wetherall D. Improved access point selection, In *4th International Conference on Mobile Systems, Applications and Services, MOBISYS'06,* June 2006.
12. Papanikos I., Logothetis M. A study on dynamic load balance for IEEE 802.11b wireless LAN. In *8th International Conference on Advances in Communications and Control, COMCON'01*, June 2001.
13. Cisco aironet 1200 series access points, Cisco Systems Inc., 2005.
14. Orinoco ap-600 access point, data sheet, Proxim Corp., 2004.
15. Bianchi G, Tinnirello I. Kalman filter estimation of the number of competing terminals in an IEEE 802.11 network. In *22nd Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM'03*,Vol. 2, April 2003; 844–852.
16. Dhou IB. A novel load-sharing algorithm for energy eficient MAC protocol compliant with 802.11 WLAN. In *IEEE 50th Vehicular Technology Conference, VTC 1999-Fall*, Vol. 2, September 1999; 1238–1242.
17. Bianchi G, Tinnirello I. Improving load balancing mechanisms in wireless packet networks. In *IEEE International Conference on Communications, ICC 2002*, Vol. 2, April 2002; 891–895.
18. IEEE 802.11 WG. Draft Supplement to Standard for Telecommunications and Information Exchange Between Systems—LAN/MAN Specific Requirements—Part 11: Wireless Medium Access Control and Physical Layer Specifications: Specification for Radio Resource Measurement, IEEE 802.11k/D9.0. New York, USA: The Institute of Electrical and Electronics Engineers, Inc., September 2007.
19. IEEE 802.11 WG. IEEE Standard for Telecommunications and Information Exchange Between Systems—LAN/MAN Specific Requirements—Part 11: Wireless Medium Access Control and Physical Layer Specifications. New York, USA: The Institute of Electrical and Electronics Engineers, Inc., June 2007.

20. Carvalho M, Gracía-Luna-Aceves J. Delay analysis of IEEE 802.11 in single-hop networks. In *11th IEEE International Conference on Network Protocols*, November 2003; 146–155.
21. Chatzimisios P, Boucouvalas A, Vistas V. Packet delay analysis of IEEE 802.11 MAC protocol, *Electronics Letters* 2003; **39**:1358–1359.
22. Raptis P, Vistas V, Paparrizos K, Chatzimisios P, Boucouvalas A. Packet delay distribution of the IEEE 802.11 distributed coordination function. In *6th IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks, WoWMoM'05*, June 2005; 299–304.
23. Bianchi G. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on Selected Areas in Communications* 2000; **18**:535–547.
24. Garcia E, Viamonte D, Vidal R, Paradells J. Achievable bandwidth estimation for stations in multi-rate IEEE 802.11 WLAN cells. In *8th IEEE Symposium on a World of Wireless, Mobile and Multimedia Networks, WoWMoM'07*, June 2007.
25. Heusse M, Rousseau F, Berger-Sabbatel G, Duda A. Performance anomaly of 802.11b. In *Proceedings of the 22nd IEEE Annual Conference INFOCOM'03*, Vol. 2, March 2003; 836–843.
26. Garcia E. Available admission capacity estimations in IEEE 802.11 access points. *Technical Report*, April 2008. Available online at: http://hdl.handle.net/2117/2045
27. Qiao D, Choi S, Shin K. Goodput analysis and link adaptationfor IEEE 802.11a wireless LANs, *IEEE transactions on Mobile Computing* 2002; **1**:278–292.
28. Paradells J, *et al*. Design of an UMTS/GPRS assisted mesh network In *WWRF17 Meeting—Serving and Managing Users in a Heterogeneous Environment*, November 2006.
29. Intel® Wireless WiFi Link drivers for Linux. http://intellinuxwireless.org/
30. 4G Systems AccessCube. http://www.meshcube.org
31. Host AP Linux driver for Intersil Prism2/2.5/3 wireless LAN cards. http://hostap.epitest.fi
32. Lopez E, Casademont J, Cotrina J. Outdoor IEEE 802.11g cellular network performance. In *Proceedings of IEEE Globecom04*, November 2004.
33. Rappaport TS. *Wireless Communications Principles and Practices* (2nd edn). Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.
34. Jain R, Chiu D, Hawe W. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems, *Technical Report TR-301*, DEC Research, September 1984.
35. Mangold S, Berlemann L. IEEE 802.11k: Improving confidence in radio resource measurements, In *The 16th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC'05*, September 2005.

AUTHORS' BIOGRAPHIES

Eduard Garcia Villegas received his MSc degree from the Technical University of Catalonia in 2003. He is assistant professor at the same university and a member of the Wireless Networks Group (WNG). He is currently working towards his PhD. He has worked in several public funded research projects. His research interests include Radio Resource Management in WLANs, wireless security and wireless mesh networks.

José Luis Ferrer received his MSc degree from the Technical University of Catalonia in 2004. He is currently a researcher in the Wireless Networks Group (WNG) and working towards its PhD. He has worked in several public funded research projects related to mobile and wireless networks. His research interests include mobility management and routing protocols design and performance in ubiquitous sensor networks.

Elena Lopez-Aguilera is an Assistant Professor at the Technical University of Catalonia (UPC) in Barcelona. She received her M.S. Degree in Telecommunications Engineering from the UPC in 2001 and the Ph.D. Degree in 2008. She joined the Wireless Networks Group (WNG) at the Telematics Department of the UPC in 2002. In 2005-2006 she was an invited researcher at the Grenoble Computer Science Laboratory (LIG). Her main research interests include the study of medium access protocols in WLANs and mesh networks. She has published papers in the area of wireless communications and MAC mechanisms.

Rafael Vidal Ferré received his MSc degree from the Technical University of Catalonia in 1997. He is assistant professor at the same university and a member of the Wireless Networks Group (WNG). He is currently working towards his PhD. He has worked in several public funded research projects. His research interests include Mobility Management and Self-Configuration in wireless networks.

Josep Paradells is professor at the Technical University of Catalonia. He is the head of the Wireless Networks Group (WNG). His expertise areas are network convergence and ambient intelligence, combining theoretical studies with real implementations. He has been participating in national and European public funded research projects and collaborating with main Spanish telecommunications companies. He has published his research results in conferences and journals.

DOI:10.1002/ett