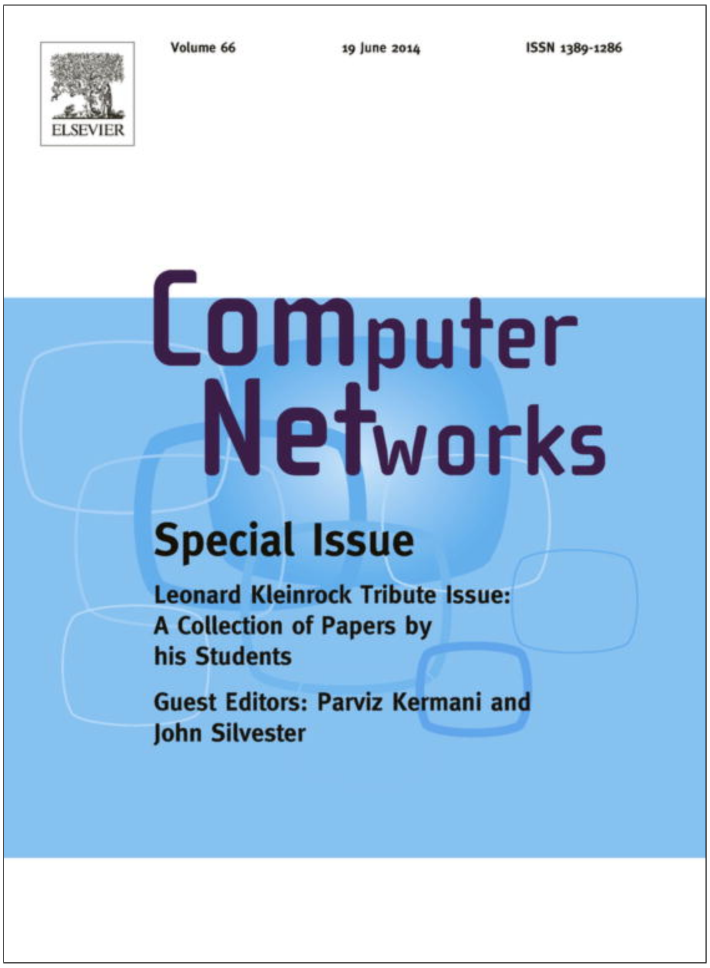
Volume 66
19 June 2014
ISSN 1389-1286
ELSEVIER
Computer Networks
Special Issue
Leonard Kleinrock Tribute Issue:
A Collection of Papers by
his Students
Guest Editors: Parviz Kermani and
John Silvester

Computer Networks 66 (2014) 82–93

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Queueing systems to study the energy consumption of a campus WLAN

Marco Ajmone Marsan [a,b,*], Michela Meo [a]

[a] Dipartimento di Elettronica e Telecomunicazioni, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
[b] IMDEA Networks Institute, Avenida del Mart Mediterraneo, 22, 28918 Leganes (Madrid), Spain

ARTICLE INFO

Keywords:
Energy efficiency
Wireless LANs
Queueing models

ABSTRACT

In this paper we exploit simple approximate queueing models to assess the effectiveness of the approaches that have been proposed to save energy in dense wireless local area networks (WLANs), based on the activation of access points (APs) according to the user demand. In particular, we look at a portion of a dense WLAN, where several APs are deployed to provide sufficient capacity to serve a large number of active users during peak traffic hours. To increase capacity, some APs are colocated and provide identical coverage; we say that these APs belong to the same *group*, and they serve users in the same area. The areas covered by different AP groups only partially overlap, so that some active users can only be served by a group of APs, but a fraction of active users can be served by more groups. Due to daily variations of the number of active users accessing the WLAN, some APs can be switched off to save energy when not all the capacity is needed. A real example of this setting is provided by a floor of one building of Politecnico di Torino in Italy, where a student library is located. The approximate analytical models indicate that the energy saving achievable with the proposed approaches is quite substantial, over 40% if at least one AP for each group is always kept on, even with no traffic, to be ready to accept incoming users, and it grows to almost 60% if all APs can be switched off at night, using a separate technology to activate an AP when the first user requests association in the morning.

1. Introduction

It first happened to our phones: initially we were accustomed to walk to them to answer a call, and to stay there for the entire conversation. We even found it natural not to be reachable while driving. Now, being tied to a cable for just a short call, makes us feel like dogs on leash, and we use our phones when reaching a friend's home, rather than ringing the door bell. Then it happened to our connections to the Internet: we are now addicted to reading email wherever we are: on a plane that just landed, in the elevator, in bed, stopping at a red light ... using our smartphones, tablets, laptops, phablets, and whatnot. This mutation from wired to wireless happened in spite of the unavoidable loss in performance, inherent in the poor characteristics of radio transmission, just because of convenience. We use our mobile(s) at work, even if the fixed phone is on the desk, and we prefer our laptop(s) to our desktop, even if downloads are slower. To mitigate this performance loss, large organizations have been consistently increasing the available bandwidth for mobile data users, by deploying more and more access points (APs) in their wireless local area networks (WLANs). This attitude has led to today's dense, centrally managed WLANs, where very large numbers of APs are installed, comparable to the

* Corresponding author at: Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy. Tel.: +39 011 0904032.
E-mail address: ajmone@polito.it (M. Ajmone Marsan).

http://dx.doi.org/10.1016/j.comnet.2014.03.012
1389-1286/

maximum number of simultaneously active users (some industrial campuses report over 10 thousand APs).

Even if the individual power consumption of APs is small, the fact that their number is very high implies a remarkable energy consumption and cost (10 W per AP, times 10 thousand APs, times 8760 h in a year, results in 876 MWh a year, with a cost of around 200 thousand euro, using the – admittedly high – kWh price in Italy). Most disturbing is the fact that the majority of this enegy is wasted, since the WLAN capacity that APs provide is not necessary 24/7; rather, capacity should be modulated according to the number of active users, or the traffic they generate, both of which vary widely over hours, days, and weeks; actually, the maximum WLAN capacity is necessary only at peak usage periods, for a limited time.

These considerations, coupled with the growing attention of the networking research community to energy efficiency, has led several research groups to investigate the energy saving which is possible with a smart management of WLAN resources, in particular by switching on and off APs according to the required capacity.

The first work in this field is [5], where the authors suggest that, in dense WLANs, APs can be clustered, based on their Euclidean distance. When the WLAN traffic (or the number of users) is low, only one AP in each cluster, the cluster-head, is switched on. When the traffic/number of users increases, additional APs can be switched on, to provide adequate capacity. Note that keeping track of the number of users associated with the APs in the cluster is easier (and more stable) than measuring the traffic load, so this quantity is preferred as a control variable in the dense WLAN central controller that turns on/off the cluster APs. The estimated energy saving can be 20–50% in less dense scenarios, whereas in more dense WLANs it can grow to 50–80%. An improvement of the AP clustering scheme was proposed in [6], by using the number and signal strength of the received beacons. This paper also suggested that a suitable metric to estimate user demand and drive the provided capacity can be the percentage of time the channel is busy due to transmission and interframe spacing.

A similar approach was proposed in [12] to reduce the number of switched-on APs, suggesting that it can be even possible to switch off all the APs in a cluster, provided the area covered by the cluster can be served by neighboring AP clusters. In this case the user demand is estimated from the number of users associated with APs, and energy savings are quantified in about 60%.

A first analytical model for the estimation of the energy gain achievable by modulating the number of switched-on APs as a function of the user demand was proposed in [2], considering just one AP cluster. Assuming as input a measured traffic trace, the analytical predictions indicate that the energy saving can be of the order of 40%.

In [10], the authors developed an ILP (Integer Linear Program) optimization model to adapt the number of switched-on APs to the number and the position of active users, achieving up to 63% power saving. In a subsequent paper [9], the same authors devised a heuristic approach which reduces the problem complexity, but also the achievable power saving.

A WLAN power saving scheme based on the maximum coverage problem was proposed in [3]. The proposed algorithm runs on a central controller, which collects the number of users associated with each AP and their data rates, and switches APs on and off dynamically, while maintaining coverage and guaranteeing user performance. Power saving of about 80% is reported at the expense of frequent user associations and significant delays.

The authors of [4] suggest that a drastic reduction of the density of APs in WLANs is possible, provided that the few APs remaining active can provide the coverage required to discover the presence of users. The key point here is that the detection of the user presence is possible with very limited AP coverage, and additional APs can then be switched on, to provide adequate capacity to users. Numerical evaluations show that up to 98% of APs can be switched off with this approach.

Other approaches exist, which are based on the presence of a secondary channel, that can be used to alert switched-off APs of the user presence. For example, the approaches proposed in [13–15] assume that all inactive APs can be switched off regardless of coverage. To react to user presence, an auxiliary low-power channel is available, that wakes the APs when users require access. The authors of [8] assumed that users are connected to a cellular network, which is capable of requesting the switch-on of a WLAN AP in the vicinity of the user. In [16] the presence of a secondary Bluetooth interface in both APs and mobile stations is assumed, and the Bluetooth interface can be used to request the activation of the WLAN APs.

In a previous paper [1] we considered a portion of a dense WLAN, where clusters of APs partially overlap in coverage, so that some active users can only be served by the APs of one cluster, but a fraction of active users can be served by APs in more than one cluster. We investigated the energy-performance trade-off with a model based on coupled queues: each queue represents a cluster of colocated APs, and customers arriving at the system might be served by one or several queues, depending on their position. We also proposed approximations based on single-queue and two-queue analysis, which provide fairly accurate performance estimates.

In this paper we expand on those single-queue approximate models, first showing that measurements of WLAN AP coverage validate the setting we consider, and that measurements of user activity justify the interest in energy saving approaches for a university campus WLAN. Then, we use simple queueing models, with and without setup times, to compare the cases in which the AP wake-up can or cannot exploit the presence of a secondary channel for the detection of user presence.

Queueing systems [7] have been for many years one of the most popular and versatile tools for the quantitative analysis and design of networks. The types of metrics traditionally derived from queueing system models are extremely diverse, ranging from very simple indicators, like the average number of waiting customers, or the average time in the queue, to more elaborate parameters, like, for example, server busy and idle period duration distributions. Of course, the metrics of interest depend on the type of system and on the objective of the quantitative analysis. Very little interest was paid in the past to metrics related to energy consumption. This is not due to a limitation in the modeling power of queues; rather, it reflects the lack of attention that in the past was paid to the energy characteristics of networks. Now, energy consumption has become an important element of the design space, and models are being developed with the objective of characterizing the energy properties of ICT systems of different types. As a result, energy-related metrics have started to appear in the context of queueing models.

This paper is organized as follows. Section 2 contains the problem statement, with the description of the considered portion of the dense WLAN. Section 3 presents the detailed queueing model, together with the considered customer and AP management algorithms. Section 4 describes the approximate queueing model and the computation of performance metrics. Section 5 presents and discusses numerical results obtained with the approximate model of the WLAN. Section 6 presents the variation of the approximate model that includes the AP activation time, and the results which it produces. Section 7 exploits the model results to quantify the energy saving achievable with the considered AP management scheme. Finally, Section 8 concludes the paper.

2. Problem statement

In this section, we describe the system we focus on, including the considered customer management (CMN) and AP management algorithms, and we present a detailed queueing model to study the system. For simplicity, we use the case of only two AP clusters in the description, but the extension to more clusters is trivial.

The most important terms of the notation used throughout the paper are summarized in Table 1.

2.1. The wireless LAN

We consider a portion of a dense WLAN in which an area $A = A_1 \cup A_2$ is covered by several access points (APs) in a manner such that a group (we use the term 'group' rather than 'cluster' because we assume that APs in a group are colocated, and provide identical coverage, over different frequency channels; in the literature, clusters define sets of APs which are not necessarily in the same physical location) G_1 of n_1 APs covers area A_1, and another group G_2 of n_2 APs covers area A_2, see Fig. 1 for a sketch of the system. Each AP can serve up to K mobile terminals (MTs); this means that no more than K MTs can be associated at the same time with the AP. APs can be switched on and off according to the number of MTs to be served. A fraction Ψ_1 of the total area A corresponds to $A_1 \cap \bar{A}_2$, so that the APs in group G_2 cannot serve MTs in this area. Similarly, a fraction Ψ_2 of the total area A corresponds to $A_2 \cap \bar{A}_1$, that cannot be served by the APs in group G_1. Finally, MTs located in area $A_1 \cap A_2 = A_0$, which corresponds to a fraction $1 - \Psi_1 - \Psi_2 = \Psi_0$ of the total area, can be served by any AP. APs can be activated and deactivated based on a given algorithm which is triggered by MT arrivals and departures. When an AP activation is triggered by the arrival of a MT, the AP switch-on time may translate into an increase of the time needed by the MT to associate with an AP. We will later look at the effect of this time.

As an example of this setting in a real WLAN, Fig. 2 shows the coverage of one floor of the building of Politecnico di Torino in which a portion of the main library and some study rooms are located. The top part of the figure refers to the estimated signal strength obtained by applying the multiwall model [11] with identical emitted power for the two AP groups; the bottom part of the figure reports direct measurements of the actual signal strength, and reveals that the APs do not actually emit the same power. Assuming a receiver threshold of -70 dBm, the number of AP groups that can be received in the different positions of the floor according to the multiwall model is depicted in Fig. 3. As can be seen, the coverage areas overlap for a large fraction, about 70% of the total floor space, meaning that, assuming a uniform distribution of the MTs, only 30% of the MTs need to be served by a specific AP group and this 30% is evenly divided among the two groups.

MTs are assumed to be uniformly distributed over the area, and to collectively request access according to a Poisson process with rate λ. The MTs remain associated with the AP for a time described by an exponentially distributed random variable with rate μ. Since the time a MT is associated with an AP depends on the user behavior and only very marginally on the system performance, the rate μ is

Table 1
Notation.

Symbol	Description
n_i	Number of APs in group i
K	Maximum number of MTs simultaneously associated with an AP
λ	Rate of association request
$1/\mu$	Mean association time
N_i	Maximum number of MTs associated with group i
α_i	Probability that a requests is originated from a MT in area A_i
ρ	System load
η	Energy consumption of an active AP [W]
ϵ	Energy needed for an AP activation/deactivation [J]
u_i	Number of unconstrained MTs in A_i
c_i	Number of constrained MTs in A_i
S	State space of the proposed model
γ_i	Approximate customer arrival rate to A_i
$1/\delta$	Mean AP switch-on time

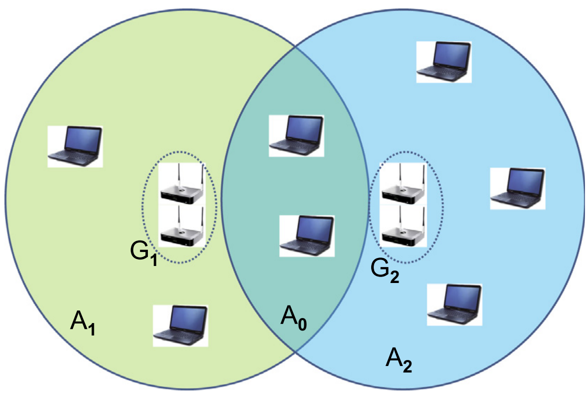

Fig. 1. Sketch of the considered portion of a dense WLAN.

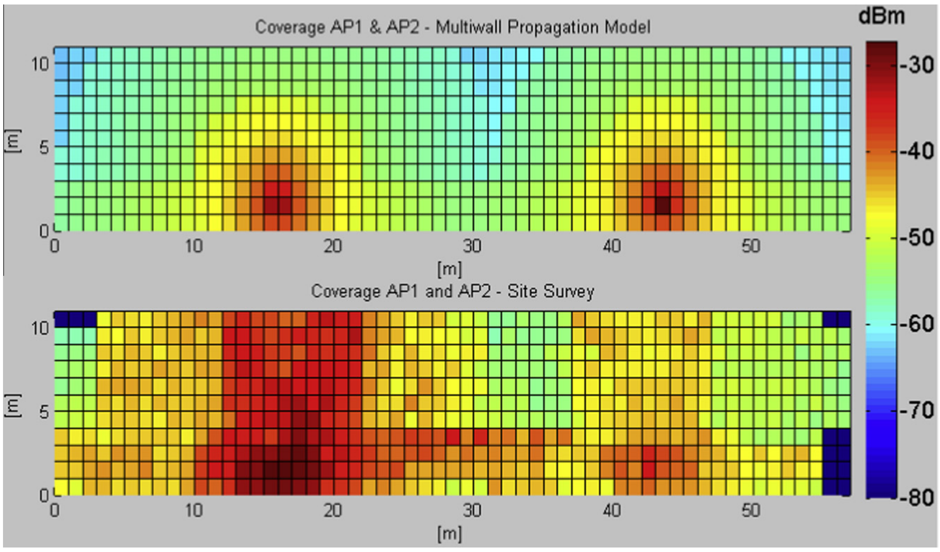

Fig. 2. Example of coverage map in a floor of a Politecnico di Torino building, multi-wall model (top) and measurement (bottom).

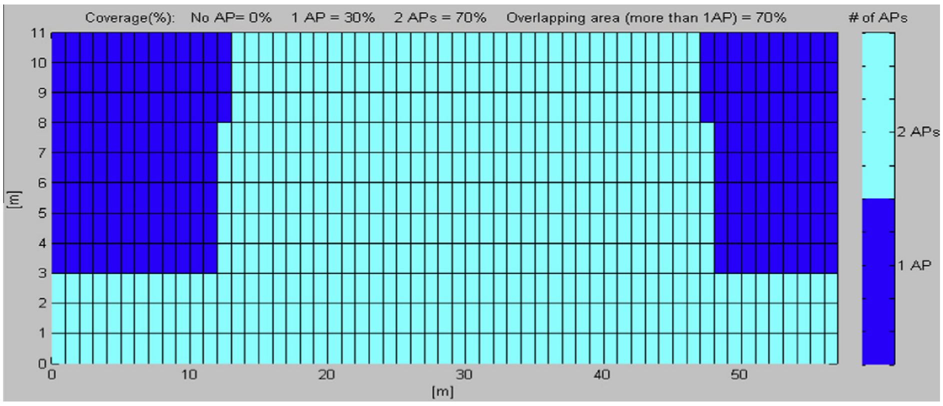

Fig. 3. Example of overlapping area in a floor of a Politecnico di Torino building.

independent on the number of MTs in the system. Indeed, the time a MT is associated with an AP is equivalent to the time a user is connected to a network and it is of the order of minutes or hours; thus, the possible effect of network performance on user's behavior can be considered negligible, and μ can be made independent on the number of MTs connected to the AP.

As an example of user behavior, we report the analysis of traces collected over the WLAN of Politecnico di Torino, which show quite significant variations over time, that justify the adoption of approaches for the provision of resources on demand. Fig. 4 reports the traffic measured during one day in different locations of the Politecnico campus, considering one AP per location type (library, classroom, etc.). Samples are taken every 5 min. The blue line corresponds to the traffic carried by an AP in the area of Fig. 2; the traffic is quite high because this area is continuously populated with students that access the study rooms. The pink lowest curve reports the case of an AP located inside a department, where most of the users have a wired connection and use the WLAN occasionally. The two intermediate curves refer, instead, to an AP in a public area and in a classroom; in both cases the traffic is quite high,

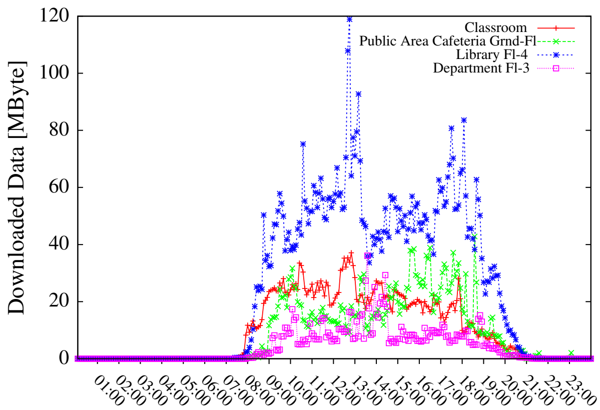

Fig. 4. Examples of the amount of traffic downloaded by APs located in different positions of the Politecnico campus; measurements are taken every 5 min.

with different behaviors during morning and afternoon, corresponding to the movements of students. Regardless the volume of traffic, the various cases share the same kind of day/night pattern. Similar behaviors can be derived by observing the number of active MTs. Fig. 5 reports the number of MTs that associate in one hour to anyone of the APs of the main campus in Politecnico (which hosts around 200 APs); the number of generated sessions in one hour is also shown. The day/night behavior is clear and evident, and justifies the use of resource on demand strategies that switch on the APs only when really needed, so that during no traffic periods (such as nights and holidays) energy can be saved by putting the unneeded APs to sleep.

2.2. The customer management algorithm

In [1] we considered several alternatives as regards the algorithm used to manage unconstrained customers, that correspond to MTs in A_0. The considered algorithms differ in two aspects: (i) the state information that is used to take decisions about MT association with APs: in the various algorithms, state information goes from the simple binary information about the system being full or not, to complete state, including, for all APs, the state and number of associated MTs and (ii) the kind of event that can trigger the algorithm: in some cases the algorithm is triggered by the request of a new MT association, in other cases both arrivals and departures, i.e., MT associations and de-associations, trigger the algorithm.

The results in [1] showed that the more complex algorithms bring only small performance improvements with respect to simpler ones, so that the conclusion was that the simplest customer management algorithm could be the wisest choice. In this paper we thus consider only the Random customer management algorithm, according to which each MT in A_0 that requests access, chooses the access point group G_1 or G_2 with equal probability. If the chosen group cannot accommodate the MT request, the MT is associated with the other group, if possible. If both groups are full with customers, the request is refused. This algorithm does not explicitly try to optimize energy consumption, allowing MTs to randomly choose the AP group. The only information required at MTs concerns the full/not full state of AP groups at arrival instants.

2.3. The AP management algorithm

The central controller of the dense WLAN collects results about the status of the different components of the WLAN, and in particular of the MT associations to APs, which allow it to control the switch-on and switch-off of APs belonging to different groups, so as to provide resources on demand to users, and save energy.

Ideally, a new AP must be switched on when a MT requires association to a group of APs where the capacity of the APs which are already on is fully used, and an AP of a group can be switched off when the number of MTs associated with the APs in the group can be handled with one less AP. To switch off an AP it is necessary that the MTs are gathered in the smallest number of APs possible; it is, thus, needed that MTs can hand-off to the desired AP.

Since the AP switch-on cannot be instantaneous, but requires switch-on times of the order of 1–2 min, it may be reasonable to introduce a threshold behavior, such that when j APs in a group are on, and the total number of MTs associated with them is $jK - T_h$, a new AP is switched on. The threshold T_h can be chosen according to the AP switch-on time, and the association request rate. Similarly, one of the APs can be switched off when the total number of MTs associated with the APs in the group is $jK - T_l$, with T_l possibly different from T_h, so as to provide a hysteresis which prevents an excessive frequency of AP switch on/off events.

3. The detailed queueing model

We model the portion of the dense WLAN with the queueing system shown in Fig. 6. The queueing system is composed of two multiserver stations with no waiting room. Each station models a group of APs. The service corresponds to the MT association with an AP. The two stations comprise N_1 and N_2 servers, respectively, where $N_i = Kn_i$ corresponds to the maximum number of MTs that can be associated with AP group i, that is composed of n_i APs, each one limited to host K MTs. As an example, Fig. 7 shows the case of station 1 with $n_1 = 2$ APs.

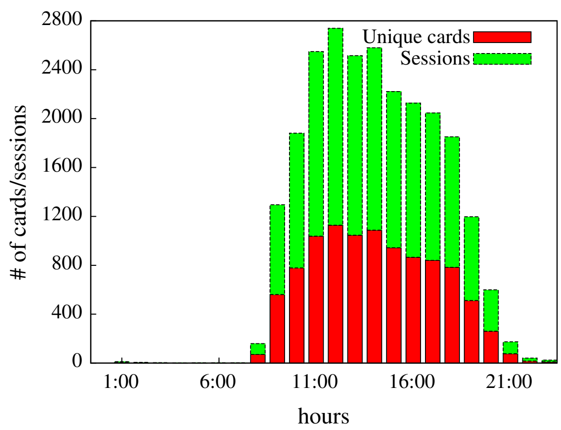

Fig. 5. Example of the number of active MTs (network interface cards) and active sessions measured in an hour in the whole main campus of Politecnico.

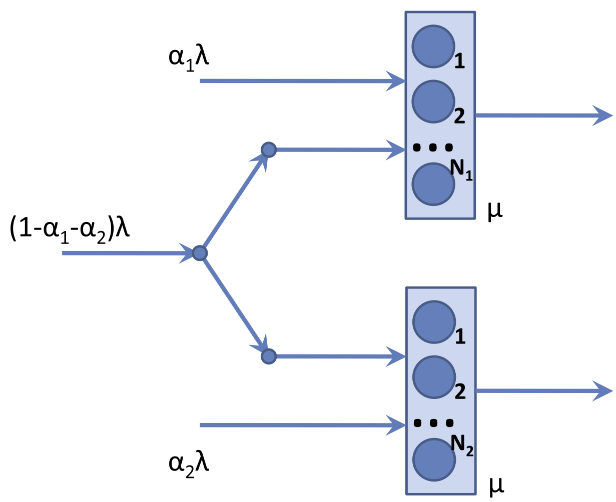

Fig. 6. Detailed queueing model of the system.

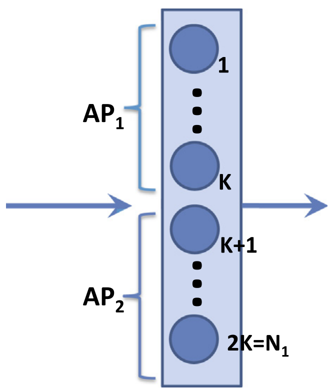

Fig. 7. Example of the queueing system modeling a group of APs.

MTs request service, i.e., they request to associate, according to a Poisson process with rate λ. This is the customer arrival process in the queueing model. With probability equal to α_1 an arriving customer is in area $A_1 \cap \overline{A}_2$ and it is, thus, constrained to proceed to station 1. By assuming uniform arrival rate in the area, the probability to arrive in A_1 is proportional to the size of the area, and it is thus $\alpha_1 = \Psi_1$. If the customer cannot be served by any AP in group 1, because the maximum number of MTs has already been reached (i.e., it cannot find an idle server), the customer is lost. The same occurs to group 2 with probability $\alpha_2 = \Psi_2$ (where $\alpha_1 + \alpha_2 < 1$). With probability $\alpha_0 = \Psi_0 = 1 - \alpha_1 - \alpha_2$, an arriving customer is unconstrained, i.e., it can be associated with either one of the two groups, according to the system CMN algorithm, and the customer is lost only when no idle server exists at either station. As an example, for the case presented in Figs. 2 and 3, in which 70% of the area is common, and the remaining is evenly distributed among the two groups, the model parameter setting corresponds to $\alpha_1 = \alpha_2 = 0.15$.

Service times are i.i.d. random variables with exponential distribution with rate μ. The system load can be defined as $\rho = \lambda/\mu$.

Each AP (and the corresponding K servers in the model) can be either *active* or *inactive*. Active servers can be either busy, i.e., serving a customer, or idle, i.e., ready to provide service to an incoming customer. In the model, the transition of an AP from inactive to active and vice versa corresponds to groups of K servers that, at a time, become idle or active. We assume that as soon as K active servers are idle, their corresponding AP is deactivated, and that if all active servers are busy when a customer arrives, K inactive servers (one inactive AP) are activated, if available. This implies setting $T_h = T_l = 0$.

We assume that each active AP (group of K servers) consumes energy at a rate η W, while an inactive AP (a group of K inactive servers) consumes negligible energy. The energy consumption may be a function of load, or of the number of active servers in the group; however, in this paper we always assume η to be constant, which is consistent with the characteristics of today's hardware. The activation or deactivation of an AP requires an energy equal to ϵ J.

The state of the queueing system is defined by a quadruple $\bar{s} = (u_1, c_1, u_2, c_2)$, with u_i the number of unconstrained customers at station i and c_i the number of constrained customers at station i. Considering that $u_i + c_i \leqslant N_i$, the cardinality of the state space S is given by:

$$| S |= \left(\frac{(N_1+1)(N_1+2)}{2}\right)\left(\frac{(N_2+1)(N_2+2)}{2}\right) \quad (1)$$

that is of the order of $(N_1N_2)^2$.

4. The approximate queueing model

As we can see from (1), the cardinality of the state space S of the queueing system grows roughly with the fourth power of the maximum number of users that can be served by a group of APs (assuming that the two groups can serve about the same number of users). If the number of AP groups is n, and each group can serve at most N users, the state space cardinality is of the order of N^{2n}. This means that the Markovian model that underlies the queueing system suffers from a combinatorial explosion of the number of states for both growing number of users and growing number of AP groups.

In order to tackle the problem of very large state spaces, in [1] we introduced approximate models that provide accurate estimates of the performance metrics, while generating state spaces of much smaller size than the detailed model discussed so far, and even allowing in some case a closed form expression for the limiting probabilities.

The simplest approximation is based on considering just one of the two queues in our system, say queue 1. The idea behind this approach is to approximate the customer arrival rate γ_1 at queue 1 from the area $A_0 = A_1 \cap A_2$, independently of the state of the other queue (queue 2). The performance metrics of interest are then computed for queue 1, independently from the state of queue 2.

The customer arrival rate in A_0 is equal to $\alpha_0\lambda$. According to the Random CMN algorithm, this rate is equally split between the two queues, so that queue 1 receives from this area a customer flow at rate $\alpha_0\lambda/2$. However, queue 1 also receives all customers that arrive in area A_0 when queue 2 is full, so that we must also account for an additional customer arrival rate equal to $P[f_2]\alpha_0\lambda/2$, where $P[f_2]$ is the probability that queue 2 is full. So, we can write:

$$\gamma_1 = (1 + P[f_2])\frac{\alpha_0\lambda}{2} \quad (2)$$

The total arrival rate at queue 1 is then:

$$\lambda_1 = \alpha_1\lambda + (1 + P[f_2])\frac{\alpha_0\lambda}{2} \quad (3)$$

Assuming that the two queues are equal (same number of APs, same size of the served area, same maximum number of users per AP), or very similar, we can approximate the value of $P[f_2]$ with the value $P[f_1]$. If the parameters of the two queues are indeed identical, we only introduce an error deriving from the fact that we are assuming independence in the behavior of the two queues. If the parameters of the two queues differ, an additional error is introduced by assuming $P[f_2] = P[f_1]$.

For the model solution, in this case we must use an iterative algorithm, because of the circular dependence of the loss probability on the arrival rate, and of the arrival rate on the loss probability. At each step of the iteration, the probability distribution vector $\vec{\pi}$ is computed, as well as $P[f_1]$, for a given value of the customer arrival rate λ_1. The new value of λ_1 is then computed from $P[f_1]$, and a new iteration is run. The algorithm stops when either a maximum number of iterations is reached, or the relative error between the values of $P[f_1]$ at two consecutive steps of the iteration is smaller than a given tolerance. The performance measures of interest are computed from the final probability distribution vector.

Note that the queue we are using in this approximation is an $M/M/N_1/0$, with arrival rate equal to λ_1, so that the state space cardinality is simply $N_1 + 1$, and the limiting probabilities can be expressed in closed form as:

$$\pi_{i,1} = \frac{\left(\frac{\lambda_1}{\mu}\right)^i \frac{1}{i!}}{\sum_{k=0}^{N_1} \left(\frac{\lambda_1}{\mu}\right)^k \frac{1}{k!}} \tag{4}$$

for $i = 0, \ldots, N_1$, and

$$P[f_1] = \pi_{N_1,1} \tag{5}$$

Hence, the iterative algorithm is extremely fast.

Note that the limiting distribution for this queue is invariant with respect to the service time distribution.

This approach can be easily extended to cases comprising more than two queues.

In [1] we validated this approximate model, and we showed that it can be quite accurate for symmetrical settings, like the one presented in Fig. 3. We also presented and validated other, more complex approximations, based on several separate queues, which are better suited to asymmetrical cases.

In what follows, we will report only the results obtained with the simplest approximation, in which one queue is used to model a group of APs. We will denote by λ the overall arrival rate at the queue, by μ the customer service rates and by π_i the probability that there are i customers in the queue, i.e., that i MTs are associated with the group of APs. Moreover, N will be the maximum number of MTs that can associate with a group that is composed of n APs, so that $N = nK$.

4.1. The performance metrics

For the WLAN and the queueing system we introduce the following performance metrics:

- average *number of customers* at each station (average number of MTs associated with the AP group),
- average *utilization* of the APs,
- average *power consumption* of the APs,
- average *energy consumed to serve one customer*,
- average and cumulative distribution function (CDF) of the *active time* of an AP, i.e., the time between the instant in which an AP is switched on and when it is switched off again.

When i customers are in the station, the amount of MTs associated with AP x of the group is given by:

$$c_x(i) = \begin{cases} 0 & \text{for} \quad i \leqslant (x-1)K \\ n \bmod K & \text{for} \quad (x-1)K < i < xK \\ K & \text{for} \quad i \geqslant xK \end{cases} \tag{6}$$

The average number of customers at AP x can then be computed as:

$$E[N_x] = \sum_{i=1}^{N} c_x(i)\pi_i \tag{7}$$

The average utilization of AP x can be computed by:

$$E[U_x] = \sum_{i=(x-1)K+1}^{N} \pi_i \tag{8}$$

The average power consumption can be computed as:

$$E[P] = E[P_s] + E[P_{act}] + E[P_{de}] \tag{9}$$

where $E[P_s]$ is the power consumed by active APs, and $E[P_{act}], E[P_{de}]$ are the powers consumed for server group activation and deactivation, respectively. We have:

$$E[P_s] = \eta \sum_{i=1}^{N} \lceil i/K \rceil \pi_i \tag{10}$$

where the term in the sum represents the number of active APs.

The power to activate and deactivate an AP is given by the product of the energy needed to activate and deactivate an AP multiplied by the frequency with which this happens. Thus:

$$E[P_{act}] = \epsilon\lambda \sum_{x=0}^{n-1} \pi_{xK} \tag{11}$$

$$E[P_{de}] = \epsilon\mu \sum_{x=0}^{n-1} (xK+1)\pi_{xK+1} \tag{12}$$

since in state xK the arrival of a customer with rate λ causes the activation of AP $x+1$, and in state $xK+1$ the departure of a customer, event that occurs with rate $(xK+1)\mu$, causes the switch off of AP x.

The average energy consumed to serve one customer can be computed as:

$$E[E_c] = \frac{E[P]E[T]}{E[N]} = \frac{E[P]}{\lambda(1 - P[loss])} \tag{13}$$

where $E[T]$ is the average time spent by a customer in the queueing system, which is computed by Little's Law as:

$$E[T] = \frac{E[N]}{\lambda(1 - P[loss])} \quad \text{with} \quad E[N] = \sum_{i=1}^{N} i\pi_i \tag{14}$$

The AP active time can be computed from the first passage time between the states in which activation and deactivation can occur.

5. Numerical results

In this section we discuss numerical results derived from the solution of the approximate queueing model. We present, in particular, an analysis of the consumed power and of the AP active times so as to derive insight into the effectiveness of using AP activation and deactivation as a way to reduce energy consumption.

We consider the case of 3 APs per group, with the maximum number of associated MTs per AP, $K = 8$. We set the time that each MT remains associated with an AP to be exponentially distributed with mean equal to 10 min, and we vary λ. The load is denoted by $\rho = \lambda/\mu$. In what follows, we call AP1 the first AP to be switched on, i.e., the one which handles the first K MTs associated with APs in the group, AP2 the second AP to be switched on, i.e., the one which handles the K MTs in excess of the K handled by AP1, and AP3 the one that handles the MTs that are associated with the group in excess of $2K$.

The energy consumption per time unit for each access point is assumed to be constant, and is set to $\eta = 10$ W. The energy spent for an AP activation or deactivation is neglected, i.e., we set $\epsilon = 0$ J.

Fig. 8 shows the average number of MTs associated with the three APs versus the load of the AP group. As expected, the average number of MTs associated with AP1 is larger than for AP2, which in turn is larger than for AP3. The number of MTs associated with AP2 is negligible for load less than 3, and the number of MTs associated with AP3 is negligible for load less than 8. Considering meaningful loads to be lower than 24 (the number of servers) we see that the maximum average utilizations of APs are around 7 for AP1 and AP2, and around 6 for AP3, for a total of about 20 MTs in the AP group on average.

Fig. 9 shows the average utilization of AP1, AP2, and AP3, as defined before, as well as the curve of the AP group being idle (no MT associated with any AP). If the APs in the group behave ideally, i.e., if they can be activated and deactivated in zero time according to the number of MT associations (no active AP if no MT is associated with APs in the group, 1 active AP if the number of associated MTs is between 1 and K, 2 active APs between $K+1$ and $2K$, and 3 active APs beyond $2K$), then the average number of active APs for each load value is given by the sum of the values of the three AP curves, and the power consumption in W is obtained by just applying a factor η.

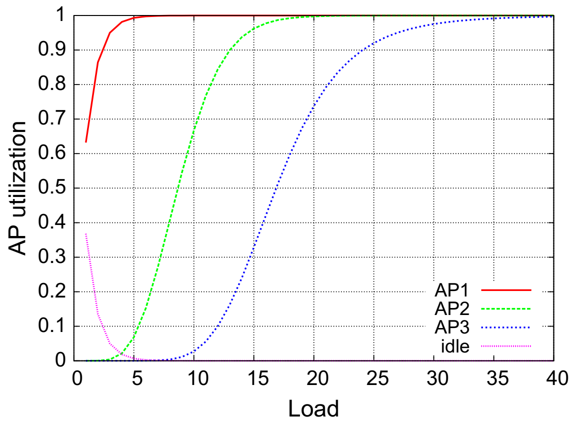

Fig. 9. Average AP utilization versus load.

Fig. 10 reports the average active time for the three APs versus the AP group load. Clearly, when the load is low, say below 6, one active AP is sufficient: the number of associated MTs is usually smaller than $K = 8$, and it is very unlikely that the second AP (AP2) switches on. In case this happens, the time that AP2 remains active is very small. When the load grows, say between 6 and 20, the typical number of associated MTs is larger than 8, so that AP1 is almost always on (this translates in extremely long active times) and AP2 is on for increasingly long periods, while AP3 switches on rarely, and when this happens it remains active for very short times only. For even higher values of the AP group load (up to 40) AP3 exhibits increasingly long active times, up to about 2 h.

The CDF of the active time of AP1 is shown in Fig. 11 for different (but low) values of the AP group load. The CDF reflects the behavior observed above. For example, when the load is very low, equal to 1, 95% of the times that AP1 switches on, it remains active for less than 1 h. When the load is higher, for example equal to 5, the active time is much longer, and only 35% of the times it is shorter than 1 h. In most cases, the probability that the active time of AP1 is shorter than 5 min is less than 20%. The fact that, even for very low loads, the probability of short active times is small, makes the AP switch-on times (that we

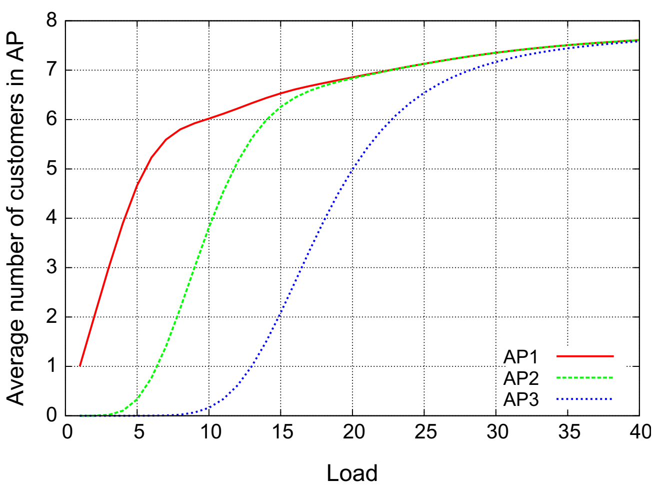

Fig. 8. Average number of MTs associated with each AP versus load.

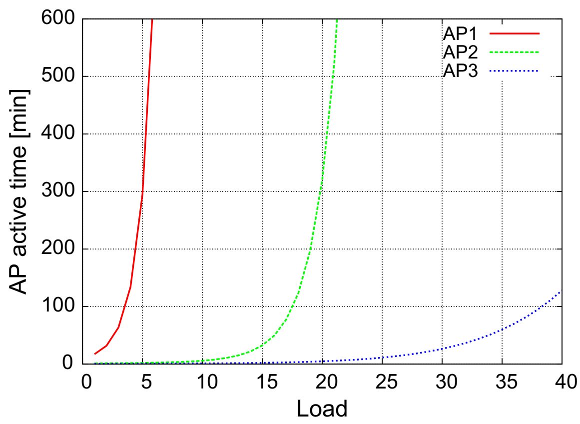

Fig. 10. Average active time for the three considered APs versus load.

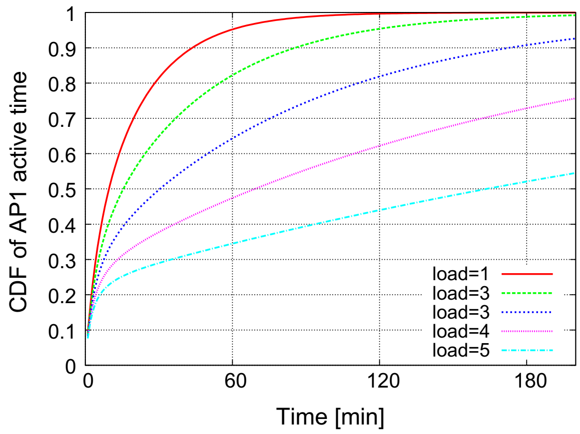

Fig. 11. Cumulative distribution of the active time of AP1 under various values of the load.

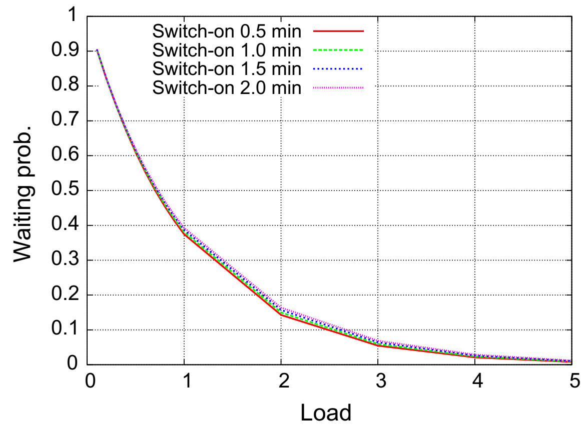

Fig. 12. Probability that a user has to wait that AP1 switches on, versus the AP group load, for variable average switch-on time.

observed in experimental setups to be of the order of 1–2 min) largely irrelevant for the overall system performance.

Nevertheless, in the next section we modify our approximate model to account for the switch-on time of AP1.

6. Model with AP1 activation time

The model that we have presented in the previous sections is based on the assumption that an AP can be switched on or off instantaneously, according to the number of MTs that associate with the AP group. This assumption is justified by the large ratio between AP active times and switch on/off times, as well as the possibility of introducing thresholds to anticipate the activation of APs when the capacity of the active APs is close to exhaustion. In this section we nevertheless investigate the impact of the switch-on time of AP1, assuming that all the APs that cover a given area are switched off in periods of no traffic. This implies the existence of a secondary method for the identification of user presence, so that AP1 is switched on for example in the morning when the first student entering the study room switches on her/his PC.

This amounts to modifying the queueing model discussed in the previous sections by introducing a delay between the arrival of the customer that finds an empty queue, and the beginning of service. If the delay can be assumed to be exponentially distributed, with rate δ, the continuous-time Markov chain (CTMC) model underlying the queueing system requires that in the state definition, in addition to recording the number of associated MTs, memory is kept about the state of AP1 (active or not). Let the state be given by the pair of values (i, s), where i is the number of customers in the system and $s \in \{0, 1\}$ represents the service being available ($s = 1$ means AP1 active) or unavailable ($s = 0$ means AP1 inactive). When the queue is empty, at the arrival of the first customer, the CTMC moves to state $(1, 0)$ in which AP1 is being activated and the customer is waiting for service. Further arrivals can occur during the activation period, with no service in progress. When AP1 finally activates, if there are i customers waiting, the CTMC moves with rate δ from $(i, 0)$ to $(i, 1)$ and services start. The state space cardinality is equal to $2N + 1$, where N is the maximum number of MTs that can be associated with the AP group.

It is also possible to assume a constant switch-on time (which can be more realistic). In this case, the queueing model can no longer be described with a CTMC, but the semi-Markov process underlying the queueing model can still be analyzed by using an embedded discrete-time Markov chain (DTMC). In the embedded chain, the idle state represents the cases in which no service is in progress. The exit from the DTMC idle state corresponds to the AP1 activation, and occurs after the arrival of the first MT followed by the constant activation time; out of the idle state the DTMC moves towards any possible active state, depending on the number of arrivals that occurred during the activation time. In this case, the state space of the DTMC comprises $N + 1$ states.

The results that we discuss next refer to the case in which the AP activation time is either exponentially distributed, with average comprised between 0.5 and 2 min, or constant, with the same range of values.

Fig. 12 shows curves of the probability that a user has to wait that AP1 switches on, versus the AP group traffic load, for values of the average switch-on time that very between 30 s and 2 min. For the same cases, Fig. 13 shows the average time that a user has to wait. We can see that the probability that a user has to wait that AP1 switches on is almost invariant with respect to the average switch-on time, and that the probability is high only for very low traffic values. We also note that the average time that a user has to wait is less than 10 s for all traffic load values larger than 3.

The next two figures, Figs. 14 and 15, compare the cases of exponentially distributed and constant AP switch-on times. The waiting probabilities computed using the two models are almost identical, while the average waiting times are slightly shorter in the (more realistic) case of constant switch-on times, as expected.

7. Energy saving

In this section we exploit the approximate models presented in this paper to assess the effectiveness of AP

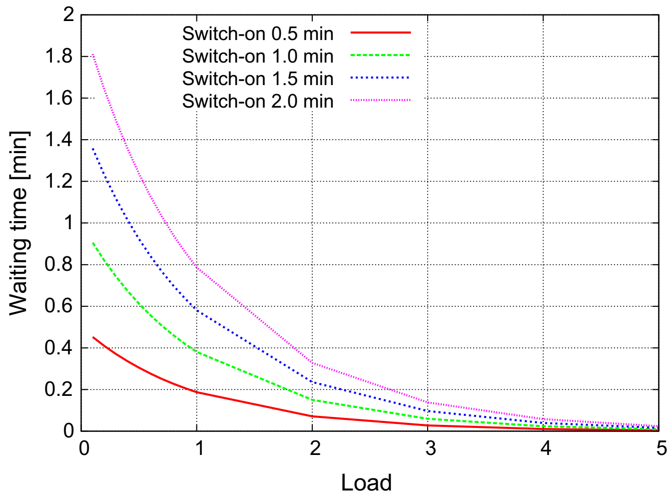

Fig. 13. Average time that a user has to wait that AP1 switches on, versus the AP group load, for variable average switch-on time.

management schemes by focusing on the achievable reduction in AP energy consumption. In order to consider

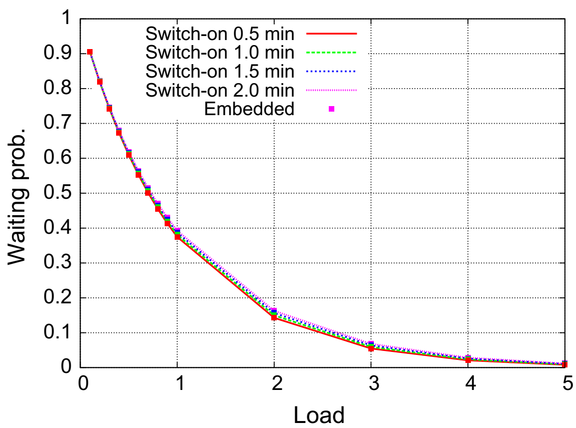

Fig. 14. Model with AP activation time, comparison between the CTMC and embedded MC model: average time that a user has to wait that AP1 switches on versus load.

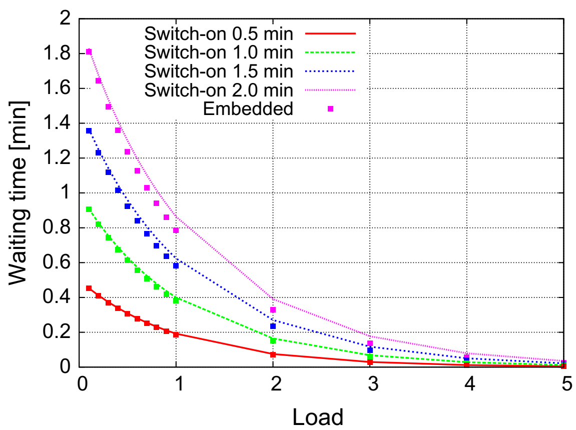

Fig. 15. Model with AP activation time, comparison between the MC and embedded MC model: average time that a user has to wait that AP1 switches on versus load.

a realistic setting, we take as input a daily traffic pattern obtained from the one reported in Fig. 4, scaling it down to be representative of the considered area, which comprises two groups of APs with 3 APs each. In particular, we assume that, at the peak hour, the traffic load is $\rho = N$. For each value of the hourly load we derive the steady-state solution of the approximate model (with no AP activation time) and we compute the energy consumption of the AP group in three cases: (i) the 3 APs are always on; (ii) one AP is always on to be ready to serve MTs entering the area, and the other two are turned on only when necessary; and (iii) all APs can be switched off when no MT requests service in the considered area, since a secondary method is available to detect the arrival of MTs requesting association.

Fig. 16 shows the average energy consumption of each one of the 3 APs for each hour, under the considered daily traffic pattern. We can see that no energy is necessary during the night, since no MT is present in the area covered by the AP group, and that the periods of activity are different for the 3 APs.

The results in Fig. 16 can be combined to generate the total energy consumption of the AP group in the three

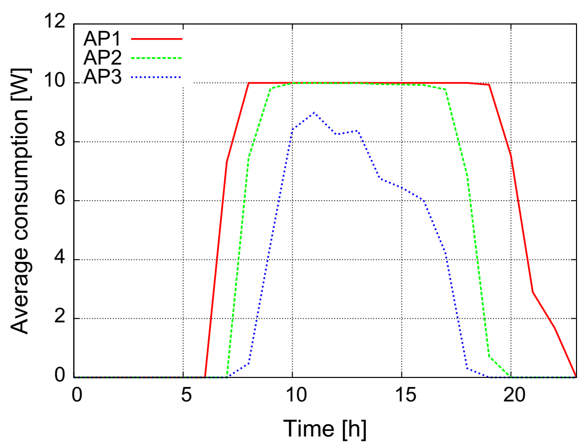

Fig. 16. Average consumption of each AP in a typical day.

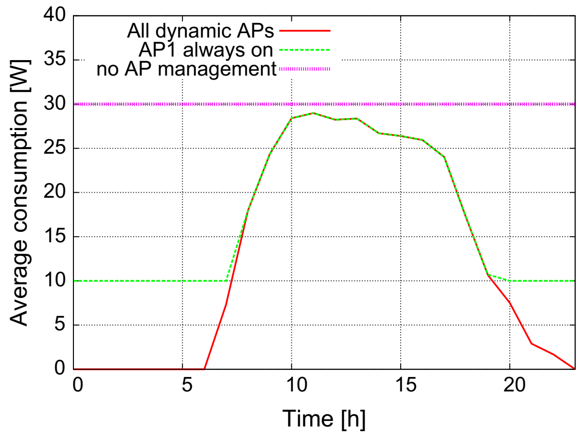

Fig. 17. Average total consumption in a typical day, with and without coverage continuity.

cases. We see in Fig. 17 that if we assume that the 3 APs are always on, the AP group power consumption is constant, equal to 30 W. If instead AP1 is always on, but the other two APs are turned on only when necessary, the power consumption is 10 W from 7 pm to 7 am, and is following the daily traffic pattern in the other half of the day. Finally, if all APs can be switched off when no MT requests service in the considered area, no power is consumed by the AP group from midnight to 6 am, and the power consumption is proportional to traffic for the rest of the day.

If we integrate the power consumption to compute the energy absorbed from the grid, which translates into cost for the network operator, we see that if we assume that the 3 APs are always on, the total daily energy consumption is equal to 0.72 kWh (about 263 kWh a year). If AP1 is always on, and the other two APs are turned on only when necessary, the total daily energy consumption is equal to 0.41 kWh. If all APs can be switched off when no MT requests service in the considered area, the total daily energy consumption is equal to 0.31 kWh. This implies a saving of 43% if AP1 is always on, and of 57% if AP1 is switched off in the periods of no traffic. It should be noted that the increased saving comes at a cost, since it requires

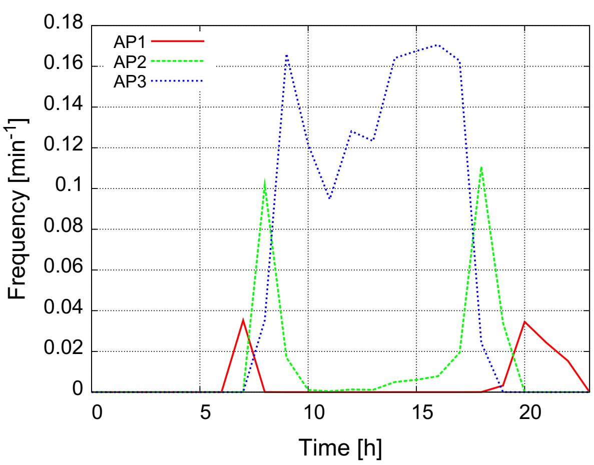

Fig. 18. Switch on frequency for each AP in a typical day.

the existence of a technology (possibly a secondary – low power – radio channel) to detect the presence of users requesting association when no AP is active.

Finally, in Fig. 18 we show, for the same setting, the frequency at which APs switch on during each hour of the day. We see that AP1 switches on only when traffic is very low: in the early morning and late evening. AP2 switches on later in the morning and earlier in the evening. AP3 instead switches on (and off) during most of the day, following the fluctuations in the number of users that request association in the considered area. By looking at the integral of the switch-on rate, we can conclude that the average number of switch-on/off for AP1 is lowest, and the one for AP3 is highest.

8. Conclusions

Sleep modes for APs are being studied by several research teams as a promising approach to improve the energy efficiency of WLANs. They can be specially effective in dense centrally-managed WLANs, where a large number of APs is deployed to provide high capacity, which is however necessary only at peak times, for a short portion of the day. In this paper, we presented simple queueing models to assess the effectiveness of AP sleep modes in dense WLANs, using the setting of one floor of the Politecnico di Torino campus as an example. We computed several different performance and energy parameters, showing that the energy saving achievable with the proposed approaches is quite substantial, of the order of 43% if at least one AP for each group always remains on, even with no traffic, to be ready to accept incoming users, and it grows to about 57% if all APs can be switched off at night, thanks to a separate technology to activate an AP when the first user requests association in the morning.

The main insight provided by our analysis is the following. First, we proved that the energy saving potential is quite substantial: more than half of the energy used to power on a campus WLAN can be saved. Second, we showed that the impact of AP switch-on/off times is negligible, since typical on and off periods are much longer (at least one order of magnitude on average, for reasonable traffic load) than the transient duration. Third, we showed that the added benefit of a secondary technology to wake up sleeping APs, thanks to which all APs can be switched off during long periods of no traffic (specially night) is important (almost 15% of the total energy).

Finally, we showed that, even in new networking scenarios, queueing models [7] remain a simple and effective tool for the investigation of the system behavior and for the selection of the most beneficial operating modes.

Acknowledgements

The research leading to these results has received funding from the EU 7th Framework Programme (FP7/2007-2013) under Grant Agreement n. 257740 (NoE TREND).

We would like to thank Christian Tipantuna for the work on the coverage maps and propagation models; Fikru

Getachew, Nanfang Li, Marco Ricca and Yi Zhang for providing the plots with measurement results.

References

[1] Marco Ajmone Marsan, Ana Paula Couto da Silva, Michela Meo, Energy-performance trade-off in dense WLANs: a queuing study, Elsevier Comput. Networks 56 (11) (2012) 2522–2537.
[2] Marco Ajmone Marsan, Luca Chiaraviglio, Delia Ciullo, Michela Meo, A simple analytical model for the energy-efficient activation of access points in dense WLANs, in: 1st International Conference on Energy-Efficient Computing and Networking, June 2010, pp. 159–168.
[3] Kwan-Wu Chin, A green scheduler for enterprise WLANs, in: Australasian Telecommunication Networks and Applications Conference (ATNAC '11), 2011, pp. 1–3.
[4] Fatemeh Ganji, Lukasz Budzisz, Adam Wolisz, Assessment of the Power Saving Potential in Dense Enterprise WLANs, Technical Report TKN-13-003, TKN Group, TU Berlin, April 2013.
[5] Amit P. Jardosh, Gianluca Iannaccone, Konstantina Papagiannaki, Bapi Vinnakota, Towards an energy-star WLAN infrastructure, in: 8th IEEE Workshop on Mobile Computing Systems and Applications (HotMobile '07), March 2007, pp. 85–90.
[6] Amit P. Jardosh, Konstantina Papagiannaki, Elizabeth M. Belding, Kevin C. Almeroth, Gianluca Iannaccone, Bapi Vinnakota, Green WLANs: on-demand WLAN infrastructures, Mobile Networks Appl. 14 (6) (2009) 798–814.
[7] Leonard Kleinrock, Queueing Systems: Volume I – Theory, Wiley Interscience, New York, 1975.
[8] SuKyoung Lee, SungHoon Seo, Nada Golmie, An efficient power-saving mechanism for integration of WLAN and cellular networks, IEEE Commun. Lett. 9 (12) (2005) 1052–1054.
[9] Josip Lorincz, Massimo Bogarelli, Antonio Capone, Dinko Begusic, Heuristic approach for optimized energy savings in wireless access networks, in: International Conference on Software, Telecommunications and Computer Networks (SoftCOM '10), 2010, pp. 60–65.
[10] Josip Lorincz, Antonio Capone, Massimo Bogarelli, Energy savings in wireless access networks through optimized network management, in: 5th International Symposium on Wireless Pervasive Computing (ISWPC '10), May 2010, pp. 449–454.
[11] Ingo Forkel, Matthias Lott, A multi wall and floor model for indoor radio propagation, in: IEEE Vehicular Technology Conference (VTC 2001-Spring), May 2001.
[12] Riichiro Nagareda, Akio Hasegawa, Tatsuo Shibata, Sadao Obana, A proposal of power saving scheme for wireless access networks with access point sharing, in: International Conference on Computing, Networking and Communications (ICNC '12), 2012, pp. 1128–1132.
[13] Suhua Tang, Hiroyuki Yomo, Yoshihisa Kondo, Sadao Obana, Exploiting burst transmission and partial correlation for reliable wake-up signaling in Radio-On-Demand WLANs, in: IEEE International Conference on Communications (ICC '12), 2012, pp. 4954–4959.
[14] Suhua Tang, Hiroyuki Yomo, Yoshihisa Kondo, Sadao Obana, Wake-up receiver for radio-on-demand wireless LANs, EURASIP J. Wireless Commun. Network. 2012 (1) (2012) 1–13.
[15] Hiroyuki Yomo, Yoshihisa Kondo, Kosuke Namba, Suhua Tang, Takatoshi Kimura, Tetsuya Ito, Wake-up ID and protocol design for radio-on-demand wireless LAN, in: IEEE 23rd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC '12), 2012, pp. 419–424.
[16] Jong-Woon Yoo, Kyu Ho Park, A cooperative clustering protocol for energy saving of mobile devices with WLAN and Bluetooth interfaces, IEEE Trans. Mobile Comput. 10 (5) (2011) 491–504.

Marco Ajmone Marsan holds a double appointment as Full Professor at the Department of Electronics and Telecommunications of the Politecnico di Torino (Italy), and Research Professor at IMDEA Networks Institute (Spain).

He earned his graduate degree in Electrical Engineering from the Politecnico di Torino in 1974 and completed his M.Sc. in Electrical Engineering at the University of California at Los Angeles (USA) in 1978. In 2002, he was awarded a "Honoris Causa" Ph.D. in Telecommunication Networks from the Budapest University of Technology and Economics.

From 2003 to 2009 he was Director of the IEIIT-CNR (Institute for Electronics, Information and Telecommunication Engineering of the National Research Council of Italy). From 2005 to 2009 he was Vice-Rector for Research, Innovation and Technology Transfer at Politecnico di Torino.

He is involved in several national and international scientific groups: He was Chair of the Italian Group of Telecommunication Professors (GTTI); the Italian Delegate in the ICT Committee and in the ERC Committee of the EC's 7th Framework Programme. He is a Fellow of the IEEE and he is listed by Thomson-ISI amongst the highly-cited researchers in Computer Science.

He has been principle investigator for a large number of research contracts with industries, and coordinator of several national and international research projects.

Michela Meo received the Laurea degree in electronic engineering in 1993, and the Ph.D. degree in electronic and telecommunications engineering in 1997, both from the Politecnico di Torino, Italy. Since November 2006, she is associate professor at the Politecnico di Torino. She coauthored almost 200 papers, about 60 of which are in international journals. She edited six special issues of international journals, including ACM Monet, Performance Evaluation, and Computer Networks. She was program cochair of two editions of ACM MSWiM, general chair of another edition of ACM MSWiM and of IEEE Online GreenComm, program co-chair of the IEEE QoS-IP, IEEE MoVeNet 2007, and IEEE ISCC 2009, IEEE Online GreenComm 2012, IEEE Infocom Miniconference 2013, and she was in the program committee of about 100 international conferences, including SIGMETRICS, INFOCOM, ICC, and GLOBECOM. Her research interests include the field of performance evaluation and modeling, green networking and traffic classification and characterization.