Modeling the Impact of Start-Up Times on the Performance of Resource-on-Demand Schemes in 802.11 WLANs

Jorge Ortín, Pablo Serrano, Carlos Donato

***Abstract*—Resource on Demand in 802.11 Wireless LANs is receiving an increasing attention, with its feasibility already proved in practice and some initial analytical models available. However, while these models have assumed that access points (APs) start up in zero time, experimentation has showed that this is hardly the case. In this work, we provide a new model to account for this time in the simple case of a WLAN formed by two APs where the second AP is switched on/off dynamically to adapt to the traffic load and reduce the overall power consumption, and show that it significantly alters the results when compared to the zero start-up time case, both qualitatively and quantitatively.**

***Index Terms*—WLAN, 802.11, Resource on Demand, Energy Consumption**

I. Introduction

One of the most effective techniques to cope with the growing traffic demand in wireless networks is to deploy more access points (APs), thus reducing the per-cell coverage and facilitating spectrum re-use. This technique, though, challenges energy-efficient operation, as a deployment planned for a high traffic load results in a huge wastage of energy at a low load if all the infrastructure is kept powered on.

To achieve energy efficient operation in very dense scenarios, the network has to implement a Resource-on-Demand (RoD) scheme by which APs are activated as the demand grows and deactivated as it shrinks. Given that, in general, mobile networks are carefully planned, owned by a single operator, and consist of equipment with very high energy demands (and, correspondingly, high energy bills), it comes to no surprise that most of the research so far in RoD has focus on the case of cellular networks [1]. For the case of Wireless LAN (WLAN), though, only a few works have addressed the problem of RoD [2]–[4].

In [2], authors demonstrate the feasibility and potential savings of RoD for 802.11 WLANs with "Survey, Evaluate, Adapt, and Repeat" (SEAR), a RoD framework based on heuristics that opportunistically powers on and off APs while maintaining coverage and user performance. In contrast to this experimental-driven approach, in [3] authors present the first analytical model for RoD, focusing on the case of "clusters" of APs (i.e., devices with overlapping coverage areas) and analyzing the impact of the strategy used to (de)activate on parameters such as the energy savings and the switch-off rate of the devices. In [4], authors extend the work of [3] to analyze the case when APs do not completely overlap their coverage areas, to understand the trade-offs when e.g. (re)associating clients from one AP to another AP in order to power down the former.

From (Power)	To (Power)	Time
OFF (0 W)	ON (2.7 W)	45 s
ON (2.7 W)	OFF (0 W)	3 s

TABLE I
TIME REQUIRED TO SWITCH FROM THE ON STATE TO THE OFF STATE (AND VICE-VERSA) IN A LINKSYS WRT54GL.

In both analytical works [3], [4], as well as in a recent follow-up analysis [5], among other simplifying assumptions, authors neglect the time required to power on an AP. However, in [2] it is reported that typical start-up times range between 12 and 35 seconds. To confirm these results, we perform an experimental characterization of the power consumed by a Linksys WRT54GL router running OpenWRT 10.03.1, which is a very popular wireless router that has been widely deployed, also measuring the average time required to power it on (i.e., the device starts broadcasting the SSID) and to power it off (i.e., no SSID is broadcasted). The results are provided in Table I. As our results confirm, these times are far from negligible, in particular when compared against inter-arrivals and/or service times. In this work we revisit this assumption and assess its impact on performance.

More specifically, in this work we address the problem of modeling the time required to start-up an AP in a RoD scenario. We consider the case of a network with two overlapping APs and show that, even in this simple scenario, considering the start-up times alters both qualitatively and quantitatively the results, as compared to the case of "immediate" boot times. Our analysis is validated by extensive event-driven simulations, which confirm the validity of the model for a variety of scenarios.

II. System Model

Our system is a simplified version of the *cluster model* analyzed in [3], consisting of two identical APs serving the same area. One of the APs is always on, in order to maintain the WLAN coverage, while the other AP is opportunistically powered on (off) as users arrive (leave) the system. However, in contrast to the model in [3], powering on the second AP takes T_{on} units of time; during this time, the second AP is not available and arriving requests are served by the first AP. Each AP consumes P_{AP} units of power when active (i.e., during start-up and when powered on) and 0 otherwise. Although

978-3-901882-70-8

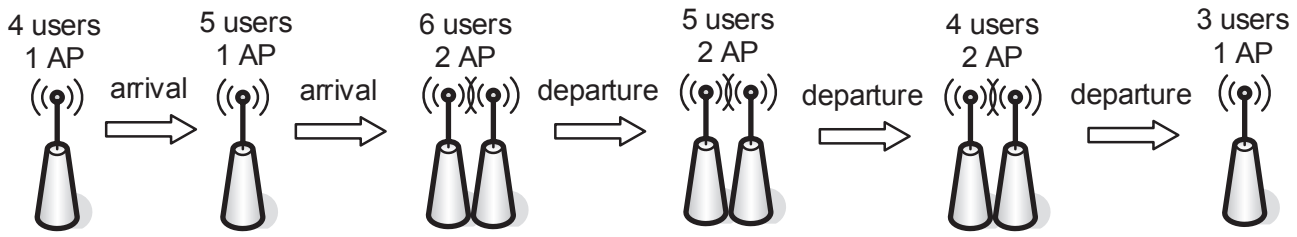

Fig. 1. Example of the powering on/off process for $N_h = 5$ and $N_l = 3$.

commodity hardware can support an intermediate state (i.e., switching on/of the wireless card), this does not bring as much savings as powering on/off the complete device [6].

In our model, a "user" is a new connection generated by a wireless client. Following [7], these are generated according to a Poisson process at rate λ and are always served by the less loaded AP. The AP bandwidth is evenly shared among all the users, which demand an exponentially distributed amount of work. We argue that although in real systems these amounts of work may deviate from the exponential distribution, this assumption serves to illustrate the impact of boot-up times on performance. Based on these assumptions, service times are also exponentially distributed, with the departure rate being μ when there is only one serving AP and 2μ when both APs are serving, i.e., we neglect the impact of channel sharing. We assume a load-balancing algorithm such that users (re)associate while they are being served, and that this (re)association time is negligible –note that this can be achieved with the recent 802.11v and 802.11r amendments [8], which supports triggering re-associations and performing fast transitions, respectively, with minor disruption of the service. Following our previous measurements, we will also neglect the time required to power off an AP.

We set the maximum number of users per AP to K. This assumption on the "hard capacity" on the number of users, also used in [3], emulates the provisioning of a minimum bandwidth (e.g., QoS) or the finite size of the address pool. Based on this, the maximum number of users allowed into the network is $2K$; however, despite there should be at most K users per AP when this maximum is reached, we allow up to $2K$ users into the first AP while the second one is being powered on, as users will re-associate once it becomes available.

In order to power on and off the second AP, we assume that there is a threshold-based policy with hysteresis: the second AP is powered on when there are N_h users associated with the first AP and another user arrives, and it is powered off when there are $N_l + 1$ users in the system and one of them leaves. Therefore, the power on-off process has a hysteresis of size $N_h - N_l$. We illustrate in Fig. 1 an example of the process of switching on/off APs for the case of $N_h = 5$ and $N_l = 3$. As the figure shows, when there are 5 users in the WLAN only one AP is powered on, but when a sixth user arrives the second AP starts to boot up (although it may take some time before it can serve users). Then, at some point a user leaves, but both APs are kept on, and even with four users no AP is deactivated. Only when the limit $N_l = 3$ is reached, the second AP is switched off and only one AP remains active. This example corresponds to a hysteresis of $N_h - N_l = 2$.

We characterize the performance of the system with three figures:

- The average power consumed by the infrastructure P
- The average time spent in the system by a user T_s.
- The probability that a user is not allowed into the system because of reaching the *hard limit* of $2K$ users, i.e., the blocking probability p_B.

The focus of the work is to model the impact of T_{on} on these variables.

III. Performance Analysis

We model our system with the *regenerative process* [9] illustrated in Fig. 2. This regenerative process is formed by three stages, which depend on the status of the second AP:

- Stage A, in which the second AP is inactive.
- Stage B, in which it is being powered on but cannot serve clients yet.
- Stage C, in which both APs are active and serving users.

Following the description of the system model, there are three transitions:

- The transition $A \rightarrow B$, which is produced when there are N_h users associated with the first AP and a new user arrives.
- The transition $B \rightarrow C$, which is triggered by the completion of the T_{on} units of time required to power on the second AP.
- The transition $C \rightarrow A$, which occurs when there are $N_l + 1$ users in the system and one of them leaves.

We note that, in case there are N_l or fewer users when the transition $B \rightarrow C$ happens (i.e., a number of users left while the second AP was switching on), we will consider that the system traverses state C with a zero sojourn time, and then transition to state A.

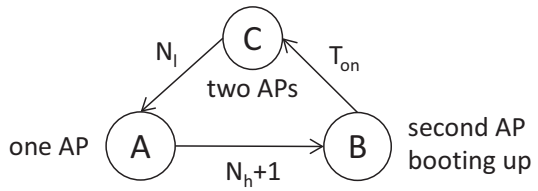

Fig. 2. Regenerative process to model the system.

In the following, we first describe how to compute the performance figures of the complete system, based on per-stage variables, and then present a model for the dynamics of the system, based on a Markov chain model for each stage. Throughout the article, we will refer with "stage" to the three states of the regenerative process illustrated in Fig. 2, and reserve the use of "state" for the description of the Markov chains. We note that this analysis of a two-AP scenario is exact as long as the assumptions on the arrival and departure processes, and the (re)association times hold.

A. Computing the overall performance figures

The average duration T of a complete cycle of the regenerative process can be computed as

$$T = T^A + T^B + T^C, \tag{1}$$

where T^j is the average sojourn time of stage j.

Note that, in our scenario, we have by definition that $T^B = T_{on}$, while the computation of T^A and T^B will be performed in the following subsection.

Based on the T^j, the average *power consumed* by the network is

$$P = \frac{P_{AP}T^A + 2P_{AP}(T^B + T^C)}{T}. \tag{2}$$

To compute the other performance figures, we need to obtain the expected amount of time that there are i users in the system during the duration of a cycle, T_i. Similarly to (1), this value can be expressed as

$$T_i = T_i^A + T_i^B + T_i^C, \tag{3}$$

where T_i^j is the average amount of time that there are i users in the system during the sojourn time of stage j. With the values of T_i and T, the probability p_i that there are i users in the system is given by

$$p_i = \frac{T_i}{T} = \frac{T_i^A + T_i^B + T_i^C}{T^A + T^B + T^C} \tag{4}$$

Based on the p_i, the *blocking probability* is equal to the probability that there are $2K$ users in the system, i.e.,

$$p_B = p_{2K}, \tag{5}$$

while the *average time* spent by a user in the system T_s is given by Little's formula:

$$T_s = \frac{N_t}{\lambda\,(1 - p_B)}, \tag{6}$$

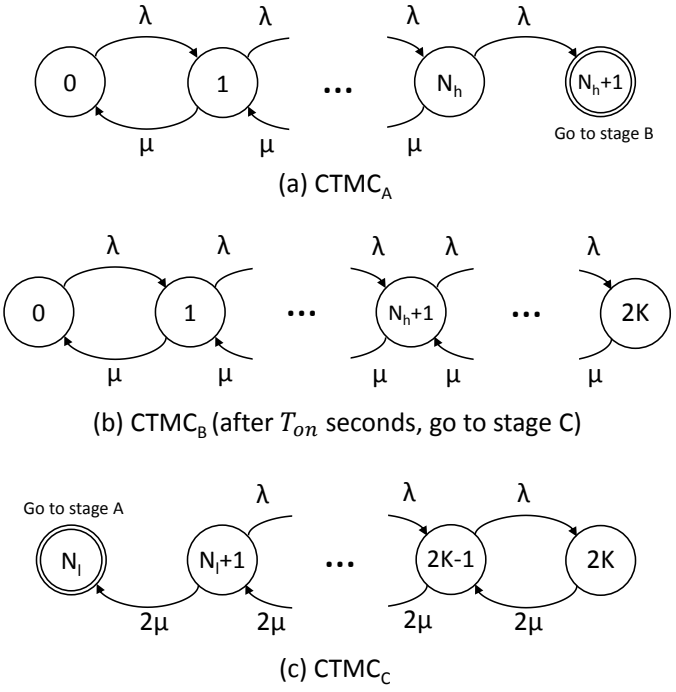

Fig. 3. CTMCs representing the different stages of the regenerative process.

where N_t corresponds to the average number of users in the system, which is computed as

$$N_t = \sum_{i=0}^{2K} i p_i. \tag{7}$$

With the above, we can compute the performance figures of the system with (2), (5) and (6), *given the times* T_i^j and T^j. We next describe how to compute these by modeling the dynamics of each stage of the regenerative process.

B. Modeling each stage of the regenerative process

The three stages of the regenerative process can be modeled with three different Continuous-Time Markov Chains (CTMCs), illustrated in Fig. 3. In all the chains, the state models the number of users being served by the system, each chain having a different number of states:

- CTMC_A models the system when only one AP is powered on, and therefore its number of states ranges from 0 (empty system) to $N_h + 1$ (the system transitions to the next stage).
- CTMC_B models the system during the T_{on} units of time it takes for the second AP to power, and therefore it can serve between 0 and the maximum number of users $2K$.
- CTMC_C models the system when the two APs are serving users, and therefore ranges between $2K$ and N_l (the system transitions to stage A).

We next analyze each of these CTMCs separately, starting with CTMC_B (the one with the largest number of states).

1) CTMC$_B$: This case is illustrated in Fig. 3b, where users arrive at a rate λ and are served at a rate μ. Our aim is to compute the expected total time the CTMC spends in each state during the interval $[0, T_{on})$. If we define $\pi_i(t)$ as the probability that a CTMC is in state i at time t, the expected total time spent in that state i during the interval $[0, t)$ is

$$L_i(t) = \int_0^t \pi_i(u)\,\mathrm{d}u, \tag{8}$$

and based on this, we can compute $T_i^B = L_i^B(T_{on})$, which is required to derive the performance figures of the system as explained in the previous section.

To compute $\pi_i(t)$, we must solve the differential equation

$$\frac{\mathrm{d}\boldsymbol{\pi}(t)}{\mathrm{d}t} = \boldsymbol{\pi}(t)\,\mathbf{Q} \tag{9}$$

where $\boldsymbol{\pi}(t)$ and $\mathbf{Q}$ are the vector of state probabilities and the generator matrix of the CTMC, respectively.

For CTMC$_\mathrm{B}$ we have that

$$\boldsymbol{\pi}^B(t) = \left[\pi_0^B(t), \ldots, \pi_{2K}^B(t)\right]$$

and

$$\mathbf{Q}^B = [q_{ij}],\ i,j \in \{0,\ldots,2K\},$$

with the elements of this matrix being

$$q_{ij} = \begin{cases} -\lambda & \text{for } i = 0, j = 0 \\ -\mu & \text{for } i = 2K, j = 2K \\ -\lambda-\mu & \text{for } i = \{1,\ldots,2K-1\}, j = i \\ \lambda & \text{for } i = \{0,\ldots,2K-1\}, j = i+1 \\ \mu & \text{for } i = \{1,\ldots,2K\}, j = i-1 \\ 0 & \text{in any other case} \end{cases} \tag{10}$$

We also need the set of initial conditions $\boldsymbol{\pi}^B(0)$ to solve (9). Given that stage B starts when there are N_h users in the system and a new arrival happens, we have that

$$\pi_i^B(0) = \begin{cases} 1 & \text{for } i = N_h + 1 \\ 0 & \text{in any other case} \end{cases}$$

With these, we can solve the system specified by (9) and compute T_i^B with (8) as explained above.[1] Note that we can also obtain $\boldsymbol{\pi}^B(T_{on})$, which is required to compute the set of initial conditions for both the next stage C and stage A, as explained next.

2) CTMC$_C$: This case is illustrated in Fig. 3c, with the departure rate being 2μ as both APs are serving users. In contrast to the previous chain, CTMC$_\mathrm{C}$ has an absorbing state, namely, N_l. When the system reaches this number of users, the second AP is powered off and the system transitions to stage A.

As in the previous case, we need to compute the expected total time the chain spends in each state during the sojourn time T^C. These values correspond to the *time until absorption* spent in each of the non-absorbing states of CTMC$_\mathrm{C}$, which are defined as $\lim_{t\to\infty} L_i(t)$ for the set of states $\{T_{l+1},\ldots,T_{2K}\}$. The times before absorption can be computed as [10]

$$\mathbf{L}^C(\infty)\,\mathbf{Q}^C = -\boldsymbol{\pi}^C(0), \tag{11}$$

where

$$\mathbf{L}^C(t) = \left[L_{N_l+1}^C(t), \ldots, L_{2K}^C(t)\right],$$
$$\boldsymbol{\pi}^C(t) = \left[\pi_{N_l+1}^C(t), \ldots, \pi_{2K}^C(t)\right],$$

and

$$\mathbf{Q}^C = [q_{ij}],\ i,j \in \{N_l+1,\ldots,2K\},$$

with

$$q_{ij} = \begin{cases} -2\mu & \text{for } i = 2K, j = 2K \\ -\lambda-2\mu & \text{for } i = \{N_l+1,\ldots,2K-1\}, j = i \\ \lambda & \text{for } i = \{N_l+1,\ldots,2K-1\}, j = i+1 \\ 2\mu & \text{for } i = \{N_l+2,\ldots,2K\}, j = i-1 \\ 0 & \text{in any other case} \end{cases} \tag{12}$$

The initial conditions $\boldsymbol{\pi}^C(0)$ are determined by the distribution of the state probabilities at the end of stage B, i.e., $\pi_i^B(T_{on})$: if there are less than N_l+1 users in the system, the second AP is immediately powered off and the system transitions to stage A; otherwise, the number of users at the end of stage B corresponds to the number of users at the beginning of stage C.

Following the above, we have that

$$\pi_i^C(0) = \begin{cases} \pi_i^B(T_{on}) & \text{for } i = \{N_l+1,\ldots,2K\} \\ 0 & \text{in any other case} \end{cases}$$

Therefore, the system will spend zero sojourn time at stage C with probability $1 - \sum_{N_l+1}^{2K} \pi_i^C(0)$.

Once (11) is solved, the sojourn time of state C can be computed as

$$T^C = \sum_{i=N_l+1}^{2K} L_i^C(\infty), \tag{13}$$

and $T_i^C = L_i^C(\infty)$ for $i = \{N_l+1,\ldots,2K\}$ and 0 elsewhere.

3) CTMC$_A$: This case, illustrated in Fig. 3a, is also modeled with a CTMC with an absorbing state, namely, N_h+1. This state triggers the activation of the second AP, which corresponds to the transition to stage B.

The times before absorption can be computed also with (11), where now we have

$$\mathbf{L}^A(t) = \left[L_0^A(t), \ldots, L_{N_h}^A(t)\right],$$
$$\boldsymbol{\pi}^A(t) = \left[\pi_0^A(t), \ldots, \pi_{N_h}^A(t)\right],$$

and

$$\mathbf{Q}^A = [q_{ij}],\ i,j \in \{0,\ldots,N_h\},$$

with

$$q_{ij} = \begin{cases} -\lambda & \text{for } i = 0, j = 0 \\ -\lambda-\mu & \text{for } i = \{1,\ldots,N_h\}, j = i \\ \lambda & \text{for } i = \{0,\ldots,N_h-1\}, j = i+1 \\ \mu & \text{for } i = \{1,\ldots,N_h\}, j = i-1 \\ 0 & \text{in any other case} \end{cases} \tag{14}$$

[1] Instead of solving (9) and then computing (8), $\pi_i(t)$ and $L_i(t)$ can be efficiently evaluated for a given $t = T_{on}$ value using the *uniformization* method.

Similarly to the case of CTMC_C, the set of initial conditions $\boldsymbol{\pi}^A(0)$ is determined by the status of the system at the end of stage B: in case there were less than N_l users once the second AP is available, the system will transition directly to stage A, i.e.,

$$\pi_i^A(0) = \pi_i^B(T_{on}), \quad \text{for } i = \{0, \ldots, N_l - 1\}, \tag{15}$$

otherwise, the transition to stage A will happen through state N_l, i.e.,

$$\pi_i^A(0) = 1 - \sum_{j=0}^{N_h - 1} \pi_j^B(T_{on}), \quad \text{for } i = N_l \tag{16}$$

and correspondingly $\pi_i^A(0) = 0$ for any other state.

Finally, the sojourn time of state A is computed as

$$T^A = \sum_{i=0}^{N_h} L_i^A(\infty), \tag{17}$$

and $T_i^A = L_i^A(\infty)$ for $i = \{0, \ldots, N_h\}$ and 0 elsewhere.

IV. Performance Evaluation

To analyze the impact of T_{on} on the performance, we assume a system in which up to $2K = 10$ users are allowed, $P_{AP} = 3.5$ W, and $\lambda = 0.1$ arrivals/s and $1/\mu = 10$ s, which corresponds to an average load of approx. 50%.[2] We consider four different activation policies:

- $N_l = N_h = 4$: no hysteresis and the activation threshold lower than the maximum number of user per AP (K).
- $N_l = N_h = 5$: no hysteresis and the activation threshold set to K.
- $N_l = 2$ and $N_h = 4$: a hysteresis of two users and an activation threshold set to $K - 1$.
- $N_l = 2$ and $N_h = 5$: a hysteresis of three users and an activation threshold equal to K.

For each considered policy, we analyze the impact of T_{on} on the service time T_s, the power consumed P, and the blocking probability p_B, with the results shown in Fig. 4, Fig. 5, and Fig. 6, respectively. In the figures, we depict with lines the values from our analytical model and with points the results of a discrete event simulator, where each point represents the average of ten simulation runs, each run consisting on more than 10^6 user departures (we do not represent the 95%-confidence intervals as their relative size is well below 1%).

There are several observations that can be drawn from the figures. First, the results from the model coincide with the simulations values for all considered configurations (we obtained the same accuracy for other configurations of the load, omitted for space reasons), which confirms the validity of our analysis. Second, the results also confirm that T_{on} has a non-negligible impact on performance: it increases delay figures by 25–35%, power consumption by up to 20%, and approximately quadruples the blocking probability for the case of $N_l = N_h = 5$.

[2]These service times can emulate a scenario where a user downloads e.g. 20 MB using 802.11g, assuming an effective throughput of approximately 15 Mbps.

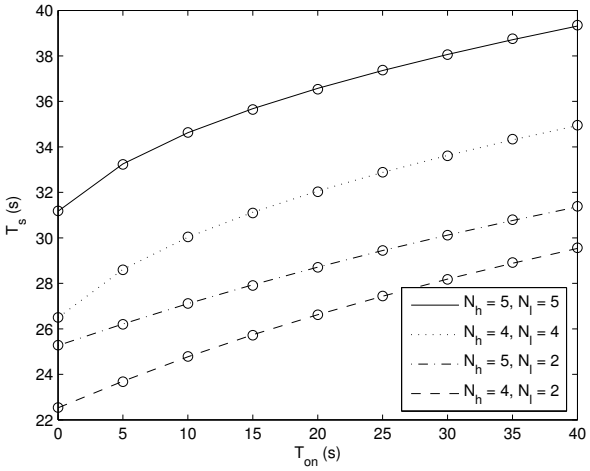

Fig. 4. Impact of the activation time T_{on} on the serving time T_s.

In addition to the above, which confirms the quantitative impact of T_{on} on performance, we note that non-zero start-up times introduce *qualitatively* different results. For instance, for the case of power consumption (Fig. 5), when $T_{on} = 0$, the less consuming scheme is $N_h = 5, N_l = 5$; however, when $T_{on} > 5$ s, the less consuming policy becomes $N_h = 5, N_l = 2$. In this way, a strategy designed for $T_{on} = 0$ can be outperformed by other policies when $T_{on} > 0$ (indeed, for $T_{on} \geq 20$ s it is outperformed by two strategies). Similarly, there is a trade-off between delay performance and power consumption for $T_{on} \approx 0$, i.e., less consuming strategies lead to the largest delays; however, when $T_{on} > 5$ s, this trade-off disappears.

Additionally, we performed further experiments that confirm the influence of T_{on} on performance for different values of N_h and N_l. In general, the smaller the hysteresis, the higher the impact of T_{on}. Therefore, we conclude that T_{on} has a notable impact on performance, both qualitatively and qualitatively, and has to be taken into account when designing a RoD policy.

V. Conclusions and Future work

In this work, we have presented an analytical model for the case of a simple RoD system, which takes into account the time required to power on an AP. The accuracy of the model has been validated via simulations, and results have showed that, even for the simple scenario considered, the time required to start-up an AP has a dramatic impact on performance. Indeed, this time alters both the quantitative and qualitative results as compared to the case of zero start-up time. We believe that results should be taken into account when designing infrastructure on demand policies in real-life deployments.

We are currently working to extend our model along two lines:

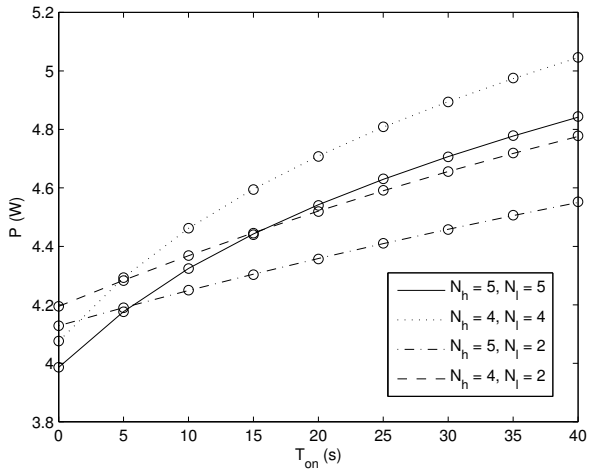

Fig. 5. Impact of the activation time T_{on} on the power consumed P.

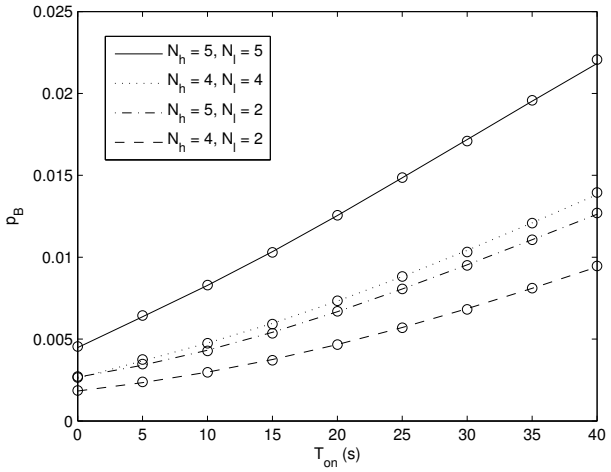

Fig. 6. Impact of the activation time T_{on} on the blocking probability p_B.

- Adding a third AP to the model, which could then be used to design configuration policies for existing WLANs in the 2.4 GHz band, when there are at most 3 non-overlapping channels.
- Adding another energy state to the Access Points, i.e., a sleep-like state in which the energy consumption is not zero but the time to power on is significantly reduced, to understand in which scenarios it is more efficient not to completely switch off the infrastructure.

Finally, we are also planning the deployment of a small-size testbed to experiment with resource-on-demand algorithms, with the aim to provide seamless operation based on current standards while minimizing energy consumption.

ACKNOWLEDGEMENTS

This work has been partly supported by the European Community through the CROWD project (FP7-ICT-318115) and by the Madrid Regional Government through the TIGRE5-CM program (S2013/ICE-2919).

REFERENCES

[1] P. Serrano, A. de la Oliva, P. Patras, V. Mancuso, and A. Banchs, "Greening wireless communications: Status and future directions," *Computer Communications*, vol. 35, no. 14, pp. 1651 – 1661, 2012.

[2] A. Jardosh, K. Papagiannaki, E. Belding, K. Almeroth, G. Iannaccone, and B. Vinnakota, "Green WLANs: On-demand WLAN infrastructures," *Mobile Networks and Applications*, vol. 14, no. 6, pp. 798–814, 2009.

[3] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "A simple analytical model for the energy-efficient activation of access points in dense wlans," in *Proceedings of e-Energy '10*. New York, NY, USA: ACM, 2010, pp. 159–168.

[4] A. P. C. da Silva, M. Meo, and M. A. Marsan, "Energy-performance trade-off in dense wlans: A queuing study," *Computer Networks*, vol. 56, no. 10, pp. 2522 – 2537, 2012.

[5] M. A. Marsan and M. Meo, "Queueing systems to study the energy consumption of a campus WLAN," *Computer Networks*, vol. 66, pp. 82–93, 2014.

[6] P. Serrano, A. Garcia-Saavedra, G. Bianchi, A. Banchs, and A. Azcorra, "Per-frame energy consumption in 802.11 devices and its implication on modeling and design," *Networking, IEEE/ACM Transactions on*, vol. PP, no. 99, pp. 1–1, 2014.

[7] M. Papadopouli, H. Shen, and M. Spanakis, "Modeling client arrivals at access points in wireless campus-wide networks," in *Local and Metropolitan Area Networks, 2005. LANMAN 2005. The 14th IEEE Workshop on*, 2005, pp. 6 pp.–6.

[8] G. R. Hiertz, D. Denteneer, L. Stibor, Y. Zang, X. P. Costa, and B. Walke, "The IEEE 802.11 Universe," *Comm. Mag.*, vol. 48, no. 1, pp. 62–70, Jan. 2010.

[9] J. Medhi, *Stochastic Models in Queueing Theory*. Academic Press, 2002.

[10] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains*. Wiley-Interscience, 2006.