# Optimal configuration of a Resource-on-Demand 802.11 WLAN with Non-Zero Start-Up Times[☆]

Jorge Ortín[a], Pablo Serrano[b,*], Carlos Donato[b,c]

[a] *Centro Universitario de la Defensa Zaragoza, Spain*
[b] *Departamento de Ingeniería Telemática, Universidad Carlos III de Madrid, Leganés, Spain*
[c] *IMDEA Networks Institute, Leganés, Spain*

**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. Our findings show that having a non-zero start up time modifies significantly the trade-offs between power consumption and performance that appears on Resource on Demand solutions. Finally, we propose an algorithm to optimize the energy consumption of the network while guaranteeing a given performance bound.

*Keywords:* WLAN, 802.11, Resource on Demand, Energy Consumption, Infrastructure on Demand

## 1. 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 [2, 3]. For the case of Wireless LAN (WLAN), though, fewer works have addressed the problem of RoD [4, 5, 6].

Very recently, two surveys have addressed the impact of sleep-mode techniques [7] and the impact of on-demand activation of resources [8] on the energy efficiency of wireless networks. Both surveys agree that the partial or full deactivation of base stations/APs with low traffic is a key enabler for energy efficiency. In this regard, two of the main conclusions in [7], whose focus is cellular networks, are: (i) dynamic approaches outperform static solutions; and (ii) there is a

[☆] This paper is an extended version of our paper [1], which was presented at the Fourth IFIP Conference on Sustainable Internet and ICT for Sustainability.

[*] Corresponding author. Address: Avenida de la Universidad 30, Edif. Torres Quevedo, E-28911 Leganés, Madrid, Spain

widespread use of over-simplified models, and further research analyzing the impact of parameters such as the time required to switch on/off base stations is needed.

The review of on-demand approaches [8] considers both WLANs and cellular networks. In this survey, more than fifty strategies are classified according to the type of network (cellular, WLAN), performance metric (user demand, coverage, QoS, energy efficiency), type of algorithm (online fast reaction, online slow reaction, offline) and control scheme (centralized, distributed, pseudo-distributed, cooperative). In this work, it is also highlighted that the delay to switch on/off equipment should be considered when implementing the algorithms in real environments.

Among the works cited in [8], in the seminal work of [4], 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 [5] 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 APs on parameters such as the energy savings and the switch-off rate of the devices. In [6], authors extend the work of [5] 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.

In both analytical works [5, 6], as well as in a recent follow-up analysis [9], among other simplifying assumptions, authors neglect the time required to power on an AP. This assumption is also made in [10], where the impact of the AP power model on the energy efficiency of a WLAN is analyzed. However, in [4] 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, following the methodology we described in [11] and 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). We note that, for this experiment, there is no traffic being transmitted or received in the ON state, which significantly impacts the energy consumed as reported in [11]. The results are provided in Table 1. 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.

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

Table 1: Time required to switch from the ON state to the OFF state (and vice-versa) in a Linksys WRT54GL.

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.

## 2. System Model

Our system is a simplified version of the *cluster model* analyzed in [5], 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 [5], 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 ac-

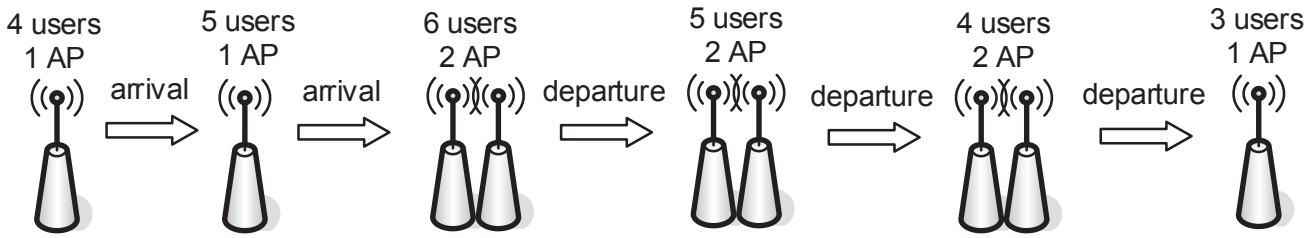


Figure 1: Example of the powering on/off process for $N_h = 5$ and $N_l = 3$.

tive (i.e., during start-up and when powered on) and 0 otherwise. Although commodity hardware can support an intermediate state (i.e., switching on/off the wireless card), this does not bring as much savings as powering on/off the complete device [11].

In our model, a "user" is a new connection generated by a wireless client. Following [12], these are generated according to a Poisson process at rate $\lambda$ 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 $\mu$ when there is only one serving AP and $2\mu$ 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 [13], which support 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 [5], 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 the following 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 rate at which the second AP is powered on/off $\omega$, which is another key variable of interest as it can affect the lifetime of the equipment.

The focus of the work is first to model the impact of $T_{on}$ on these variables, then to understand the different trade-offs in performance, and finally to derive the optimal configuration of an RoD scheme based on an optimization criterion.

## 3. Performance Analysis

We model our system with the *regenerative process* [14] 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 higher than the hysteresis left while the second AP was switching on), we will consider that the system traverses state $C$ with a zero sojourn time, and then transitions to state $A$.

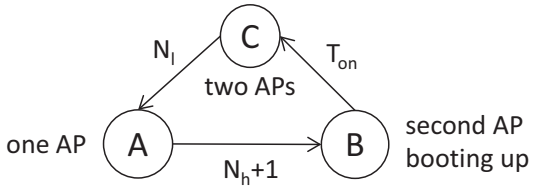


Figure 2: Regenerative process to model the system.

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.

### *3.1. 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 $T^C$ 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

$$p_i = \frac{T_i}{T} = \frac{T_i^A + T_i^B + T_i^C}{T^A + T^B + T^C}. \quad (4)$$

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

$$p_B = p_{2K}, \quad (5)$$

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

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

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

$$N_t = \sum_{i=0}^{2K} i p_i. \quad (7)$$

Finally, the derivation of the deactivation rate of the second AP $\omega$ is almost immediate, given that it corresponds to the inverse of a complete power on–power off cycle, i.e., the average duration of a cycle of the regenerative process. Therefore, it can be computed as

$$\omega = \frac{1}{T^A + T^B + T^C}. \quad (8)$$

With the above, we can compute the performance figures of the system with (2), (5), (6), and (8), given the times $T_i^j$ and $T^j$. We next describe how to compute these times by modeling the dynamics of each stage of the regenerative process.

### *3.2. 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:

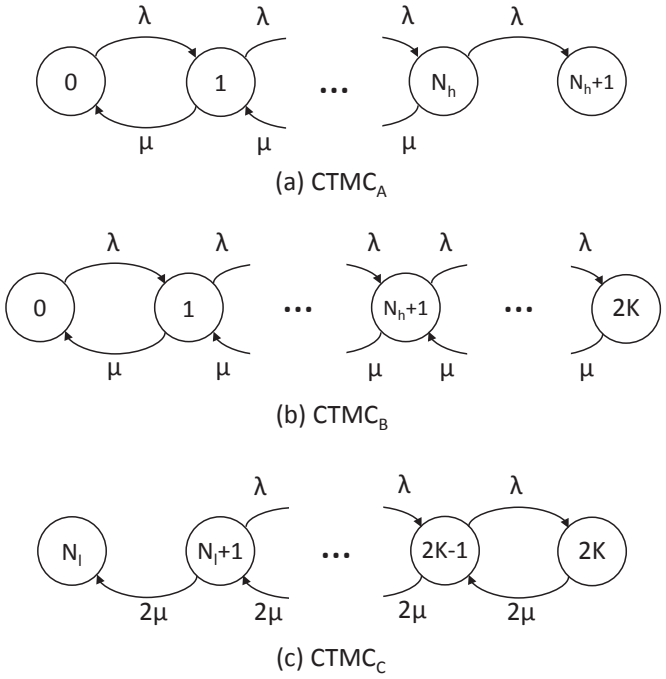


Figure 3: CTMCs representing the different stages of the regenerative process.

- $\text{CTMC}_\text{A}$ 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).
- $\text{CTMC}_\text{B}$ 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 $2K$.
- $\text{CTMC}_\text{C}$ models the system when the two APs are serving users, and therefore ranges between $N_l$ and $2K$ (the system transitions to stage A).

We next analyze each of these CTMCs separately, starting with $\text{CTMC}_\text{B}$ (the one with the largest number of states).

#### *3.2.1. CTMC$_B$*

This case is illustrated in Fig. 3b, where users arrive at a rate $\lambda$ and are served at a rate $\mu$. 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

$$L_i(t) = \int_0^t \pi_i(u)\,\mathrm{d}u, \tag{9}$$

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

$$\frac{\mathrm{d}\boldsymbol{\pi}(t)}{\mathrm{d}t} = \boldsymbol{\pi}(t)\,\mathbf{Q}, \tag{10}$$

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

For $\mathrm{CTMC_B}$ 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

$$q_{ij} = \begin{cases} -\lambda & \text{for } i = 0 \text{ and } j = 0 \\ -\mu & \text{for } i = 2K, j = 2K \\ -\lambda - \mu & \text{for } i = \{1, \ldots, 2K-1\} \\ & \text{and } j = i \\ \lambda & \text{for } i = \{0, \ldots, 2K-1\} \\ & \text{and } j = i+1 \\ \mu & \text{for } i = \{1, \ldots, 2K\} \\ & \text{and } j = i-1 \\ 0 & \text{in any other case} \end{cases} \tag{11}$$

We also need the set of initial conditions $\boldsymbol{\pi}^B(0)$ to solve (10). 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 (10) and compute $T_i^B$ with (9) 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.

### *3.2.2. CTMC$_C$*

This case is illustrated in Fig. 3c, with the departure rate being $2\mu$ as both APs are serving users. In contrast to the previous chain, $\mathrm{CTMC_C}$ 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 $\mathrm{CTMC_C}$, which are defined as $\lim_{t\to\infty} L_i(t)$ for the set of states $\{T_{N_l+1}, \ldots, T_{2K}\}$. The times before absorption can be computed as [15]

$$\mathbf{L}^C(\infty)\,\mathbf{Q}^C = -\boldsymbol{\pi}^C(0), \tag{12}$$

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

$$q_{ij} = \begin{cases} -2\mu & \text{for } i = 2K, j = 2K \\ -\lambda - 2\mu & \text{for } i = \{N_l+1, \ldots, 2K-1\} \\ & \text{and } j = i \\ \lambda & \text{for } i = \{N_l+1, \ldots, 2K-1\} \\ & \text{and } j = i+1 \\ 2\mu & \text{for } i = \{N_l+2, \ldots, 2K\} \\ & \text{and } j = i-1 \\ 0 & \text{in any other case} \end{cases} \tag{13}$$

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

[1] Instead of solving (10) and then computing (9), $\pi_i(t)$ and $L_i(t)$ can be efficiently evaluated for a given $t = T_{on}$ value using the *uniformization* method.

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 (12) is solved, the sojourn time of stage $C$ can be computed as

$$T^C = \sum_{i=N_l+1}^{2K} L_i^C(\infty), \tag{14}$$

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

*3.2.3. 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 (12), 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

$$q_{ij} = \begin{cases} -\lambda & \text{for } i = 0 \text{ and } j = 0 \\ -\lambda-\mu & \text{for } i = \{1, \ldots, N_h\} \text{ and } j = i \\ \lambda & \text{for } i = \{0, \ldots, N_h-1\} \text{ and } j = i+1 \\ \mu & \text{for } i = \{1, \ldots, N_h\} \text{ and } j = i-1 \\ 0 & \text{in any other case} \end{cases} \tag{15}$$

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.,

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

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

$$\pi_i^A(0) = 1 - \sum_{j=0}^{N_l-1} \pi_j^B(T_{on}), \quad \text{for } i = N_l \tag{17}$$

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

Finally, the sojourn time of stage $A$ is computed as

$$T^A = \sum_{i=0}^{N_h} L_i^A(\infty), \tag{18}$$

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

## 4. Impact of $T_{on}$ on performance

To analyze the impact of $T_{on}$ on the performance, we assume a system in which up to $2K = 10$ users are allowed, a fixed value of $P_{AP} = 3.5$ W,[2] and $\lambda = 0.1$ arrivals/s and $1/\mu = 10$ s, which corresponds to an average load of approx. 50%.[3] 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 the activation threshold set to $K-1$.

[2] For simplicity, we assume a constant power consumption figure throughout all scenarios. Following our previous work of [11], the consumption of a Linksys AP ranges between approx. 2.7 W when there is no activity and 4.4 W when the activity is maximum. The average of these figures results approx. 3.5 W, which is the value used.

[3] 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.

- $N_l = 2$ and $N_h = 5$: a hysteresis of three users and the activation threshold equal to $K$.

When presenting the results, we depict with lines the values from our analytical model and with points the results of a discrete event simulator that is written in C and whose operation we validated thoroughly. Each point represents the average of ten simulation runs, and 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%).

### *4.1. Total delay $T_s$*

We first analyze, for the four considered policies listed above, the impact of $T_{on}$ on the total delay $T_s$, with the results shown in Fig. 4. There are several observations that can be drawn from the figure. 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, as it increases delay figures by 25–35% as compared to the case of zero start-up times. Finally, the policy more reluctant to power on the second AP (i.e., $N_h = 5, N_l = 5$) results in the largest delays for all values of $T_{on}$, while the policy more eager to power on the second AP (i.e., $N_h = 4, N_l = 2$) results in the smallest delay values.

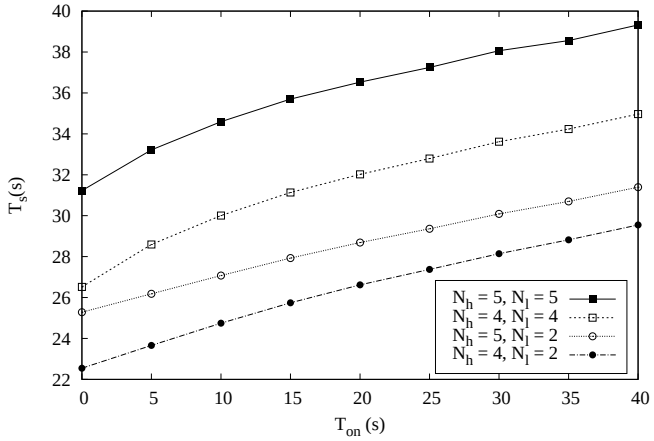


Figure 4: Impact of the activation time $T_{on}$ on the total delay $T_s$.

### *4.2. Power consumed P*

We next analyze the impact of $T_{on}$ on the total power consumed by the network with the four considered policies, with the results shown in Fig. 5. First, as in the previous case, it is clear that $T_{on}$ has significant impact on the performance w.r.t. this variable as well, as it increases power consumption by up to 20%. 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, when $T_{on} = 0$, the less consuming scheme is $N_h = 5, N_l = 5$ (which is inline with intuition, given that the system spends most of the time in stage $A$ and therefore the term $P_{AP}T^A$ prevails in (2)); however, when $T_{on} > 5$ s, the less consuming policy becomes $N_h = 5, N_l = 2$. We also note that the policy that resulted in the smallest delays ($N_h = 4, N_l = 2$) has the largest power consumption only for $T_{on} < 5$ s. More specifically, 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 partially under some configurations (we will further explore these trade-offs in the next section). In this way, a strategy designed to minimize the power consumption 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).

### *4.3. Blocking probability $p_B$*

Concerning the results on the probability $p_B$ that a user is not allowed into the system, because the maximum capacity $2K$ has been reached, they are presented in Fig. 6. The observed behavior in this case is expected, given the results on $T_s$ presented in Fig. 4 and the relationship between $T_s$ and $p_B$ given in (6), with the relative order of the different activation policies being the same: the higher the delays (because the second AP is powered off for relatively longer periods of time),

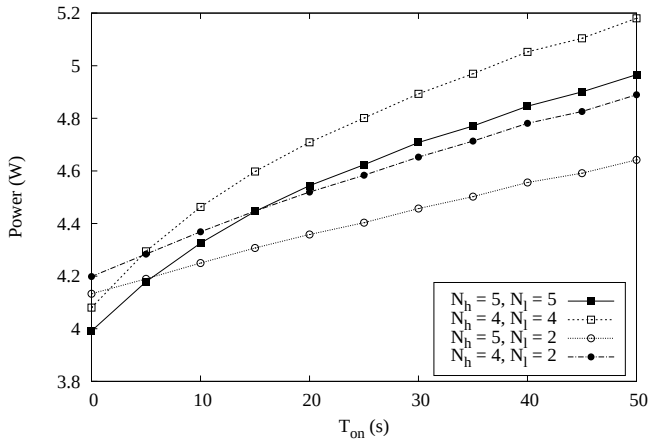


Figure 5: Impact of the activation time $T_{on}$ on the power consumed $P$.

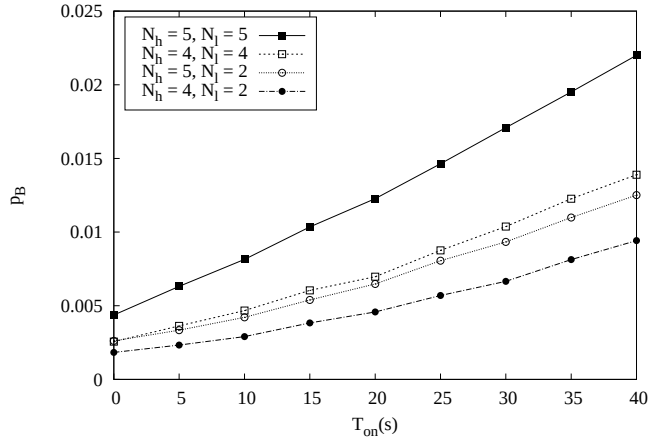


Figure 6: Impact of the activation time $T_{on}$ on the blocking probability $p_B$.

the higher the probability that a user cannot be admitted into the system.

### *4.4. Activation rate $\omega$*

Finally, we analyze the impact of $T_{on}$ on the rate at which the second AP is powered on and off $\omega$, with the results being illustrated in Fig. 7. As expected, the results show that, in general, the longer it takes the AP to boot (and consequently, the longer the system remains in stage $B$), the lower the activation rate will be, as expressed in (8). The figure also shows that those policies with more hysteresis (i.e., $N_l = 2$) obtain lower activation rates and result less sensitive to $T_{on}$, the reason being that the hysteresis increases the average sojourn times of stages $A$ and $C$, thus decreasing

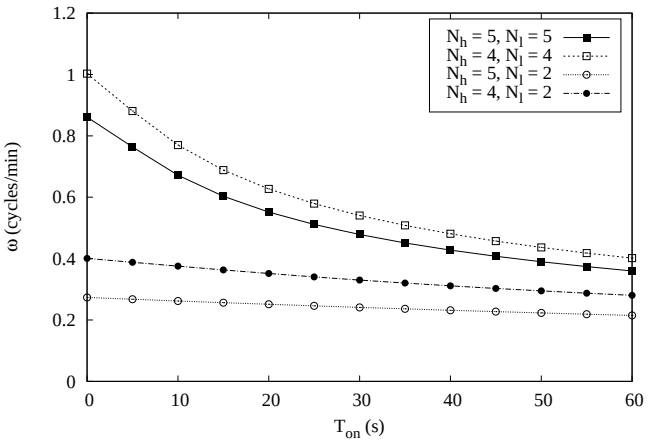


Figure 7: Impact of the activation time $T_{on}$ on the activation rate $\omega$.

the influence of the term $T^B$ in (8). Finally, given a specific hysteresis, the policies more reluctant to power on the second AP also obtain lower activations rates, since the condition to switch on the second AP (i.e., reaching $N_h$ users) is harder to satisfy.

## 5. On the trade-offs in a RoD scheme

Building on the previous results, in this section we analyze the different trade-offs that appear in a network that implements a resource on demand scheme. More specifically, in the previous section we have seen that, depending on the $N_h$ and $N_l$ configuration, and the value of $T_{on}$, the performance in terms of $T_s$, $P$, $p_B$ and $\omega$ changes both qualitatively and qualitatively. Now we want to further explore these changes, considering also different values of the system load $\rho$.

Throughout this section we will focus our findings on the most illustrative trade-offs, namely:

1. Total delay ($T_s$) vs. power consumption ($P$). This trade-off serves to represent the cost in terms of power consumption for a given gain in terms of performance, e.g., how many watts costs a given reduction in seconds.
2. Power consumption ($P$) vs. activation rate ($\omega$). This trade-off illustrates that the resource consumption has two dimensions that must be carefully considered when configuring the RoD scheme, since a decrease of the

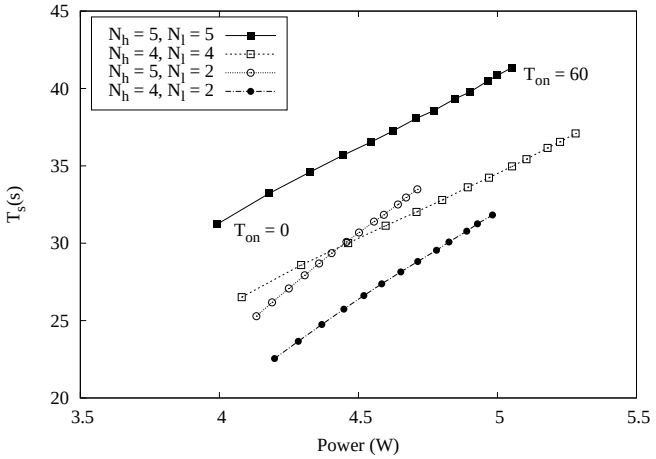


Figure 8: Impact of the activation time $T_{on}$ on the $T_s$ vs. $P$ trade-off.

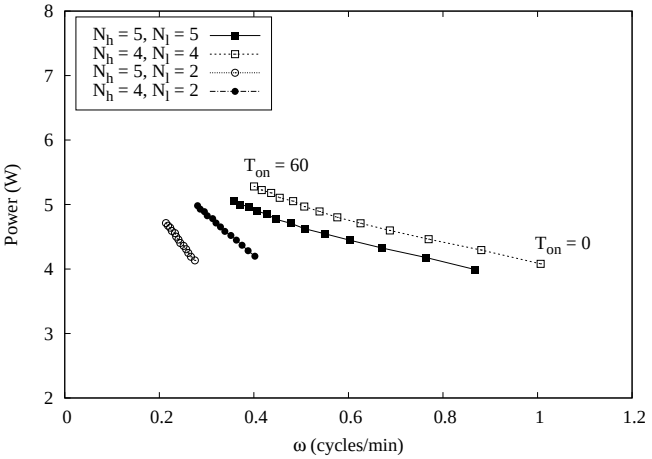


Figure 9: Impact of the activation time $T_{on}$ on the $P$ vs. $\omega$ trade-off.

power consumption may lead to an undesirable increase of the activation rate of the second AP.[4]

### *5.1. Impact of $T_{on}$*

We start our analysis with the impact of $T_{on}$ on the two considered trade-offs. To this aim, we build on the results from the previous section (i.e., $\rho = 1/2$), and depict them in Figs. 8 and 9, where each point corresponds to a pair of values ($\{P, T_s\}$ for Fig. 8, and $\{\omega, P\}$ for Fig. 9) for a different value of $T_{on}$.[5]

On the one hand, Fig. 8 illustrates that, as already showed in the previous section, the longer the value of $T_{on}$, the worse the performance of the system both in terms of $T_s$ and $P$, and that different $N_h, N_l$ configurations result in different performance figures, both quantitatively and qualitatively. The figure also shows another effect of non-zero $T_{on}$, namely, that there are some configurations worse than others *under all circumstances*. Indeed, while for the case of $T_{on} = 0$, if one configuration results in a lower delay than other it also results in a larger power consumption, this is no longer true when $T_{on} > 0$. For instance, when $T_{on} = 60$ s, the configurations $N_h = N_l = 5$ and $N_h = N_l = 4$ obtain worse figures of *both* $P$ and $T_s$ than the other two configurations.

On the other hand, Fig. 9 reveals one "positive" aspect of a longer $T_{on}$: given that it takes longer to complete a cycle of the regenerative model, the power consumption increases but the activation rate decreases, which could help to extend the lifetime of the second AP.

### *5.2. Impact of $\rho$*

We next analyze how the trade-offs varies for different values of the network load $\rho$. To this aim, we set $T_{on}$ equal to 30 s and plot the two considered trade-offs for $\rho$ values ranging between 0.05 and 0.95 in steps of 0.05, with the results being depicted in Figs. 10 and 11. For the case of the $T_s$ vs. $P$ trade-off (Fig. 10), we can derive the following main results:

- When the load is very low, there is very little difference between the (de)activation policies, as only one AP is on almost all the time.
- When the load is very high, though, there are non-negligible differences in terms of power (approx. 4%) and, in particular, delay (approx. 30%) between the best and worst performing case. In all cases, the power consumption is very close to 7 W, hinting that

[4] For instance, in our previous experimental works (e.g. [16, 17]) we have experienced faulty behaviour from the power sources due to frequent rebooting of the devices. As a high rate of switching on/off an AP may impact its lifetime, we consider as "very large" values of $\omega$ those in the same order of magnitude as $1/T_{on}$.

[5] Given the tight matching between analytical and simulation values, from now we will only represent the values corresponding to the analysis.

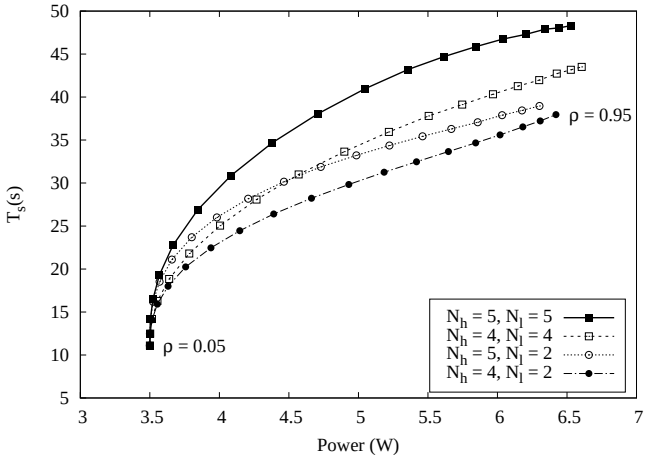


Figure 10: Impact of the load $\rho$ on the $T_s$ vs. $P$ trade-off.

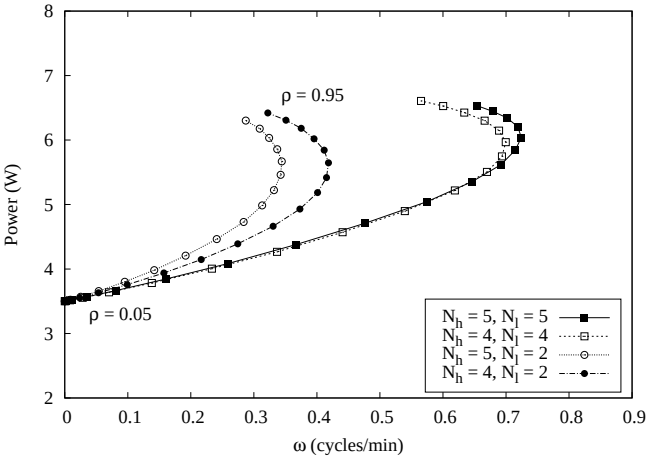


Figure 11: Impact of the load $\rho$ on the $P$ vs. $\omega$ trade-off.

the second AP is on most of the time, either booting (stage $B$) or activated (stage $C$). The difference in delay performance depends on the value of $N_l$: the smaller this value, the longer the system stays in stage $C$, thus providing users with better service.

- Like for the case of $T_{on}$, seen in the previous section, there are qualitatively variations in the $T_s$ vs. $P$ trade-off when $\rho$ changes, as the lines corresponding to different $N_h, N_l$ configurations not only change their slope but also might cross each other.

- Given a $N_h$ value, a configuration without hysteresis ($N_l = N_h$) obtains a poorer performance *both* in terms of $T_s$ and $P$ than the configuration with hysteresis ($N_l < N_h$) for *all* the values of $\rho$.

We next analyze how the $P$ vs. $\omega$ trade-off varies with $\rho$, which is illustrated in Fig. 11 and shows a very-different behavior as compared with the case of the variation with $T_{on}$. Based on the figure, we can derive the following main conclusions:

- Again, for small $\rho$ values, there are little differences between configurations, with $P$ being very close to the use of only one AP and $\omega$ being close to 0.

- As $\rho$ increases, performance worsens for both variables, i.e., there is an increase of the power consumption and the activation rate, the former already seen in the previous figure, while the later being caused by the crossing of the $N_h$ threshold due to the larger load.

- However, once a certain $\rho$ threshold is crossed, power consumption keeps increasing but the activation rate decreases: this is because, the higher the load, the less likely the $N_l$ threshold will be reached, and therefore the system will increase the amount of time with both APs active, which results in a smaller $\omega$.

- Like in the previous case, strategies without hysteresis obtain worse figures than those with hysteresis, since the total power is similar for both types of RoD schemes but the activation rate is much higher when no hysteresis is employed.

### *5.3. Impact of $N_l$*

Finally, we analyze the impact of the configuration of the RoD on performance. To this aim, we fix $\rho = 1/2$ and perform a sweep on $N_l$ for two values of $N_h = \{4, 5\}$. To further analyze the impact of non-zero start-up times on performance, we first consider the case of $T_{on} = 0$, and then the case of $T_{on} = 30$ s.

*Scenario I: $T_{on} = 0$.* The results corresponding to this configuration are depicted in Fig. 12 ($T_s$ vs. $P$) and Fig. 13 ($P$ vs. $\omega$). In both cases, the

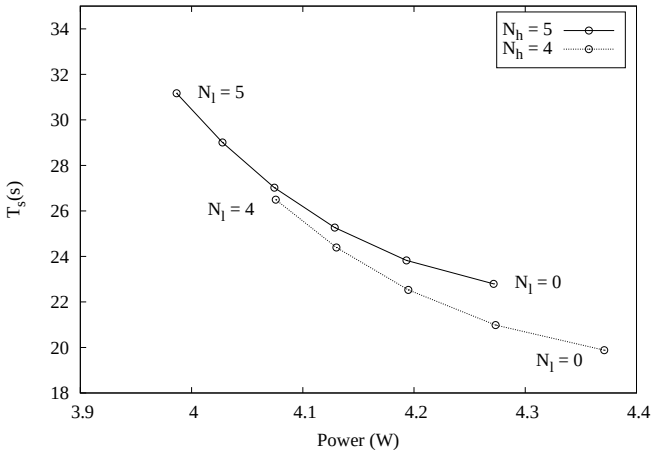


Figure 12: Impact of $N_l$ on the $T_s$ vs. $P$ trade-off, $T_{on} = 0$.

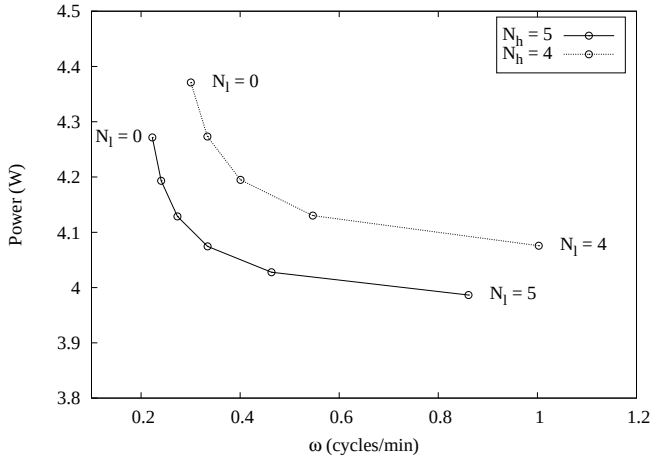


Figure 13: Impact of $N_l$ on the $P$ vs. $\omega$ trade-off, $T_{on} = 0$.

trade-offs are monotonous: given a $N_h$ configuration, decreasing the delay implies increasing the power consumption and, similarly, decreasing the power consumption implies a higher activation rate. Additionally, a higher $N_h$ value results in smaller power consumptions and activation rates.

*Scenario II:* $T_{on} = 30$ *s.* We represent the results corresponding to this configuration in Fig. 14 ($T_s$ vs. $P$) and Fig. 15 ($P$ vs. $\omega$). In this case, there is no monotonous behavior for any of the trade-offs, which further confirms the qualitative impact of having a non-zero $T_{on}$. More specifically, when there is a notable hysteresis (i.e., $N_l \leq 2$), the previous trade-offs still exist, with a decrease of delay (power) resulting in an increase of power (activation rate). However, when there is no or

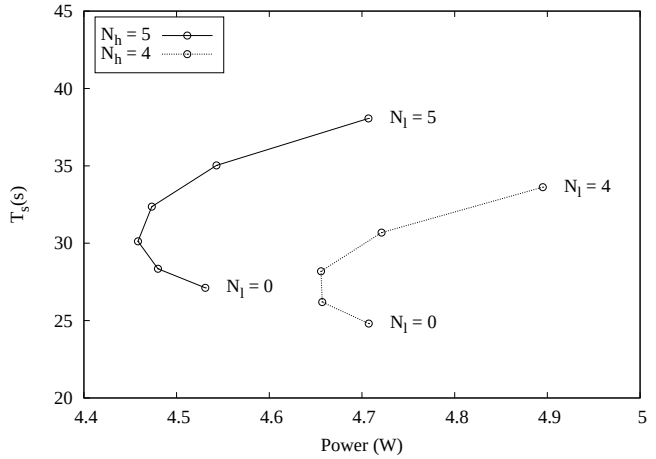


Figure 14: Impact of $N_l$ on the $T_s$ vs. $P$ trade-off, $T_{on} = 30$ s.

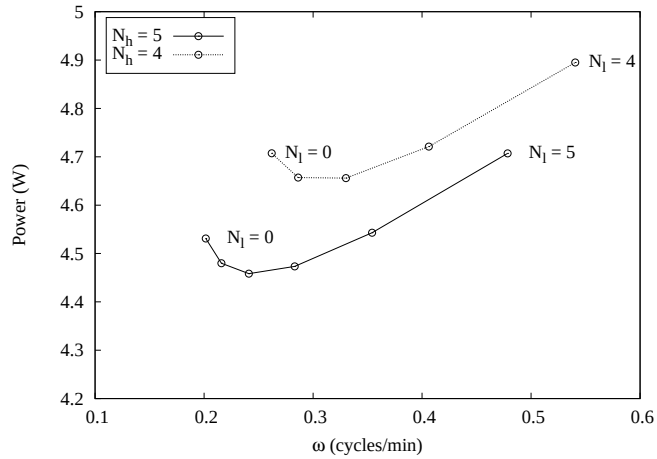


Figure 15: Impact of $N_l$ on the $P$ vs. $\omega$ trade-off, $T_{on} = 30$ s.

very small hysteresis (i.e., $N_l$ closer to $N_h$), the performance worsens for both the delay (power) and the power consumption (activation rate). Additionally, the graph shows that, in general, given a delay bound, a higher value of $N_h$ is preferable since it leads to smaller power consumption.

## 6. Optimal configuration of a RoD scheme

Our model not only serves to analyze the trade-offs in a WLAN implementing a RoD scheme, but also can be used to derive the optimal configuration of its parameters (namely, $N_h$ and $N_l$) for a given scenario (in terms of $\rho$ and $T_{on}$), as we illustrate next. We note that there are many different algorithms and configuration criteria that

**Algorithm 1** Centralized Adaptive Control algorithm

```
Compute T_s^*
Set P_min = ∞
for N_h = 0 ... K do
    for N_l = −1 ... N_h − 1 do
        Compute T_s with (6)
        if T_s < T_s^*(1 + α/100) then
            Compute P with (2)
            if P < P_min then
                P_min ← P
                {N_h^*, N_l^*} ← {N_h, N_l}
            end if
        end if
    end for
end for
Return: {N_h^*, N_l^*}
```

could be used to configure the WLAN, and therefore that our proposal only serves to illustrate one approach.

### *6.1. Optimization Algorithm*

Our optimization criterion is as follows. For our 2-AP setting, the best performance in terms of delay for any $\rho$ value is the one provided when both APs are always active. We denote this minimum average delay as $T_s^*$. Then, we assume that the network administrator is willing to trade-off an increase of this average delay by e.g. $\alpha$ % in exchange for a better power consumption by means of a RoD scheme. To this aim, we sweep on all possible values of $N_h$ and $N_l$, and select that configuration with the minimum power consumption (denoted as $P_{\min}$) among all the ones with an average delay smaller than $(1+\alpha/100)T_s^*$. We denote this configuration as $\{N_h^*, N_l^*\}$.

We summarize the operation of this scheme in Algorithm 1, whose computational complexity is quadratic and relatively small (i.e., it consists on two sweeps over a small number of possible configurations). We note that the sweep includes the configuration with $N_h = 0$ and $N_l = -1$, i.e., the case of the two APs always on, and therefore the algorithm will always provide at least this configuration as a result.

### *6.2. Results*

Fig. 16 shows the optimal configuration for $T_{on} = 0$ (top) and $T_{on} = 30$ s (bottom). In both cases, when the load is low ($\rho < 0.2$), the most efficient strategy is to use large values of $N_h$ (and $N_l$), since the probability that the system reaches a high number of users is very small and therefore there is no need to power on the second AP. These values are almost the same for both the zero and non-zero start-up case.

As load increases, the value of $N_h$ decreases, to ensure that the total delay does not exceed the imposed threshold. Similarly, $N_l$ also decreases to ensure that the second AP is active for enough time to maintain the total delay below the threshold. We note that here the configuration between the zero and non-zero cases changes: when $T_{on}$ is 30 s, higher values of hysteresis (and lower values of $N_h$) are required to keep the total delay below the threshold while minimizing the consumed power, which also leads to a lower activation rate of the second AP. Additionally, the value of $N_h$ is also lower than in the zero start-up case, to prevent situations in which there are many users in the system while the second AP is being powered on. In contrast, when $T_{on}$ is 0 the optimal values of both $N_h$ and $N_l$ are higher, thanks to the better dynamics of the system.

For high loads ($\rho > 0.8$), the value of $N_h^*$ is increased in one unit. This is because $N_h$ and $N_l$ cannot be but natural numbers, and therefore this "rounding" has a notable impact on the resulting configuration. For our considered system and $\alpha$ value, when $\rho$ is above 0.8 both the resulting optimal $N_h^*$, and $N_h^* - 1$ fulfil the condition on the delay (note that it is a relative condition), while the former results in smaller values of the power consumption (this can be seen in Fig. 17, as the increase in the power consumption changes slightly).

We next analyze the power consumption of the optimal strategies for $T_{on} = 0$ and $T_{on} = 30$ s, with the results being depicted in Fig. 17. We note that the configuration with minimum delay corresponds to both APs powered on, i.e., a 7 W consumption for all values of $\rho$, and that the optimal strategy allow an increase of this mini-

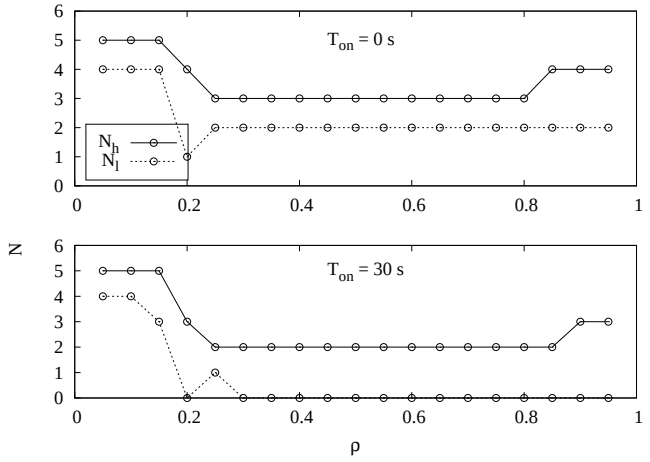


Figure 16: Optimal configuration of $N_h$ and $N_l$ vs. $\rho$ with $T_{on} = 0$ and $T_{on} = 30$ s.

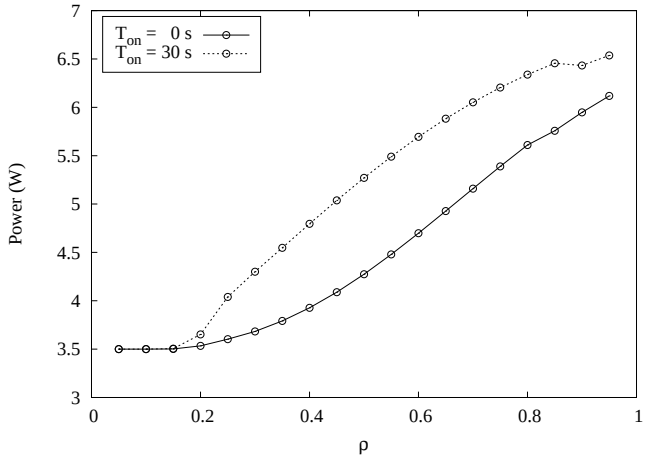


Figure 17: Power consumption with $\{N_h^*, N_l^*\}$, $T_{on} = 0$ and $T_{on} = 30$ s.

mum delay of (at most) 10% in order to minimize power consumption. For low values of the load ($\rho < 0.2$), performance is identical for the zero and non-zero $T_{on}$ cases, as only one AP is powered on. When the load is larger ($\rho \geq 0.2$), there is a notable difference between the two cases, this being caused by the lower values of $N_h$ and $N_l$ for the $T_{on} = 30$ s case, that increase the time spent by the system in stages $B$ and $C$. Compared to the configuration with minimum delay, the savings range between approx. 10% (high load) and 50% (low load), which further motivates the use of RoD schemes.

## 7. 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 have also obtained the optimal configuration of a simple RoD scheme taking into account the start-up time, finding that this time modifies the optimal parameters of the RoD system. As a consequence, we believe that the start-up time should be taken into account when designing infrastructure on demand policies in real-life deployments.

We are currently extending our model to account for a larger number of APs. To this aim, we are building on a semi-Markov process similar to the one illustrated in Fig. 2, but extended for $2N+1$ stages, with $N$ being the number of APs, and with two types of stages: one type when there is one AP being powered on (where the system stays for $T_{on}$), and one when there are no APs being powered on. Our preliminary results show a good accuracy between simulation figures and the numerical analysis, whose complexity is significantly higher than the one presented in this paper.

## Acknowledgments

The work of J. Ortín was partly supported by the Centro Universitario de la Defensa through project CUD2013-05, Gobierno de Aragon (research group T98) and the European Social Fund (ESF). The work of P. Serrano and C. Donato was partly supported by the European Commission under grant agreement H2020-ICT-2014-2-671563 (Flex5Gware) and by the Spanish Ministry of Economy and Competitiveness under grant agreement TEC2014-58964-C2-1-R (DRONEXT)

## References


[1] J. Ortin, P. Serrano, C. Donato, Modeling the Impact of Start-Up Times on the Performance of Resource-on-Demand Schemes in 802.11 WLANs, in: Sustainable Internet and Internet for Sustainability, SustainIT 2015, The 4th IFIP Conference on.

[2] P. Serrano, A. de la Oliva, P. Patras, V. Mancuso, A. Banchs, Greening wireless communications: Status and future directions, Computer Communications 35 (2012) 1651 – 1661.

[3] Y. S. Soh, T. Quek, M. Kountouris, H. Shin, Energy efficient heterogeneous cellular networks, Selected Areas in Communications, IEEE Journal on 31 (2013) 840–850.

[4] A. Jardosh, K. Papagiannaki, E. Belding, K. Almeroth, G. Iannaccone, B. Vinnakota, Green WLANs: On-demand WLAN infrastructures, Mobile Networks and Applications 14 (2009) 798–814.

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

[6] A. P. C. da Silva, M. Meo, M. A. Marsan, Energy-performance trade-off in dense wlans: A queuing study, Computer Networks 56 (2012) 2522 – 2537.

[7] J. Wu, Y. Zhang, M. Zukerman, E.-N. Yung, Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey, Communications Surveys Tutorials, IEEE 17 (2015) 803–826.

[8] L. Budzisz, F. Ganji, G. Rizzo, M. Marsan, M. Meo, Y. Zhang, G. Koutitas, L. Tassiulas, S. Lambert, B. Lannoo, M. Pickavet, A. Conte, I. Haratcherev, A. Wolisz, Dynamic resource provisioning for energy efficiency in wireless access networks: A survey and an outlook, Communications Surveys Tutorials, IEEE 16 (2014) 2259–2285.

[9] M. A. Marsan, M. Meo, Queueing systems to study the energy consumption of a campus WLAN, Computer Networks 66 (2014) 82–93.

[10] R. G. Garroppo, G. Nencioni, G. Procissi, L. Tavanti, The impact of the access point power model on the energy-efficient management of infrastructured wireless lans, Computer Networks in press (2015).

[11] P. Serrano, A. Garcia-Saavedra, G. Bianchi, A. Banchs, A. Azcorra, Per-frame energy consumption in 802.11 devices and its implication on modeling and design, Networking, IEEE/ACM Transactions on PP (2014) 1–1.

[12] M. Papadopouli, H. Shen, 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, pp. 6 pp.–6.

[13] G. R. Hiertz, D. Denteneer, L. Stibor, Y. Zang, X. P. Costa, B. Walke, The IEEE 802.11 Universe, Comm. Mag. 48 (2010) 62–70.

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

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

[16] P. Serrano, P. Patras, A. Mannocci, V. Mancuso, A. Banchs, Control theoretic optimization of 802.11 wlans: Implementation and experimental evaluation, Computer Networks 57 (2013) 258 – 272.

[17] P. Salvador, L. Cominardi, F. Gringoli, P. Serrano, A first implementation and evaluation of the ieee 802.11aa group addressed transmission service, SIGCOMM Comput. Commun. Rev. 44 (2013) 35–41.