Ortin, J., Donato, C., Serrano, P. & Banchs, A. (2016). Resource-on-Demand Schemes in 802.11 WLANs With Non-Zero Start-Up Times. *IEEE Journal on Selected Areas in Communications, 34*(12), 3221–3233.

DOI: 10.1109/jsac.2016.2624158

Resource-on-Demand Schemes in 802.11 WLANs with Non-Zero Start-Up Times

Jorge Ortín, Carlos Donato, Pablo Serrano *Senior member, IEEE*, Albert Banchs, *Senior member, IEEE*

***Abstract*—Increasing the density of access points is one of the most effective mechanisms to cope with the growing traffic demand in wireless networks. To prevent energy wastage at low loads, a resource-on-demand (RoD) scheme is required to opportunistically (de)activate access points as network traffic varies. While previous publications have analytically modelled these schemes in the past, they have assumed that resources are immediately available when activated, an assumption that leads to inaccurate results and might result in inappropriate configurations of the RoD scheme.**

In this paper, we analyse a general RoD scenario with N access points and non-zero start-up times. We first present an exact analytical model that accurately predicts performance but has a high computational complexity, and then derive a simplified analysis that sacrifices some accuracy in exchange for a much lower computational cost. To illustrate the practicality of this model, we present the design of a simple configuration algorithm for RoD. Simulation results confirm the validity of the analyses, and the effectiveness of the configuration algorithm.

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

I. Introduction

To cope with the growing demand of wireless traffic, one of the most effective approaches is to increase the density of access points (AP) in the network. The side effect of this strategy, though, is the increase of the power consumed, which can result in energy wastage if all the infrastructure is kept powered on when the load is low [1], [2]. Techniques to "green" the operation of the network include the design of more energy efficient hardware [3], the optimization of the radio transmission chain [4], or the implementation of resource-on-demand (RoD) strategies that dynamically adapt to the network load, activating resources as it grows and deactivating them when it shrinks [5].

RoD schemes are relatively easy to deploy, as they do not require the introduction of major changes in the network, and have been proposed to decrease the energy consumption of base stations in "traditional" mobile networks (GSM, UMTS) [6], [7], as these devices account for up to 60% of the total energy consumed [3]. Following [8], RoD policies can be divided into static and dynamic, depending on whether the switching on/off of the devices is scheduled or it follows real-time traffic patterns. In general, dynamic approaches are more efficient than static solutions, although they require higher switching on/off rates. In [5], a number of dynamic approaches are classified according to the wireless technology, performance metric, reaction time of the algorithm and control scheme (centralized or distributed).

Regarding WLAN networks, it has been shown that RoD techniques can potentially provide substantial gains in energy savings when the number of considered APs increases (gains of approx. 37%, or 26 kW, can be achieved for a university campus [9] or even higher [10]). Several publications have shown the feasibility of RoD policies in campus networks [11], [12]. The first analytical model for these techniques [13] focuses on the case of "clusters" of APs covering the same area, and studies 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 [14], the work is extended to account for the case when APs do not completely overlap their coverage areas. Following this interest in RoD schemes for WLANs, new publications analyse the performance when some assumptions are relaxed,[1] e.g., in [15] authors analyse the impact of using an accurate energy consumption model on performance.

In this paper, we analyse the impact of start-up times on the performance of a RoD scheme. By "start-up time" we mean the time it takes between the AP is activated until the WLAN is announced. According to the seminal work of [16], typical start-up times of an AP range between 12 and 35 seconds, yet they have not been considered in previous analytical models. However, in our previous work [17], we already showed that for the simple case of 2 APs, start-up times have a notable impact on performance, both qualitatively and quantitatively, as compared to the ideal case of immediate boot times. In that work, we also confirmed that the time to power on an AP is is on the same order of magnitude –approx. 45 s– and practically constant.

We analyse now the general case of a RoD scenario consisting of N overlapping APs with non-zero start-up times. We present an analytic *exact* model that accurately predicts

J. Ortín is with the Centro Universitario de la Defensa and the Aragón Institute of Engineering Research, Universidad de Zaragoza, Spain. E-mail: jortin@unizar.es

C. Donato and A. Banchs are with the Institute IMDEA Networks and the Dept. of Telematics Engineering, Universidad Carlos III de Madrid, Spain. E-mail: carlos.donato@imdea.org, banchs@it.uc3m.es

P. Serrano is with the Dept. of Telematics Engineering, Universidad Carlos III de Madrid, Spain. Email: pablo@it.uc3m.es

The work of C. Donato, P. Serrano and A. Banchs was partly supported by the European Commission in the framework of the H2020-ICT-2014-2 project Flex5Gware (Grant agreement no. 671563) and by the Madrid Regional Government through the TIGRE5-CM program (S2013/ICE-2919). The work of J. Ortín has been supported by Gobierno de Aragón (research group T98) and the European Social Fund (ESF), the EU H2020 Wi-5 project (Grant Agreement No: 644262) and Centro Universitario de la Defensa (project CUD2013-05).

[1] In fact, in both [5] and [8] it is noted that current RoD policies are made using over-simplified models.

performance in terms of energy consumption and service time, but with a high computational complexity. Because of this complexity, we then present a *simplified* model that sacrifices some numerical accuracy in exchange for more affordable computational times. Finally, we present one possible use of this simplified model, namely, the design of a simple configuration algorithm for RoD, based on the minimization of the average service time. As the results show, the simplified model supports the design of optimisation policies that trade-off performance for significant gains in energy efficiency.

II. System Model

A. Scenario

We consider a *cluster model* like the one analysed in [13], consisting of N identical APs serving the same area but using non-overlapping channels. Although in typical high-density deployments the APs may not be located exactly at the same position, the high level of overlapping allows making this assumption, which simplifies the theoretical analysis. Indeed, as will be seen in Section V, this assumption does not impact the validity of the RoD strategy.

The need for a dense deployment such as the one addressed in this paper is motivated by the current trends in the increase of traffic demand. This trend has been forecasted by a number of sources. According to [18], the number of devices and connections per user is steadily growing, which increases user densities; in addition, the throughput required per user is also increasing, as new services such as HD video streaming are becoming ubiquitous. Along the same lines, the forecasts for future 5G networks,[2] predict data rates 100 times higher than today's. Even now, a recent research estimates that typical densities in the deployment of APs may exceed 4000 APs per square kilometre [10].

The scenario considered could be mapped to a very-dense 802.11a setup, where there are many available channels in the 5 GHz band (the specific number depending on the country). One of the APs is always on, in order to maintain the WLAN coverage, while the other APs are opportunistically powered on (off) as users arrive (leave) the system. Powering an AP takes a deterministic time T_{on} and, during this time, the AP is not available, so arriving requests are served by any of the other APs. We neglect the time required to power off an AP.

Each AP consumes P_{AP} units of power when active (i.e., during start-up and when powered on) and zero 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 [19]. A "user" is a new connection generated by a wireless client. Following [20] and [12], these are generated according to a Poisson process at rate λ and are always served by the less loaded AP. Also following [12], we further assume that users' demands are exponentially distributed (i.e., each user downloads an amount of data that is exponentially distributed) and that the AP bandwidth is evenly shared among all the users.[3]

Based on the above assumptions, service times (i.e., the time elapsed since a user arrives to the WLAN until it has fully downloaded its demanded data) would be exponentially distributed (with mean $1/\mu$) if every user got all the bandwidth of an AP, and the service rate (i.e., the inverse of the average service time) is μ when there is only one serving AP, 2μ when there are two APs serving, etc. (i.e., we neglect the impact of channel sharing via contention). The total load is given by $\rho = \lambda/N\mu$.

We also 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 [23], which support triggering re-associations and performing fast transitions, respectively, with minor disruption of the service.

B. Resource on Demand policy

In order to power on/off the APs we assume there is a "target" number $M > 1$ of users per AP, i.e., the system will opportunistically power on/off APs in order to keep that "target" number across resources (except for one AP that will be always on, to guarantee coverage). Based on this, we assume a threshold-based policy with hysteresis, namely:

- An AP will be powered on when the number of user per AP is ρ_h higher than this target value.
- An AP will be powered off when the number of user per AP is ρ_l lower than this target value.

In this way, with K APs powered on, the $K+1$-th AP will be powered on when the number of users reaches

$$\text{Threshold to power on another AP } (N_K)\text{: } \lceil (1+\rho_h)KM \rceil$$

while with K APs powered on, one AP will be powered off when the number of user reaches

$$\text{Threshold to power off an AP } (n_K)\text{: } \lfloor (1-\rho_l)KM \rfloor$$

We next impose some conditions on these thresholds to support an efficient operation. On the one hand, with K APs powered on we impose that there are at least K associated users, so all APs are serving traffic. This results in that the threshold to power off an AP with K users has to be at least K, i.e.,

$$n_K = \lfloor (1-\rho_l)KM \rfloor \geq K \ , \tag{1}$$

which results on the following condition for ρ_l

$$\rho_l < 1 - \frac{1}{M} \ . \tag{2}$$

On the other hand, to prevent (or, at least, reduce) "flip-flop" effects in the WLAN (i.e., to power on an AP and, once

[2] http://5g-ppp.eu/.

[3] The assumption on the Poissonian nature of user arrivals is aligned with the characterisation driven by measurements provided by [21] and [22]. Furthermore, [12] shows that, while the distribution of the duration of user connections is not a memory-less process, it can still be approximated by an exponential distribution with reasonable accuracy. In the numerical evaluation, we will rely on a more accurate traffic model in order to assess the impact of the simplifying assumptions upon which our analysis relies.

active, immediately power it off), we assume that the ρ_h and ρ_l thresholds are set such that the number of users to power on an AP when K of them are already serving traffic is larger than the number of users required to power off an AP when $K+1$ are serving traffic, i.e.,

$$\lceil (1+\rho_h)KM \rceil > \lfloor (1-\rho_l)(K+1)M \rfloor \ . \tag{3}$$

Based on the above condition, and neglecting the rounding operations, to prevent the flip-flop effects the following condition between ρ_h and ρ_l should hold

$$\rho_h > \frac{1-\rho_l}{K} - \rho_l \ , \tag{4}$$

where the rhs of (4) is maximum for $K = 1$, i.e., the case of one AP, and therefore the ρ_h threshold should be set to at least[4]

$$\rho_h > 1 - 2\rho_l \ . \tag{5}$$

In addition to the above, for analytical tractability we introduce the following restriction on the RoD policy: at any point in time there will be at most one AP being powered on. More specifically, while one AP is powering on there will be no decisions taken w.r.t. powering on or off other resources, and only once the AP is available the system will decide on the amount of resources needed.

III. Exact Analysis

A. Model overview

We model the system with the semi-Markov process illustrated in Fig. 1. The label in each arrow corresponds to the number of users (or range of users) in the system that triggers the transition between the corresponding stages. There are four types of *stages*, depending on the transitions that could happen between them:

- Stage 1, which is the initial situation with only one AP active. The only possible transition is to stage $1*$ (another AP is powered on), that is triggered when the number of users reaches N_1.
- Stage N, when all APs are active and serving traffic. The only possible transition is to stage $N-1$ (one AP is powered off), what happens when the number of users is n_N.
- Stages K (with $1 < K < N$), where there are K active APs. In this case there are two possible transitions: one to stage $K*$, triggered when the number of users reaches N_K and another AP is powered on (label N_K in Fig. 1); and other to stage $K-1$, triggered when the number of users in the system is n_K and one AP is powered off (label n_K in Fig. 1).
- Stages $K*$ (with $1 \leq K < N$), where in addition to the K active APs there is another AP booting up. For this type of stage there is a larger number of possible transitions, which are determined by the number of users in the system after T_{on}:
  - If there are N_{K+1} or more users, the system will move to stage $K+1*$, as the number of users is already above the threshold to switch on an additional AP. These transitions are marked with the label $\geq N_{K+1}$ in Fig. 1.
  - If there are between $n_{K+1}+1$ and $N_{K+1}-1$ users, the system will move to stage $K+1$. These transitions are marked with the label (n_{K+1}, N_{K+1}) in Fig. 1.
  - If there are n_{K+1} users or less, the next stage will depend on whether the number of users is also less than or equal to n_2 (and therefore the next stage will be '1'), between n_2+1 and n_3 (the next stage will be '2'), and so on. These transitions are marked with the labels $\leq n_2, (n_2, n_3], \ldots$ in Fig. 1.

[4]Note that for simplicity our policy is set on fixed values of ρ_h and ρ_l, i.e., they do not change with the number of active APs.

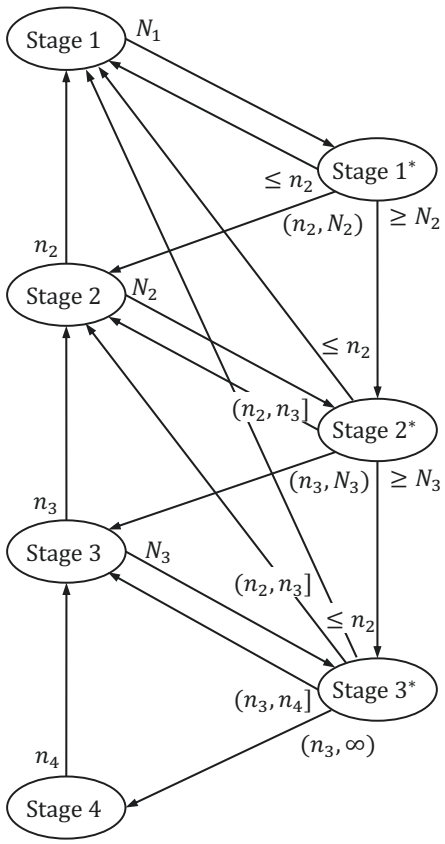


Fig. 1. Semi-Markov process for an IoD scheme with $N = 4$ APs.

With the above, we have introduced the different stages of the semi-Markov process. We next analyse each type of stage, their holding times, and the transition probabilities between them.

B. Modelling the stages of the semi-Markov model

1) Stage 1 (S_1)*:* For the initial situation with only one AP active, following our assumptions the system can be modelled with the continuous-time Markov chain (CTMC) illustrated in Fig. 2, where each state represents the number of users being served and therefore reaching the absorbing state N_1 corresponds to the case when another AP will be powered on (and stage 1 will be left).

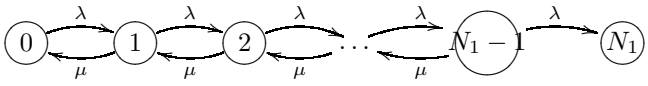


Fig. 2. CTMC for stage 1: one active AP and no AP powering on

Fig. 3. CTMC for Stage K: K active APs and no AP powering on

The average time in this stage H_1 corresponds to the *time until absorption* of the Markov chain, i.e., the time since the system arrived to the chain until it reaches the absorbing state N_1. If we def ne $L_i(t)$ as the expected total time that a CTMC spends in state i during the interval $[0, t)$, H_1 can be expressed as the sum of the terms $L_i(t)$ for all the non-absorbing states of the CTMC when $t \to \infty$ [24]

$$H_1 = \sum_{i=0}^{N_1-1} L_i^{(1)}(\infty) \ . \tag{6}$$

The values of $L_i^{(1)}(\infty)$ (the superscript (1) indicates that we are referring to the CTMC modelling stage 1) can be obtained solving the following system of equations:

$$\mathbf{L}^{(1)}(\infty)\mathbf{Q}^{(1)} = -\boldsymbol{\pi}^{(1)}(0) \ , \tag{7}$$

where

$$\mathbf{L}^{(1)}(\infty) = \left[L_0^{(1)}(\infty), L_1^{(1)}(\infty), \dots L_{N_1-1}^{(1)}(\infty)\right] \ , \tag{8}$$

$$\boldsymbol{\pi}^{(1)}(0) = \left[\pi_0^{(1)}(0), \pi_1^{(1)}(0), \dots \pi_{N_1-1}^{(1)}(0)\right] \ , \tag{9}$$

with $\pi_i^{(1)}(0)$ the initial probability of state i, and $\mathbf{Q}^{(1)}$ a $N_1 \times N_1$ matrix with the following non-zero elements:

$$q_{ij} = \begin{cases} -\lambda & i = 1, j = 1 \\ -\lambda - \mu & i = 2, \dots, N_1, j = i \\ \lambda & i = 1, \dots, N_1 - 1, j = i + 1 \\ \mu & i = 2, \dots, N_1, j = i - 1 \end{cases} \tag{10}$$

The computation of $\boldsymbol{\pi}^{(1)}(0)$ is not straightforward, as it depends on the stage the system was before arriving to stage 1, which could be stage 2 or any other stage $K*$, with $K \geq 1$. We detail how to compute $\boldsymbol{\pi}^{(1)}(0)$ for this and the other cases in the next section, after we present the modelling of the other stages of the semi-Markov process.

Finally, let $P(S_T \mid S_{T'})$ denote the transition probability from stage T' to stage T, with T and T' referring indistinctly to any stage K, including stages 0 and N, or $K*$. For the case of stage 1, we have that

$$P(S_{1*} \mid S_1) = 1 \ . \tag{11}$$

2) Stages K (S_K), $1 < K < N$: For these stages, the resulting CTMC is illustrated in Fig. 3. In this case, while the arrival rate is also λ, the service rate accounts for the total number of powered-on APs, which is constant and equal to $K \cdot \mu$ for all states.[5] As described above, there are two absorbing states: one corresponding to the powering on of another AP (when the system reaches N_K users), and another corresponding to the de-activation of one AP (when the number of users is n_K).

Similarly to the previous case, the average time in a stage K can be computed as

$$H_K = \sum_{i=n_K+1}^{N_K-1} L_i^{(K)}(\infty) \ . \tag{12}$$

In order to compute $L_i^{(K)}(\infty)$ we use the same expression as in the previous case

$$\mathbf{L}^{(K)}(\infty)\mathbf{Q}^{(K)} = -\boldsymbol{\pi}^{(K)}(0) \ , \tag{13}$$

where

$$\mathbf{L}^{(K)}(\infty) = \left[L_{n_K+1}^{(K)}(\infty), L_{n_K+2}^{(K)}(\infty), \dots L_{N_K-1}^{(K)}(\infty)\right] \ , \tag{14}$$

$$\boldsymbol{\pi}^{(K)}(0) = \left[\pi_{n_K+1}^{(K)}(0), \pi_{n_K+2}^{(K)}(0), \dots \pi_{N_K-1}^{(K)}(0)\right] \ , \tag{15}$$

and $\mathbf{Q}^{(K)}$ is a $(N_K - n_K - 1) \times (N_K - n_K - 1)$ matrix, whose non-zero elements are

$$q_{ij} = \begin{cases} -\lambda - K\mu & i = 1, \dots, N_K - n_K - 1, j = i \\ \lambda & i = 1, \dots, N_K - n_K - 2, j = i + 1 \\ K\mu & i = 2, \dots, N_K - n_K - 1, j = i - 1 \end{cases} \tag{16}$$

Again, the computation of $\boldsymbol{\pi}^{(K)}(0)$ requires the knowledge of the stage of the system before entering stage K, which we will address in the next section.

To f nalise the analysis of this stage, we have to compute the two transition probabilities from this stage to stage $K*$ (when the chain ends in the absorbing state N_K) and to stage $K-1$ (when the chain falls into the absorbing state n_K), denoted as $P(N_K)$ and $P(n_K)$ respectively,

$$P(S_{K*} \mid S_K) = P(N_K) \ , \tag{17}$$

$$P(S_{K-1} \mid S_K) = P(n_K) \ . \tag{18}$$

These can be computed as [25]

$$[P(n_K)\ P(N_K)] = \boldsymbol{\pi}^{(K)}(0)\mathbf{B}^{(K)} \ , \tag{19}$$

with $\mathbf{B}^{(K)}$ a $(N_K - n_K - 1) \times 2$ matrix whose element b_{ij} is the probability of ending in the absorbing state j, given that the chain starts in the transient state i. This matrix can be computed as

$$\mathbf{B}^{(K)} = \left[\mathbf{I} - \mathbf{T}^{(K)}\right]^{-1} \mathbf{R}^{(K)} \ , \tag{20}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{T}^{(K)}$ is a $(N_K - n_K - 1) \times (N_K - n_K - 1)$ matrix with the transition probabilities between non-absorbing states, and $\mathbf{R}^{(K)}$ is a $(N_K - n_K - 1) \times 2$ matrix

[5] Note that we have imposed $n_K > K$ with (1).

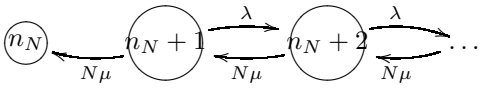


Fig. 4. Markov chain when all APs are active

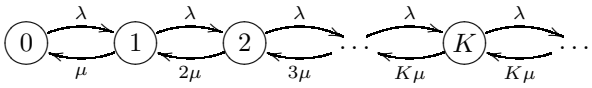


Fig. 5. Markov chain for the case of K active APs and one AP powering on

denoting the transition probabilities from non-absorbing to absorbing states. Both matrices are obtained from the associated discrete-time Markov chain (DTMC) of the CTMC, and their non-zero elements are

$$t_{ij} = \begin{cases} \frac{\lambda}{\lambda+K\mu} & i = 1,\ldots,N_K - n_K - 2, j = i+1 \\ \frac{K\mu}{\lambda+K\mu} & i = 2,\ldots,N_K - n_K - 1, j = i-1 \end{cases} \quad (21)$$

$$r_{ij} = \begin{cases} \frac{K\mu}{\lambda+K\mu} & i = 1, j = 1 \\ \frac{\lambda}{\lambda+K\mu} & i = N_K - n_K - 1, j = 2 \end{cases} \quad (22)$$

3) Stage N (S_N): When all APs are active and serving traffic the resulting CTMC is the one depicted in Fig. 4, where the arrival rate is λ and the service rate is $N \cdot \mu$. As in the case of stage 1, there is only one absorbing state, the one corresponding to the switching off of one AP when there are n_N users in the system and all the APs are on, but now the chain has an infinite number of states.

To compute the holding time in this stage, we assume that the system is stable (i.e., $\lambda < N\mu$), so there is a state n_D with $n_D > n_N$ such that

$$\sum_{i=n_D+1}^{\infty} L_i^{(N)}(\infty) \approx 0 \ , \quad (23)$$

and therefore the holding time is

$$H_N \approx \sum_{i=n_N+1}^{n_D} L_i^{(N)}(\infty) \ , \quad (24)$$

where $\mathbf{L}^{(N)}(\infty)$ is obtained from

$$\mathbf{L}^{(N)}(\infty)\mathbf{Q}^{(N)} = -\boldsymbol{\pi}^{(N)}(0) \ , \quad (25)$$

with

$$\mathbf{L}^{(N)}(\infty) = \left[L_{n_N+1}^{(N)}(\infty), L_{n_N+2}^{(N)}(\infty), \ldots L_{N_D}^{(N)}(\infty)\right], \quad (26)$$

$$\boldsymbol{\pi}^{(N)}(0) = \left[\pi_{n_N+1}^{(N)}(0), \pi_{n_N+2}^{(N)}(0), \ldots \pi_{n_D}^{(N)}(0)\right], \quad (27)$$

and $\mathbf{Q}^{(N)}$ is a $(n_D - n_N) \times (n_D - n_N)$ matrix with the following non-zero elements

$$q_{ij} = \begin{cases} -\lambda - N\mu & i = 1,\ldots,n_D - n_N, j = i \\ \lambda & i = 1,\ldots,n_D - n_N - 1, j = i+1 \\ N\mu & i = 2,\ldots,n_D - n_N, j = i-1 \end{cases} \quad (28)$$

The computation of $\boldsymbol{\pi}^{(N)}(0)$ is described in the next section. From this stage, the only possible transition is to stage $N-1$, i.e.,

$$P(S_{N-1} \mid S_N) = 1 \ . \quad (29)$$

4) Stages $K*$ (S_{K*}): For the stages with K active APs and one AP being powered on, the resulting Markov chain is illustrated in Fig. 5. In these stages there are no absorbing states that trigger the transition to other stages, since this happens when the amount of time spent in the stage is T_{on}. Because of this, the number of users that can be in the system during this stage varies between zero an infinity. Additionally, the service rate depends on the number of users as there should be at least one user per active AP for the total rate to be $K\mu$.[6] For completeness, the time spent in a stage $K*$ is given by

$$H_{K*} = T_{on} \ . \quad (30)$$

In this case, we need to obtain the expected time that the system spends in each state i during the T_{on} seconds that a stage $K*$ lasts, $L_i^{(K*)}(T_{on})$, and the probability of each state after T_{on}, $\pi_i^{(K*)}(T_{on})$. These terms are required to compute the transition probabilities from stage $K*$ to the other stages and to obtain the performance figures of the system. The values for $L_i^{(K*)}(T_{on})$ and $\pi_i^{(K*)}(T_{on})$ can be obtained with the expressions of the transient analysis of an M/M/K queue, which are

$$L_i^{(K*)}(t) = \int_0^t \pi_i^{(K*)}(u)du \ , \quad (31)$$

where $\pi_i^{(K*)}(t)$ is the probability of being in state i at time t, which is determined by the fundamental equations of the CTMC

$$\frac{d\boldsymbol{\pi}^{(K*)}(t)}{dt} = \boldsymbol{\pi}^{(K*)}(t)\,\mathbf{Q}^{(K*)} \ , \quad (32)$$

with $\boldsymbol{\pi}^{(K*)}(t) = [\pi_i^{(K*)}(t)]_i$ the transient state probability vector and $\mathbf{Q}^{(K*)}$ the infinitesimal generator matrix of the CTMC. The non-zero elements of $\mathbf{Q}^{(K*)}$ are

$$q_{ij} = \begin{cases} -\lambda - (i-1)\mu & i = 1,\ldots,K, j = i \\ -\lambda - K\mu & i = K+1,\ldots, j = i+1 \\ \lambda & i = 1,...,j = i+1 \\ (i-1)\mu & i = 2,\ldots,K, j = i-1 \\ K\mu & i = K+1,\ldots, j = i-1 \end{cases} \quad (33)$$

Note that to solve (31) and (32), we need again the vector of initial state probabilities $\boldsymbol{\pi}^{(K*)}(0)$. On the other hand, as there are no closed expressions for the transient behaviour of an M/M/K queue, we need to use approximate methods (such as *uniformization* [24]) to solve it and compute $L_i(T_{on})$ and $\pi_i(T_{on})$. Like in the previous cases, the computation of $\boldsymbol{\pi}^{(S_K*)}(0)$ is described in the next section.

[6] In fact, the CTMC corresponds to the classic M/M/K queue

Finally, as noted before, from stage $K*$ the system can go to any other stage K', with $K' \leq K+1$, and to stage $K+1*$. In this way, after T_{on} the system can have K or less APs powered on, with the following probabilities

$$P(S_{K'} \mid S_{K*}) = \begin{cases} \sum_{i=0}^{n_2} \pi_i^{(K*)}(T_{on}), & K' = 1 \\ \sum_{i=n_{K'+1}}^{n_{K'+1}} \pi_i^{(K*)}(T_{on}), & 1 < K' \leq K \end{cases} \tag{34}$$

while the probability of having more APs on (or being powered on) after T_{on} depends on whether there are more APs to be powered on, i.e., if $K < N-1$, or not ($K = N-1$). For the former case, we have that

$$P(S_{K+1} \mid S_{K*}) = \sum_{i=n_{K+1}+1}^{N_{k+1}-1} \pi_i^{(K*)}(T_{on}) \ , \tag{35}$$

$$P(S_{K+1*} \mid S_{K*}) = \sum_{i=N_{k+1}}^{\infty} \pi_i^{(K*)}(T_{on}) \ , \tag{36}$$

while for the case of $K = N-1$ there are no more APs to activate, and therefore

$$P(S_N \mid S_{N-1*}) = \sum_{i=n_N+1}^{\infty} \pi_i^{(N-1*)}(T_{on}) \ . \tag{37}$$

C. Computing the steady-state distribution

To complete the analysis of the steady-state distribution of the semi-Markov process, we have to express the set of initial conditions for every stage in terms of the final state probabilities of the other stages. To this end, we denote $\pi_i^{(T)}(0)$ as the probability that the initial state is i for the stage T (again we use T for generalization purposes, with stage T we refer indistinctly to any stage K, including stages 0 and N, or $K*$). This probability can be computed with the law of total probability as

$$\pi_i^{(T)}(0) = \sum_{S_{T'} \in \mathcal{T}_T} \pi_i^{(T|T')}(0) P_T(T') \ , \tag{38}$$

where

- $\mathcal{T}_T$ is the set of stages that can reach stage T in one transition between stages,
- $\pi_i^{(T|T')}(0)$ is the probability that the initial state of the CTMC modelling stage T is i, given that the system was in stage T' and transitioned to state T,
- $P_T(T')$ is the probability that the system was in stage T' before the stage transition, given that it is now in stage T.

The set $\mathcal{T}$ can be easily derived for each stage from the Semi-Markov model described in Section III-A. Specifically we have,

$$\mathcal{T}_1 = \{S_2, S_{1*}, \dots S_{N-1*}\} \, , \tag{39}$$

$$\mathcal{T}_K = \{S_{K+1}, S_{K-1*}, \dots S_{N-1*}\} \ \text{for } 1 < K < N, \tag{40}$$

$$\mathcal{T}_N = \{S_{N-1*}\} \, , \tag{41}$$

$$\mathcal{T}_{1*} = \{S_1\} \, , \tag{42}$$

$$\mathcal{T}_{K*} = \{S_K, S_{K-1*}\} \ \text{for } 1 < K < N-1 \ . \tag{43}$$

As an example, for the case of Fig. 1 with 4 APs, we have $\mathcal{T}_1 = \{S_2, S_{1*}, S_{2*}, S_{3*}\}$, $\mathcal{T}_2 = \{S_3, S_{1*}, S_{2*}, S_{3*}\}$, $\mathcal{T}_3 = \{S_4, S_{2*}, S_{3*}\}$, $\mathcal{T}_4 = \{S_{3*}\}$, $\mathcal{T}_{1*} = \{S_1\}$, $\mathcal{T}_{2*} = \{S_2, S_{1*}\}$, $\mathcal{T}_{3*} = \{S_3, S_{2*}\}$.

The computation of $\pi_i^{(T|T')}(0)$ depends on whether stage T' corresponds to a stage with an AP being powered on or not. For the latter case, the transition is triggered because the number of stations reached a (de)activation threshold (i.e., an absorbing state), and therefore we have

$$\pi_i^{(K*|K)}(0) = \begin{cases} 1, & i = N_K \\ 0, & \text{otherwise} \end{cases} \tag{44}$$

for $1 \leq K < N$ (note that we have included here the transition from stage 1 to stage $1*$ as well), and

$$\pi_i^{(K-1|K)}(0) = \begin{cases} 1, & i = n_K \\ 0, & \text{otherwise} \end{cases} \tag{45}$$

for $1 < K \leq N$ (we have included the transition from stage N to stage $N-1$ as well). On the other hand, when stage T' is a $K*$ stage, there are multiple states that can result in a transition to a stage, which results in the following cases:
(i) If the transition is to stage 1 ($S_T = S_1$), then

$$\pi_i^{(1|K*)}(0) = \begin{cases} \frac{\pi_i^{(K*)}(T_{on})}{\sum_{j=0}^{n_2} \pi_j^{(K*)}(T_{on})}, & 0 \leq i \leq n_2 \\ 0, & n_2 < i < N_1 \end{cases} \tag{46}$$

(ii) If the transition is to a stage $1 < K' \leq K$, then

$$\pi_i^{(K'|K*)}(0) = \begin{cases} \frac{\pi_i^{(K*)}(T_{on})}{\sum_{j=n_{K'+1}}^{n_{K'+1}} \pi_j^{(K*)}(T_{on})}, & n_{K'} < i \leq n_{K'+1} \\ 0, & n_{K'+1} < i < N_{K'} \end{cases} \tag{47}$$

(iii) If $K < N-1$ (i.e. $S_{T'} \neq S_{N-1}$) and the transition is to stage $K+1$, then

$$\pi_i^{(K+1|K*)}(0) = \frac{\pi_i^{(K*)}(T_{on})}{\sum_{j=n_{K+1}+1}^{N_{K+1}-1} \pi_j^{(K*)}(T_{on})} \ . \tag{48}$$

(iv) If $K < N-1$ (again $S_{T'} \neq S_{N-1}$) and the transition is to stage $K+1*$, then

$$\pi_i^{(K+1|K*)}(0) = \begin{cases} 0, & 0 \leq i \leq N_{K+1} \\ \frac{\pi_i^{(K*)}(T_{on})}{\sum_{j=N_{K+1}}^{\infty} \pi_j^{(K*)}(T_{on})}, & i > N_{K+1} \end{cases} \tag{49}$$

(v) If $K = N-1$ and the transition is to stage N, then

$$\pi_i^{(N|N-1*)}(0) = \frac{\pi_i^{(N-1*)}(T_{on})}{\sum_{j=n_N+1}^{\infty} \pi_j^{(K*)}(T_{on})} \ . \tag{50}$$

Finally, the computation of $P_T(T')$ can be done with the law of total probability again

$$P_T(T') = \frac{P(S_T \mid S_{T'}) \phi_{T'}}{\sum_{S_Q \in \mathcal{T}_T} P(S_T \mid S_Q) \phi_Q} \ , \tag{51}$$

where $P(S_T \mid S_{T'})$ denotes the stage transition probability computed in (11), (17), (18), (29), (34)-(37), and ϕ_T is the stationary probability of stage T in the embedded Markov

chain of the semi-Markov process. The computation of ϕ_T is done via the system

$$\boldsymbol{\phi} = \boldsymbol{\phi}\mathbf{P} \ , \tag{52}$$

where $\boldsymbol{\phi}$ is a row vector whose components are the values of ϕ_T, and $\mathbf{P}$ is a matrix composed of the stage transition probabilities of the embedded Markov chain.

With the above, we have completed the analysis that enables in the next section the computation of the steady state probabilities of the semi-Markov model. We also address there how to compute performance f gures based on these probabilities.

D. Performance f gures

We characterise the performance of the system with two f gures:

- The average power consumed by the infrastructure P.
- The average service time of a user T_s.

The average power consumed by the infrastructure can be expressed in terms of the average number of APs that are powered on, N_{AP}, as follows

$$P = N_{AP} P_{AP} \ . \tag{53}$$

N_{AP} is computed as the weighted sum of the number of APs powered on in each stage times the probability of being in that stage

$$N_{AP} = \sum_{K=1}^{N} K\mathcal{P}_K + \sum_{K=1}^{N-1} (K+1)\mathcal{P}_{K*} \ , \tag{54}$$

where $\mathcal{P}_K$ and $\mathcal{P}_{K*}$ are the stationary probabilities of the stages of the semi-Markov process, i. e., the probability of being in stage K (including stages 0 and N) or $K*$ at a specif c moment. These probabilities are related to the stage probabilities of the embedded Markov chain as follows

$$\mathcal{P}_T = \frac{H_T \phi_T}{\sum_{K=1}^{N} \phi_K H_K + \sum_{K=1}^{N-1} \phi_{K*} H_{K*}} \ . \tag{55}$$

The average service time T_s, which corresponds to the time between the instant when a user generates a service request and when this request is completely served, can be obtained via Little's formula

$$T_s = \frac{N_u}{\lambda} \ , \tag{56}$$

with N_u the average number of users in the system. This can be computed with the law of total probability as follows

$$N_u = \sum_{i=1}^{\infty} i \left(\sum_{K=1}^{N} \pi_i^{(K)} \mathcal{P}_K + \sum_{K=1}^{N-1} \pi_i^{(K*)} \mathcal{P}_{K*} \right) \ , \tag{57}$$

where $\pi_i^{(K)}$ and $\pi_i^{(K*)}$ are the average probabilities of having i users, given that the system is in stage K or $K*$, respectively. This can be computed, for each type of stage, as

$$\pi_i^{(K)} = \frac{L_i^{(K)}(\infty)}{H_K} \ , \tag{58}$$

$$\pi_i^{(K*)} = \frac{L_i^{(K*)}(T_{on})}{T_{on}} \ . \tag{59}$$

Algorithm 1 Solution to the exact model

1: Set initial estimations of $\boldsymbol{\pi}^{(K)}(0)$ and $\boldsymbol{\pi}^{(K*)}(0)$
2: **repeat**
3: Compute $\mathbf{L}^{(K)}(\infty)$ with (7), (13) and (25)
4: Obtain H_K with (6), (12) and (24)
5: Solve (31)-(32) to obtain $\mathbf{L}^{(K*)}(T_{on})$ and $\boldsymbol{\pi}^{(K*)}(T_{on})$
6: Compute $P(T \mid T')$ with (17)-(20) and (34) - (37)
7: Solve (52) to obtain $\boldsymbol{\phi}$
8: Obtain $P_T(T')$ with (51)
9: Compute $\pi_i^{(T|T')}(0)$ with (44)-(50)
10: Update $\boldsymbol{\pi}^{(K)}(0)$ and $\boldsymbol{\pi}^{(K*)}(0)$ with (38)
11: **until** Stopping criterion is met
12: Obtain $\pi_i^{(K)}$ and $\pi_i^{(K*)}$ with (58) and (59)
13: Compute $\mathcal{P}_T$ with (55)
14: Obtain N_u with (57) and N_{AP} with (54)
15: Compute T_s with (56) and P with (53)

As can be seen, all the performance metrics depend on the variables ϕ_T, H_T, $L_i^{(K)}(\infty)$ and $L_i^{(K*)}(T_{on})$, whose relationships have been described through Sections III-B to III-C. In order to obtain an exact solution for them, we should solve a system of non-linear equations with the additional problem that there are no closed expressions for the transient analysis of the CTMC modelling stages $K*$. To solve this, we propose the iterative algorithm described in *Algorithm 1*. In this algorithm, the initial values of $\boldsymbol{\pi}^{(K)}(0)$ can be set assuming that all the states with non-zero probabilities according to (46)-(50) have the same initial probability. Regarding $\boldsymbol{\pi}^{(K*)}(0)$, a good starting guess is to assume that $\pi_i^{(K*)}(0) = 1$ for $i = N_K$ and 0 otherwise (this is what would happen if $T_{on} = 0$). Finally, a common stopping criterion is that the norm of the vector difference between the old and updated version of vectors $\boldsymbol{\pi}^{(K)}(0)$ and $\boldsymbol{\pi}^{(K*)}(0)$ is below a threshold ϵ.

IV. Simplified Analysis

A. Motivation and simplif cation

The main weaknesses of the model derived in the previous section is that the initial probabilities of a stage depends on the "f nal" probabilities of the rest of stages, which depend in turn of their initial probabilities. This causes a loop that requires the use of an iterative algorithm with non-negligible computational complexity as the one described above. We next describe how to simplify the analytical model of Section III to enable an eff cient computation of the performance f gures, at the cost of some numerical inaccuracy.

The proposed simplif cation affects exclusively the transitions from stages $K*$. As can be seen in Fig. 1, from these stages the system could go to stage $K+1*$ or any other stage K', with $K' \leq K+1$. The direct transitions between stages $K*$ make that the initial state probabilities for these stages $\pi_i^{(K*)}(0)$ could be non-zero for $i \geq N_K$.

To break the coupling between stages $K*$, we assume that the initial state probabilities of stages $K*$ are f xed and equal to

$$\pi_i^{(S_{K*})}(0) = \begin{cases} 1, & i = N_K \\ 0, & \text{otherwise} \end{cases} \tag{60}$$

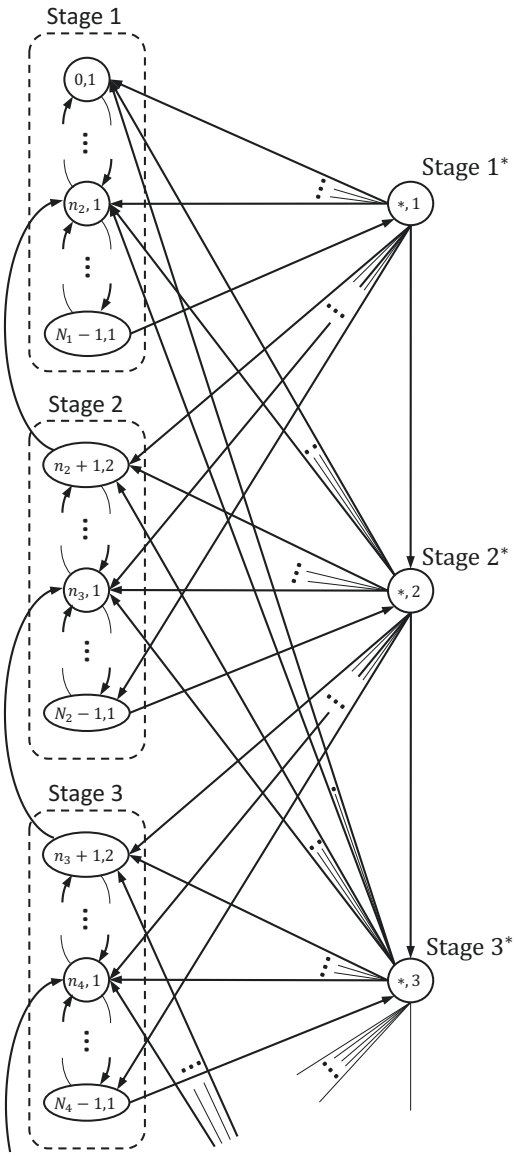


Fig. 6. Simplified model.

This implies that the system enters into stages $K*$ always with N_K users. This assumption holds as long as the transition probability between a stage $K*$ and the stage $K+1*$ is small, which is true for typical T_{on} values.

Once this assumption is made, the transition probabilities from stages $K*$ to other stages are fixed and independent of the initial state probabilities of the rest of stages. Now, we also have to tackle the same apparent coupling for the initial state probabilities of stages K. To solve this, we build a new semi-Markov model derived from the one depicted in Fig. 1 substituting stages K by the embedded DTMC of their corresponding CTMC. The description of this new model is performed in the next Section.

B. Model description

Fig. 6 shows the embedded DTMC of the semi-Markov process described above. The leftmost states, which model stages K (including 0 and N), are defined by the pair (i, K), with i the number of users in the system and K the number of powered-on APs. The holding time of these states is an exponential random variable with mean $(\lambda + K\mu)^{-1}$. The rightmost states model the stages $K*$ and their holding time is constant and equal to T_{on}. To keep a uniform notation, we note these states as $(*, K)$. The non-null transitions probabilities are described in (61).

The first two equations model the transitions between states of Stage 1, the first one corresponds to the departure of a user and the second one its arrival. The third and fourth equations model the transitions between states of stages $2 \leq N \leq N-1$ and the fifth and sixth the transitions between states of stage N. The seventh equation corresponds to the switch off of an AP when a user departures and stage K remains with n_K users, which triggers the transition to stage $K-1$. The eighth equation models the switching on of a new AP (i.e. the transition to stage $K*$) when the N_K-th user arrives and K APs are on. Note that in all the cases the transition probabilities only depend on the parameters λ, μ and the number of APs that are serving traffic at the moment of the transition.

The next equations model the transitions from states $(*, K)$, (i.e., from stages $K*$). Now the transition probabilities are of the form $\pi_i^{(K*)}(T_{on})$ and can be computed solving (31) and (32) assuming the initial state probabilities given in (60). Specifically, the ninth equation corresponds to the transition from stage $K*$ to stage $K+1*$ because the system reaches N_{K+1} users during the booting-up of the $K+1$-th AP. The tenth equation models the transition from stage $K*$ to a state where there are $K+1$ APs powered on and a number of users ranging between $n_{K+1}+1$ and $N_{K+1}-1$. The eleventh equation is similar to the previous one but for stage $N-1*$. In this case, there is no upper limit in the number of users since there is not any remaining AP to boot up. The twelfth equation models the transition from stage $K*$ to states where the number of APs on is below $K+1$. This implies that during the booting up of the $K+1$-th AP several users have left forcing the system to switch off some APs. The last equation is similar to the previous one and corresponds to transitions to states where only one AP is on.

With the previous equations, the DTMC can be easily solved to obtain the stationary distribution of the state probabilities, that we name $P(i, K)$ (or $P(*, K)$) hereafter. With these, the stationary probability of each state of the semi-Markov process, $\Phi(i, K)$ (or $\Phi(*, K)$) are

$$\Phi(i,1) = \frac{P(i,1)}{\Omega \cdot (\lambda + \mu)}, \quad 0 \leq i < N_1 \tag{62}$$

$$\Phi(i,K) = \frac{P(i,K)}{\Omega \cdot (\lambda + K\mu)}, \; n_K < i < N_K, \; 1 < K < N \tag{63}$$

$$\Phi(i,N) = \frac{P(i,N)}{\Omega \cdot (\lambda + N\mu)}, \quad i > n_N \tag{64}$$

$$\Phi(*,K) = \frac{P(*,K)T_{on}}{\Omega}, \quad 1 \leq K < N \tag{65}$$

$$\begin{cases} P(i-1,0 \mid i,0) = \frac{\mu}{\lambda+\mu} & i = \{1,\ldots,N_1-1\} \\ P(i+1,0 \mid i,0) = \frac{\lambda}{\lambda+\mu} & i = \{0,\ldots,N_1-2\} \\ P(i-1,K \mid i,K) = \frac{K\mu}{\lambda+K\mu} & i = \{n_K+2,\ldots,N_K-1\}, K = \{2,\ldots,N-1\} \\ P(i+1,K \mid i,K) = \frac{\lambda}{\lambda+K\mu} & i = \{n_K+1,\ldots,N_K-2\}, K = \{2,\ldots,N-1\} \\ P(i-1,N \mid i,N) = \frac{N\mu}{\lambda+N\mu} & i = \{n_N+2,\ldots\} \\ P(i+1,N \mid i,N) = \frac{\lambda}{\lambda+N\mu} & i = \{n_N+1,\ldots\} \\ P(n_K,K-1 \mid n_K+1,K) = \frac{K\mu}{\lambda+K\mu} & K = \{2,\ldots,N\} \\ P(*,K \mid N_K-1,K) = \frac{\lambda}{\lambda+K\mu} & K = \{1,\ldots,N-1\} \\ P(*,K+1 \mid *,K) = \sum_{i=N_{K+1}}^{\infty} \pi_i^{(K*)}(T_{on}) & K = \{1\ldots,N-2\} \\ P(i,K+1 \mid *,K) = \pi_i^{(K*)}(T_{on}) & i = \{n_{K+1}+1,\ldots,N_{K+1}-1\}, K = \{1\ldots,N-2\} \\ P(i,N \mid *,N-1) = \pi_i^{(N-1*)}(T_{on}) & i = \{n_N+1,\ldots,\} \\ P(i,K' \mid *,K) = \pi_i^{(K*)}(T_{on}) & i = \{n_{K'}+1,\ldots,n_{K'+1}\}, K = \{2\ldots,N-1\}, K' = \{2\ldots,K\} \\ P(i,1 \mid *,K) = \pi_i^{(K*)}(T_{on}) & i = \{0,\ldots,n_2\}, K = \{1\ldots,N-1\} \end{cases} \quad (61)$$

with

$$\Omega = \sum_{j=0}^{N_1-1} \frac{P(j,1)}{\lambda+\mu} + \sum_{K'=2}^{N-1} \sum_{j=n_{K'}+1}^{N_{K'}-1} \frac{P(j,K')}{\lambda+K'\mu} + \sum_{j=n_N+1}^{\infty} \frac{P(j,N)}{\lambda+N\mu} + \sum_{K'=1}^{N-1} P(*,K')T_{on} \ . \quad (66)$$

The stationary probabilities of stages $K*$ are directly $\mathcal{P}_{K*} = \Phi(*,K)$, while for stages K we have

$$\mathcal{P}_1 = \sum_{i=0}^{N_1-1} \Phi(i,1) \ , \quad (67)$$

$$\mathcal{P}_K = \sum_{i=n_K+1}^{N_K-1} \Phi(i,K) \ , \quad (68)$$

and

$$\mathcal{P}_N = \sum_{i=n_N+1}^{\infty} \Phi(i,N) \ . \quad (69)$$

Once these terms are known, we can compute the average power P with (55) and (54). The average service time T_s is obtained with (56) as well, but in this case N_u is

$$N_u = \sum_{i=1}^{\infty} i \left(\sum_{K=1}^{N} \Phi(i,K) + \sum_{K=1}^{N-1} \pi_i^{(K*)} \mathcal{P}_{K*} \right) \ , \quad (70)$$

with $\pi_i^{(K*)}$ the same as in (59).

To end this Section, we present in *Algorithm 2* the different steps required to obtain the performance figures of the system. As can be seen, in this case we avoid the presence of loops.

Algorithm 2 Solution to the approximate model

1: Set $\boldsymbol{\pi}^{(K*)}(0)$ with (60)
2: Solve (31) and (32) to obtain $\mathbf{L}^{(K*)}(T_{on})$ and $\boldsymbol{\pi}^{(K*)}(T_{on})$
3: Solve the DTMC with transitions given by (61) to obtain $P(i,K)$ and $P(*,K)$
4: Compute (62)-(66) to obtain $\Phi(i,K)$ and $\Phi(*,K)$
5: Obtain $\mathcal{P}_1$, $\mathcal{P}_K$ and $\mathcal{P}_N$ with (67)-(69)
6: Obtain N_u with (70) and N_{AP} with (54)
7: Compute T_s with (56) and P with (53)

V. Numerical Results

We next present a numerical evaluation of a RoD system in terms of the performance figures considered, namely, the average service time T_s and the power consumed by the infrastructure P. To this end, we compute these two variables for a variety of scenarios, these being defined in terms of the network load or the configuration of the RoD scheme (given by the parameters M, ρ_h, ρ_l). In the simulation results presented, we compare the results of our approximate model against the ones obtained via simulation,[7] while in Section V-C we assess the computational complexity of this model against the accurate one.

Throughout all simulations, we consider the following scenario:[8] (i) various APs can be simultaneously activated (instead of only one, as assumed in the analysis); (ii) there is no complete overlap of the coverage areas: we assume a deployment centred around one AP with a 10 m coverage radio that is always on, and $N-1$ APs with the same coverage radios that are randomly deployed within a 4 m circle centred around the first AP and that will be opportunistically (de)activated; and (iii) users are not static but follow the classical *random waypoint model* [26], selecting a novel destination at random after reaching the previous one, and moving at a speed that is randomly chosen between 0.3 and 0.7 m/s. We further assume

[7] Our approximate model is solved numerically using `Octave` (https://www.gnu.org/software/octave/), while simulation results are obtained from a discrete event simulator written in C++.

[8] Note that this scenario relaxes some of the simplifying assumptions behind our model, and thus allows to assess the impact of such assumptions on the results.

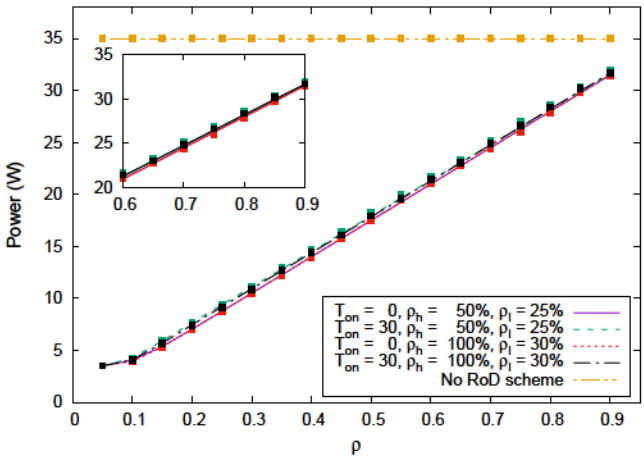


Fig. 7. Average power consumption vs. network load.

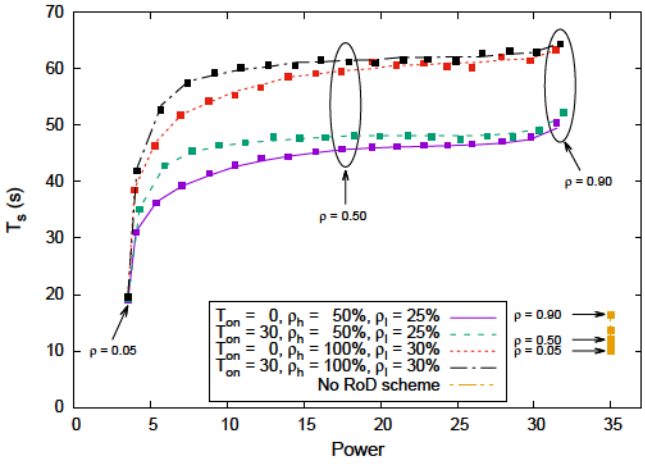


Fig. 8. Average service time vs. average power consumption.

that there are up to $N = 10$ APs available,[9] that a single AP consumes 3.5 W when active (which corresponds to the average power consumed by a Linksys device [19]) and zero otherwise, and that $\mu = 0.1$ s^{-1}.

A. Impact of network load

We first analyse the power consumption as the network load $\rho = \lambda/(N\mu)$ varies. To this end, we fix a target distribution of $M = 5$ users per AP and the following two configurations of the (de)activation thresholds $\{\rho_h, \rho_l\}$: {100%, 30%} and {50%, 25%}, the former being more "reluctant" to increase the number of APs when the network load increases. To understand the impact of T_{on} on performance, we consider the cases of zero and 30 s start-up times. We plot the computed figures of P in Fig. 7, where we use squares for the simulation values (average of 10 simulation runs, each consisting of more than 100k users) and lines for the analysis.

According to the results, the power consumption is monotonously increasing with the network load, with the analysis practically coinciding with the simulation values, with some minor deviations (approx. 1.8%) for high loads (we depict a zoomed version of the figure for these values). Considering the relative performance of each configuration, for the case of $T_{on} = 0$ the results overlap, while for the case of $T_{on} = 30$ s, the policy that is "more eager" to power APs leads to higher power consumption.

We next analyse the performance in terms of service time and the trade-off with power consumption. To this end, we plot T_s vs. P in Fig. 8, with each simulation point corresponding to a different value of ρ, which varies from 0.05 to 0.9 in steps of 0.05. Here we also provide for comparison the "ordinary" case of no RoD scheme (all APs always on), which leads to the smallest service times and the largest power consumptions. As in the previous case, the analysis accurately predicts simulation results, with differences below 2.5%. The figure also illustrates that the service time is a monotonous increasing function of the load: steep for $\rho \leq 0.3$, which is caused by the "drastic" impact of powering on an AP when the number of active resources is relatively low, and then more gradual until $\rho \approx 0.9$. Concerning the impact of the considered configurations, for the same value of the $\{\rho_l, \rho_h\}$ parameters, the non-zero start time has an impact of approx. 5 s for the more dynamic configuration, and approx. 2 s for the more "reluctant" configuration, while the impact of the activation policy results in differences of approx. 12 s.

B. Impact of RoD configuration

Next, we consider the case of a fixed value of $\rho = 0.5$, and compute the service time and power consumed for the two considered $\{\rho_h, \rho_l\}$ configurations and different values of the target number of users per AP M. We plot the service time and the power consumption as a function of M, with the results being depicted in Fig. 9.

For the case of the service time (Fig. 9, top), again simulation results practically coincide with the analysis. The larger M is, the longer the service times are, as users are more likely to share the capacity of a single AP before activating new resources. In fact, the relation is practically linear, e.g., when M changes from 5 to 10, the service time doubles for all considered scenarios: as there are, on average, more users per AP, the service times will be longer.

For the case of the power consumption, the resulting values are depicted in Fig. 9 (bottom). Here, we note that the results for both RoD configurations for $T_{on} = 0$ overlap, and result in a constant power consumption regardless of the value of M. The reason for this behaviour is that, as M increases, more users per AP are required to power on additional resources, but also longer service times will result, leading to more users in the system. In fact, the power consumption of 17.5 W implies that, on average, 5 out of the 10 available APs are on, which matches the $\rho = 0.5$ load. When the start-up times are non-zero, there is a small reduction of P as M increases. The reason for this is that, on average, the system is less likely to power on additional APs, which incurs in the

[9] This is a reasonable number for dense scenarios: for instance, an auditorium with 360 users, each of them demanding 3 Mbps for HD video, would require 31 802.11n APs with a throughput of 35 Mbps (data taken from [18]). Results of the same order of magnitude are obtained in [27], [28].

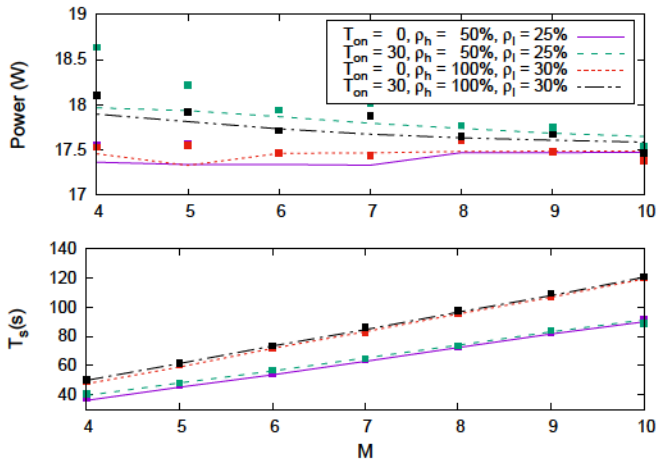


Fig. 9. Average service time (top) and power consumption (bottom) vs. target number of users per AP.

overhead of the start-up process. Finally, we also note that simulation values are very close to those from the analysis, with relative differences of approx. 4% (the smaller M is, the larger the differences are, as the impact of non-perfect overlap of coverage areas is more noticeable for a small number of users).

C. Computational complexity

We next estimate the computational complexity of obtaining the numerical solution for the exact and the simplified analysis. To this end, we assume a scenario with $N = 10$ APs, fix $M = 4$, and consider different configurations of T_{on}, ρ, and $\{\rho_h, \rho_l\}$ parameters. For each set of parameters, we compute the average service time T_s and power consumption P using the exact and the simplified analysis, as well as the time required to compute these values for each case. We note that we use `Octave` to compute the numerical solution for these analysis, running over an Intel® Xeon® X5550 @2.67GHz with 48 GB RAM, and therefore our comparison serves to illustrate the relative differences in complexity, and not absolute values.

We provide in Table I the results of the above computation. More specifically, we provide in the Table, for each considered configuration, the relative difference between the two analyses in terms of service time (denoted as ΔT_s) and power consumption (denoted as ΔP), and the corresponding computation times. There are two main observations from the results: (i) on the one hand, for both power and service time figures, the resulting differences between the numerical analyses are at most 3%, and in many cases well below 1%; and (ii) on the other hand, for the computational times, there are two orders of magnitude of difference between them in all but for two cases. Finally, it is also worth noting that, for the case of the exact analysis, computational times grow with T_{on}, which confirms to some extent that the $K*$ stages are responsible for the computational burden.

TABLE I
RELATIVE DIFFERENCES AND COMPUTATIONAL TIMES OF THE EXACT AND SIMPLIFIED ANALYSES.

T_{on} (s)	ρ_h, ρ_l	ρ	Error		Comp. time (s)	
			ΔT_s	ΔP	Exact	Simpl.
0	{0.5, 0.75}	0.25	≈ 0%	≈ 0%	104.28	2.35
		0.75	≈ 0%	≈ 0%	102.34	2.35
	{1, 0.7}	0.25	≈ 0%	≈ 0%	102.78	2.40
		0.75	0.06%	0.04%	104.58	2.30
15	{0.5, 0.75}	0.25	0.22%	0.14%	167.69	3.62
		0.75	0.91%	0.58%	269.43	5.77
	{1, 0.7}	0.25	0.02%	0.01%	166.94	3.9
		0.75	0.17%	0.14%	276.20	5.65
30	{0.5, 0.75}	0.25	1.97%	1.05%	320.52	6.75
		0.75	3.09%	1.80%	519.96	11.36
	{1, 0.7}	0.25	0.41%	0.23%	316.62	7.03
		0.75	1.17%	0.66%	541.04	11.72

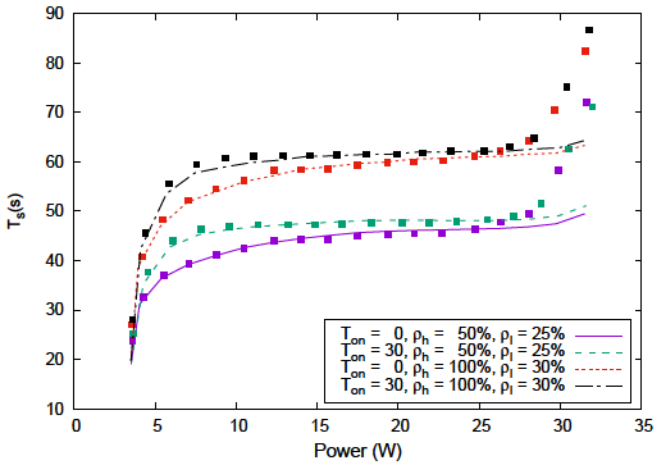


Fig. 10. Average delay vs. power consumed with non-exponential service demands.

D. Realistic traffic model

To analyse the impact of the simplifying assumptions on the traffic model of our analysis, in the following we compare the results obtained with our analysis against those obtained from simulations with a "realistic" traffic model. In particular, we follow [29] and assume that when a station joins the WLAN, it performs a random number of download requests that follows a BiPareto distribution. The length of each download also follows a BiPareto distribution, and the interarrival time of requests follows a lognormal distribution. We fix the average number of requests to 10, with the following parameters of the BiPareto distribution: $\alpha = 0.06$, $\beta = 1.73$, $c = 6.61$ and $k = 1$; the lognormal distribution is simulated with parameters $\mu = 0.34$ and $\sigma = 0.63$; and the request lengths are initially modelled with parameters $\alpha = 0.0, \beta = 2.13, c = 20.0$ and $k = 1.5$ (which leads to an average download size of 30 MB), while the user arrival rate is Poissonian at a rate ranging from 0.05 to 0.9 s^{-1}.

We show in Fig. 10 the resulting average service time vs. power consumption for different configurations of the RoD scheme and values of T_{on}, where (like in Fig. 8) we vary the load from 0.05 to 0.9 in steps of 0.05. We observe that the accuracy of the model worsens as the service rate increases: the deviations are smaller than 5% for $\rho < 0.8$ but notably

TABLE II
PERFORMANCE AN OPTIMAL CONFIGURATIONS OF A ROD SCHEME.

$\mu(s^{-1})$	ρ	$T_{on}(s)$	M	ρ_h	ρ_l	$T_s(s)$	$P(W)$
0.05	0.25	0	3	1.20	0.55	75.93	8.76
		15	4	0.75	0.30	79.17	8.96
		30	3	1.20	0.30	74.77	9.16
	0.5	0	3	1.20	0.30	76.89	17.33
		15	3	1.15	0.30	76.60	17.55
		30	3	1.15	0.30	78.41	17.81
	0.75	0	3	1.20	0.30	77.44	25.35
		15	3	1.20	0.30	79.75	26.00
		30	2	0.95	0.45	53.00	25.34
0.10	0.25	0	3	1.20	0.55	37.96	8.76
		15	3	1.20	0.30	37.38	9.16
		30	3	1.20	0.30	39.98	9.50
	0.5	0	3	1.20	0.30	38.44	17.33
		15	3	1.15	0.30	39.21	17.81
		30	2	0.95	0.45	28.41	17.88
	0.75	0	3	1.20	0.30	38.72	25.35
		15	2	0.95	0.45	26.50	26.34
		30	2	0.95	0.45	28.20	25.98
0.20	0.25	0	3	1.20	0.55	18.98	8.76
		15	3	1.20	0.30	19.99	9.50
		30	3	0.80	0.30	19.87	10.12
	0.5	0	3	1.20	0.30	19.22	17.33
		15	2	0.95	0.45	14.21	17.88
		30	2	1.00	0.45	16.36	17.61
	0.75	0	3	1.20	0.30	19.36	25.35
		15	2	0.95	0.45	14.10	25.98
		30	2	0.95	0.45	15.90	25.42

higher as the system gets closer to saturation. We conclude from these results that overall the accuracy of the model is reasonable for the range of loads of interest (i.e., sufficiently far from congestion).

E. Optimal Configuration of a RoD scheme

While the exact analysis incurs in a notable complexity, we have seen that the simplified analysis is able to compute the performance figures of a RoD scheme in an affordable manner while keeping a notable accuracy. In this way, it can be used, for instance, to compute the optimal configuration of a RoD algorithm, given a set of estimated network conditions, these being expressed in terms of λ and μ. In the following, we present one example of such configuration algorithms, although we restrict ourselves for simplicity to the considered RoD policy (although there could be many others) and a simple optimisation criterion. Our optimisation scheme works as follows. Given an estimation of the network conditions, we set a bound on the maximum service time $T_{\max}$, and perform a sweep on the configuration space $\{M, \rho_h, \rho_l\}$ to look for the configuration that minimises power while guaranteeing an average service time T_s below $T_{\max}$. In our search, M goes from 2 to 10 in steps of one, while ρ_h and ρ_l go from 0.05 to 1.25 in steps of 0.05.

The configuration resulting from this search and the corresponding performance figures are given in Table II for three different service rates $\mu = \{0.05, 0.1, 0.2\}\ s^{-1}$ and the corresponding three service time bounds $T_{\max} = \{80, 40, 20\}\ s$, respectively. If we compare the consumed power with a reference scenario of the 10 APs always on (i.e., consuming 35 W), the reduction is quite considerable, ranging between 25% and 75% depending on the network load. Finally, it is also worth remarking that T_{on} has a non-negligible effect on the resulting configuration parameters.

VI. CONCLUSIONS

Resource-on-Demand schemes are required in dense networks to adapt to the varying load while maintaining an energy efficient performance. In this paper, we have developed an analytical model of these schemes that, in contrast to previous publications, accounts for the non-zero start-up times of real hardware. We have also presented a simplified model, whose computational times are appox. 50x shorter while maintaining relative errors below 3%. We have illustrated the practicality of this simplified model with a simple algorithm to derive the optimal configuration of a RoD scheme.

REFERENCES

[1] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 55–62, August 2011.

[2] Y. Chen, S. Zhang, S. Xu, and G. Li, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, June 2011.

[3] C. Han *et al.*, "Green radio: Radio techniques to enable energy-efficient wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 46–54, June 2011.

[4] G. Lim, C. Xiong, L. Cimini, and G. Li, "Energy-efficient resource allocation for OFDMA-based multi-RAT networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2696–2705, May 2014.

[5] L. Budzisz *et al.*, "Dynamic resource provisioning for energy efficiency in wireless access networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2259–2285, 4th Quart. 2014.

[6] 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.

[7] Y. S. Soh, T. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 840–850, May 2013.

[8] J. Wu, Y. Zhang, M. Zukerman, and E.-N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 803–826, 2nd Quart. 2015.

[9] W. Sun, H. Li, and J. Wu, "Study on real energy consumption of large-scale campus wireless network," in *Proc. 2013 International Conference on Computing, Networking and Communications (ICNC)*, Jan 2013, pp. 605–609.

[10] F. Ganji, L. Budzisz, and A. Wolisz, "Assessment of the power saving potential in dense enterprise WLANs," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sept 2013, pp. 2835–2840.

[11] F. Ganji, L. Budzisz, F. G. Debele, N. Li, M. Meo, M. Ricca, Y. Zhang, and A. Wolisz, "Greening campus WLANs: Energy-relevant usage and mobility patterns," *Computer Networks*, vol. 78, pp. 164–181, 2015.

[12] F. Debele, M. Meo, D. Renga, M. Ricca, and Y. Zhang, "Designing resource-on-demand strategies for dense WLANs," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2494–2509, Dec 2015.

[13] 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 *Proc. ACM International Conference on Future Energy Systems (ACM e-Energy)*, May 2010, pp. 159–168.

[14] 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.

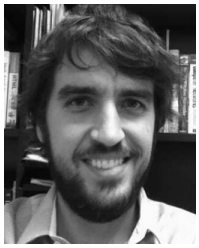
[15] R. G. Garroppo, G. Nencioni, G. Procissi, and L. Tavanti, "The impact of the access point power model on the energy-efficient management of infrastructured wireless LANs," *Computer. Networks*, vol. 94, pp. 99–111, Jan. 2016.

[16] 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.

[17] J. Ortin, P. Serrano, and C. Donato, "Optimal conf guration of a resource-on-demand 802.11 WLAN with non-zero start-up times," *Computer Communications*, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S014036641630161X
[18] J. Florwick, J. Whiteaker, A. C. Amrod, and J. Woodhams, "Wireless LAN design guide for high density client environments in higher education," Cisco, Tech. Rep. [Online]. Available: http://www.cisco.com/c/dam/en_us/solutions/industries/docs/education/cisco_wlan_design_guide.pdf
[19] 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," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1243–125, Aug. 2015.
[20] M. Papadopouli, H. Shen, and M. Spanakis, "Modeling client arrivals at access points in wireless campus-wide networks," in *Proc. 14th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN).*, April 2005, pp. 6–12.
[21] V. Paxson and S. Floyd, "Wide area traff c: the failure of Poisson modeling," *IEEE/ACM Trans. Netw.*, vol. 3, no. 3, pp. 226–244, Jun 1995.
[22] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, "The changing nature of network traff c: Scaling phenomena," *SIGCOMM Comput. Commun. Rev.*, vol. 28, no. 2, pp. 5–29, Apr. 1998.
[23] G. R. Hiertz, D. Denteneer, L. Stibor, Y. Zang, X. P. Costa, and B. Walke, "The IEEE 802.11 Universe," *IEEE Comm. Mag.*, vol. 48, no. 1, pp. 62–70, Jan. 2010.
[24] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains*. Wiley-Interscience, 2006.
[25] O. C. Ibe, *Markov Processes for Stochastic Modeling*. Academic Press, 2009.
[26] D. B. Johnson and D. A. Maltz, *Mobile Computing*. Boston, MA: Springer US, 1996, ch. Dynamic Source Routing in Ad Hoc Wireless Networks, pp. 153–181.
[27] "High density wi-f deployment guide," Cisco, Tech. Rep. [Online]. Available: https://documentation.meraki.com/MR/WiFi_Basics_and_Best_Practices/High_Density_Wi-Fi_Deployment_Guide
[28] "High-density wireless networks for auditoriums, version 3.3," Aruba Networks, Tech. Rep. [Online]. Available: http://community.arubanetworks.com/aruba/attachments/aruba/Aruba-VRDs/21/1/High-Density%20Wireless%20Networks%20for%20Auditoriums.pdf
[29] F. Hernández-Campos, M. Karaliopoulos, M. Papadopouli, and H. Shen, "Spatio-temporal modeling of traff c workload in a campus WLAN," in *Proc. 2nd Annual International Workshop on Wireless Internet (WICON '06)*, Boston, MA, USA, Aug. 2006.

Jorge Ortín received his Telecommunication Engineering degree and his PhD from the Universidad de Zaragoza in 2005 and 2011 respectively. He works at Centro Universitario de la Defensa Zaragoza since 2013, where he currently holds the position of Associate Professor. He was a postdoc research fellow at the Universidad Carlos III of Madrid in 2012 and a visiting researcher at the Politecnico de Milano in 2016. Research interests include wireless communications systems, specif cally the analysis of resource allocation problems for cellular, WLAN and sensor networks.

Carlos Donato received his B.Sc. in telecommunication engineering from Universidad Carlos III de Madrid (UC3M) in 2014, and his M.Sc. in telematics engineering from the same university in 2015. He is currently pursuing his PhD while working at Institute IMDEA Networks. His research focuses on performance evaluation and centralised optimisation of wireless networks.



Pablo Serrano (M'04 – SM'15) got his Telecommunication Engineering degree and his PhD from the Universidad Carlos III de Madrid (UC3M) in 2002 and 2006, respectively. He has been with the Telematics Department of UC3M since 2002, where he currently holds the position of Associate Professor. He was a visiting researcher at the Computer Network Research Group at Univ. of Massachusetts Amherst in 2007 under a Jos Castillejo grant, at Telefonica Research Center in Barcelona in 2013, at the School of Computer Science and Statistics at Trinity College Dublin in 2015, and at Universita degli Studi di Brescia in 2016. He has over 70 scientif c papers in peer-reviewed international journal and conferences. He serves on the Editorial Board of IEEE Communications Letters.

Albert Banchs (M'04 – SM'12) received the M.Sc. and Ph.D. degrees from the Polytechnic University of Catalonia (UPC-BarcelonaTech) in 1997 and 2002, respectively. He is currently a Full Professor with the University Carlos III of Madrid (UC3M), and has a double aff liation as Deputy Director of the IMDEA Networks institute. Before joining UC3M, he was at ICSI Berkeley in 1997, at Telefonica I+D in 1998, and at NEC Europe Ltd. from 1998 to 2003. Prof. Banchs is Editor of IEEE Transactions on Wireless Communications and IEEE/ACM Transactions on Networking. His research interests include the performance evaluation and algorithm design in wireless and wired networks.