Computer Networks 000 (2016) 1–12

Contents lists available at ScienceDirect

Computer Networks

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

Minimizing the impact of the handover for mobile users in WLAN: A study on performance optimization

E. Zola[a,*], A.J. Kassler[b]

[a] Universitat Politècnica de Catalunya (UPC), C. Jordi Girona 1-3, Barcelona, Spain
[b] Karlstad University (KAU), Universitetsgatan 2, Karlstad, Sweden

ARTICLE INFO

Article history:
Received 27 August 2015
Revised 16 March 2016
Accepted 23 March 2016
Available online xxx

Keywords:
IEEE802.11
MILP
Handover
Multi-objective optimization

ABSTRACT

IEEE 802.11 based Wireless LANs are an important piece in today's communication infrastructure in order to provide high speed wireless Internet access to static or quasi mobile users. For large WLAN deployments (i.e., Campus or enterprise WLAN), it is important to understand the impact of user mobility and handovers on the system performance. In this article, we have developed a performance model for a set of networked 802.11 based WLAN Access Points, which is based on a Mixed Integer Linear Program (MILP). The objective function tries simultaneously to maximize the total system rate while at the same time minimizing the number of handovers for a configurable handover signaling rate. Because of the conflicting nature of the two objective functions, such multi-objective optimization is difficult to explore. A detailed evaluation of the model using several scenarios involving both different numbers of static and mobile users shows that our formulation allows trading off those two objectives in a robust way.

1. Introduction

Recently, the deployment of 802.11 Access Points (APs) increased in public areas such as campuses, malls, airports, etc., in order to provide better coverage and anytime anywhere Internet access to moving users. When there are multiple APs in reach, it becomes difficult to determine which is the best AP to connect to. From the user perspective, the best association strategy would be the one that guarantees the highest throughput. However, in conventional vendor implementations, a client selects the AP with the highest received signal strength indicator (RSSI) to associate with. Such association strategy may cause unbalanced distribution of the users among neighboring APs and unfair distribution of the rates among the users, thus leading to non-optimal network performance. This is also because the achievable rate depends not only on the perceived RSSI but also on the load that other users impose on the given AP [1]. Also, this association strategy may lead to the so-called *ping-pong* effect, where a user continuously changes its association among neighbouring APs whose signal strength fluctuates over time. This behaviour has been observed by many authors in real wireless local area networks (WLANs) [2–4] and may have a negative impact on the quality perceived by the user.

According to the IEEE 802.11 standard, a station (STA) willing to use the service through a new AP must authenticate and associate with the new AP; this process is time consuming. Several works deal with the long latency during the handover (HO) process in WLAN [5–7], which may increase the probability of service interruption as: 1) every HO may lead to TCP timeouts, and 2) stock implementations of IEEE 802.11 and Internet protocols typically require several seconds to (re-)establish a connection to the AP. This may severely hamper performance, especially when TCP starts to backoff. However, the authentication and association process in WLAN also consumes bandwidth (BW), as signaling messages has to be exchanged between the user and the AP. For this, although the latency during the HO process has a great and clear impact on the QoS perceived in a WLAN, due to the shared nature of the 802.11 channel we argue that also the BW consumed during the HO process should be taken into account in 802.11 network performance modeling.

While the cost of the necessary signaling is taken into account when considering a mobile network [8] or for the vertical HO problem [9], to the best of our knowledge none has addressed the cost of signaling in a WLAN so far. Due to their non-coordinated deployment, the APs in a WLAN typically do not provide seamless HO. In addition, the background load of the users is typically not controlled. Thus, deciding about a good time when to HO is much more difficult. In [10] we introduced the HO cost in terms of the BW consumed by a HO user and presented a multi-objective optimization function for determining the association pattern of

* Corresponding author.
E-mail addresses: enrica@entel.upc.edu (E. Zola), andreas.kassler@kau.se (A.J. Kassler).

http://dx.doi.org/10.1016/j.comnet.2016.03.014
1389-1286/

moving users. The model introduced a second objective in the optimization function proposed in [11] aimed at minimizing the number of HOs performed in the system while keeping the total system download rate high, hence maximizing the overall network performance. This second objective was motivated after observing a high number of HOs even in static scenarios (i.e., static users and no new arrivals in the network), which are typically ascribed to the *ping-pong* effect [4,12]. However, a deep insight on the implications of the new objective function was not provided. This paper fills this gap by thoroughly analyzing the tradeoff between the two conflicting objectives - minimizing the number of HOs while, at the same time, maximizing throughput and fairness. Minimizing the number of HOs is an important aspect of system design in order to come to a stable operation of the deployed WLAN infrastructure, because every HO involves the potential of failure and dropped connections. Moreover, frequent HOs may lead to severe battery consumption, thus unnecessarily reducing the lifetime of the device. The main contribution of this paper is to develop a mathematical model that minimizes the number of HOs in the network while maintaining the aggregate download rate of all the STAs high. We develop a MILP and use the weighted sum method to understand the impact of higher weights on one of the two objective functions. Using different scenarios varying the number of users and mobility parameters such as the user speed, we evaluate our method and conclude that this model allows to effectively tradeoff the number of HOs for reasonable penalty in achievable throughput.

The remainder of the paper is organized as follows. Section 2 provides a brief overview of the related literature. The network model for a IEEE 802.11 WLAN is presented in Section 3. The parameters used in the evaluation are described in Section 4. The tradeoff between the two objectives is studied in Section 5 by varying the weights in the optimization function and considering different scenarios. The impact of the HO parameters is shown in Section 6. Finally, Section 7 concludes the paper and brings some ideas for future work.

2. Related work

The problem of the optimal distribution of the users among the APs so to maximize the overall performance of the WLAN has been extensively studied in the literature. Authors in [13] present a centralized optimal association policy for IEEE 802.11 WLANs. Their optimal strategy may better use the WLAN resources and improve the QoS. Baid et al., [14] study the impact of the interference among multiple overlapping WLANs on the intra-network association optimization problem and propose a cooperative optimization scheme to mitigate the interference. Their problem is formulated as a non-linear program, which can be solved faster through an approximation algorithm they have proposed. Li et al., [15] formulate a non-linear program that takes into account proportional fairness and maximizes the total user bandwidth utilities in the whole network. Several studies can be found in the literature that tackle the problem to find a good association strategy that guarantees load balancing and fairness. For example, authors in [16] proved that the objective of fairness and load balancing in the AP selection can be achieved simultaneously if one of them is achieved. Based on this, they proposed a max-min fair BW allocation problem. [17] proposed a max-min throughput AP association scheme based on the estimation of the load of the APs. However, [18] proved that greedy selection of the least-loaded AP does not guarantee optimum AP association.

Common to all these previous works, however, is the fact that they do not take into account the cost of a reconfiguration of the network (i.e., when a user has to associate to a new AP). Recently, Dely et al., [11] proposed a MILP that optimizes the associations among different APs in a IEEE 802.11 WLAN by maximizing the download rate of each STA while still trying to provide fairness among the STAs. In their dynamic model, the authors take into account the cost of reconfiguration for the network, imposed by e.g. HOs and routing updates. However, their cost function for the HO considers only the communication interruption, disregarding the signaling that the STA has to exchange in order to perform a HO. Our proposal goes a step further by also considering the impact of the signaling on the optimal distribution of the users in the WLAN, while still guaranteeing an optimal max-min fair allocation of the rates among the STAs.

Other works can be found in the literature that propose an optimization problem aimed at minimizing the number of HOs in the network. In 2016, Sathya et al., [19] proposed a MILP model for femto-cell deployment in enterprise buildings aimed at guaranteeing a certain minimum SINR and also at minimizing the number of femto-cells needed for minimal coverage constraints. By exploiting the capability of the long term evolution (LTE) system, where the UE regularly reports to its serving cell the RSSI measurements obtained from the neighbouring cells, they integrate their model with the HO constraint in order to minimize the number of HOs. When the HO constraint is also used, 30% of the unnecessary HOs are reduced (i.e., the *ping-pong* effect in corridors is minimized). The problem of the challenges in HO operations has been also investigated in [20] for a LTE cellular system where femto-cells may overlap with irregular patterns. The authors propose a multi-objective HO solution that considers multiple parameters in the selection of the optimal target cell (e.g., RSSI, available BW). Considerable improvement is achieved in the blocking rates, queuing delays, HO latency and throughput during the HO. Finally, the resource allocation problem of users association has been recently addressed in future millimeter-wave access networks. For example, the authors in [21] propose a MILP aimed at minimizing the maximum AP utilization in the network and at ensuring a fair load distribution. Due to the lower stability of the millimeter-wave channel, the problem the authors address is more challenging as several events can violate the efficient operation of the network (e.g., line-of-sight may be compromised by moving obstacles).

3. Network model

The main goal of this paper is to derive a mathematical model of a WLAN network that allows estimation of the performance (download rate) of WLAN users while at the same time modeling the HO process from one AP to another. According to the IEEE 802.11 standard, when a STA wants to connect to a given AP, it has to discover surrounding APs through either passive or active scanning, select one AP (e.g., the one with the highest RSSI), authenticate and then associate with it. This process is time consuming [5]. However, on the system perspective, this process is also resource consuming, since it prevents other users which are already associated with the given AP from downloading data while the AP is exchanging those messages to the HO user. A model to derive the optimal association strategy for STAs moving in such a multi-AP scenario was developed in [10]. The model introduced some modifications on the dynamic network model presented in [11]. In this paper, we build upon those models and analyze the impact of the weighting factor between the two objectives in the optimization function. Therefore, we define our objective function that tries to balance two important but conflicting criteria. We want to simultaneously 1) maximize the aggregate download rate of all the STAs over the whole runtime; 2) minimize the number of HOs. Minimizing the number of HOs can be seen as an important aspect to model the network stability as every HO may fail, which at the end may penalize achievable download rates.

Please cite this article as: E. Zola, A.J. Kassler, Minimizing the impact of the handover for mobile users in WLAN: A study on performance optimization, Computer Networks (2016), http://dx.doi.org/10.1016/j.comnet.2016.03.014

Table 1
Model notation along with evaluation parameters.

Symbol	Description	Type	Value
A	Set of APs a	P	{1, 2, .., *NumAP*}
α	Min user download rate	V	
b_a	Wired link capacity	P	100 Mbps[a]
$c_{as}(t)$	Connected state	V	{0, 1}
$\hat{c}_{as}(t)$	Connecting state	V	{0, 1}
c_{HO}	HO cost	P	$\frac{219.24}{D}$ kbps[b]
$D_a^{(s)}$	Num. slots for HO	P	{1, 2, 4, 8}
η	MAC efficiency	P	1[a]
$I(t)$	Set of interf. links	P	From topology
κ	Fairness parameter	P	10^{-8} [a]
λ	Weighting factor	P	{0, .., 1}
M	Big number	P	10^{+8} [a]
NumAP	Max num of APs	P	13 [a]
NumS	Max num of STAs	P	{20, 40}
$p_{as}(t)$	PHY rate	P	{0, 6, .., 54} Mbps[a]
$r_{as}(t)$	Download rate	V	
S	Set of STAs s	P	{1, 2, .., *NumS*}
$sign_{as}(t)$	Signaling state	V	{0, 1}
$sp_{as}(t)$	Sign. PHY rate	P	6 Mbps
sr	Sign. download rate	P	{0, 10, 1000} kbps
T	Set of timeslots t	P	{1, 2, .., 120}*
t_{slot}	Timeslot duration	P	1 second[a]
$u_s(t)$	User activity	P	Generated randomly[a]

[a] Values are set as in [11].
[b] From Eq. (11).

In this section we specify the parameters, variables, constraints and objective function of the model. The notation used for the model is summarized in Table 1, along with some of the values used for the evaluation. As the model presented here is built upon the models in [10,11], many values are set as in the previous studies for an easier comparison (i.e., values marked with an asterisk).

3.1. Parameters and variables

We define a network model for a IEEE 802.11 WLAN with a set A of APs and a set S of STAs. Each STA s aims at downloading data for a given period of time inside the set of timeslots T. The user request for a STA s in a given timeslot t is described by a binary variable $u_s(t)$. The system state is described by these service requests, link capacities (b_a) and link interference conditions. The MILP model works under the assumption that the future network state is known.[1] When a STA s is requesting for download in timeslot t (i.e., $u_s(t) = 1$), first it has to associate with an AP. Three states are defined: "*connecting*", "*connected*" and "*signaling*". They are modelled through three binary variables: $\hat{c}_{as}(t)$, $c_{as}(t)$, and $sign_{as}(t)$, respectively. $D_a^{(s)}$ represents the HO cost in terms of the time needed for a STA s to associate with AP a. If a STA s is in the *connecting* state but not in the *connected* state, then s is in the *signaling* state (see Eq. 7), meaning that it is exchanging management frames with the AP a to which it is associating.

In [10] we introduced a new parameter, the signaling download rate sr. In cellular networks it is common to consider the BW consumed during signaling (new users arriving at a base station or HO users); however, to the best of our knowledge, this factor is not taken into account when modelling the performance of IEEE802.11 WLANs. Besides, a signaling STA is also competing for accessing the shared medium with the other STAs, thus preventing them from download while its association process is taking place. For this reason we think it is important to take this parameter into account in our model. This parameter can be tuned according to the characteristics of the network and scenario under study.

3.2. Model constraints

In our model, we will consider that a STA s can start connecting to an AP a only when this STA is requesting service (service request constraint in Eq. 1, see Table 2). According to IEEE 802.11, a STA s can only download data from AP a when it is in the *connected* state with a (i.e., $c_{as}(t) = 1$). This is expressed through the download constraint in Eq. 2, where the download rate $r_{as}(t)$ between AP a and STA s is bigger than zero only when s is connected to a. The large number M makes sure that the download rate for a connected station is not bounded by this constraint. A STA s can only enter the *connected* state if and only if it has been in the *connecting* state for a given amount of timeslots $D_a^{(s)}$ (Eq. 3). More information on how to model the relationship among the connecting and connected states and the parameter $D_a^{(s)}$ can be found in [10].

Eq. 4 represents the capacity constraint of the wireless channel; that is, the normalized data and signaling rates ($p_{as}(t)$ and $sp_{as}(t)$, respectively) of a link and of all the interfering links $I(t)$ cannot exceed the efficiency of the MAC layer protocol η. As defined in [11], η cannot exceed one. We also need to make sure that the capacity b_a of the wired link (i.e., the Internet link to which each AP is connected) must not be exceeded (Eq. 5). Eq. 6 guarantees that a STA can only be connecting, connected and signaling to at maximum one AP in each timeslot, as required in IEEE 802.11. Finally, Eq. 7 sets the relationship among the three states; a STA not connected but in connecting state with a given AP a is also signaling with a, while a STA connected with a is also in connecting state, but not signaling.

3.3. Objective

The objective function (Eq. 8) has been defined to find a balance between: 1) the maximum aggregate download rate of all the STAs over the whole runtime (i.e., total download volume, hereafter); and 2) the minimum number of associations in the system (thus, the minimum number of HOs). λ is the weighting factor between the two objectives. κ is a fairness parameter that ensures a fair allocation of the rates among the users [11]. When κ is set to 0, the minimum download rate is maximized. However, according to Eq. 8, it might occur that some download rates are not maximized beyond α, even if they could be increased without decreasing α. By increasing κ, more focus is put on overall network performance and less on fairness; hence, α might be lower (α is the minimum download rate that each user must receive, as from Eq. 9). In the rest of the paper we set $\kappa = 10^{-8}$, as in [11], to enforce a high level of fairness and to make sure that download rates are maximized beyond α.

Furthermore, for each STA s, Eq. 10 defines the total signaling cost $sis(s)$ (i.e., the opportunity cost of all the associations that STA s performs in the network with all the APs). $c_{HO}(a)$ expresses the cost, in terms of bps, for each HO performed in the network.[2] According to Fig. 1, the time needed to exchange the authentication and the association frames can be roughly estimated to be 4.06 ms.[3] If no association was taking place in the network, this time

[1] An extension of this model to cope with unknown future states has been provided in [11] but without taking into account HO related signaling and the tradeoffs that this paper considers.

[2] For simplicity, we also include here the signaling cost when a STA first associates to the network; however, the first association is not counted as a HO in the evaluation section.

[3] Values here refer to 802.11g standard. However, the optimization model is independent on the standard used and can be easily applied to more recent and/or future versions of the standard by tuning some parameters.

Please cite this article as: E. Zola, A.J. Kassler, Minimizing the impact of the handover for mobile users in WLAN: A study on performance optimization, Computer Networks (2016), http://dx.doi.org/10.1016/j.comnet.2016.03.014

Table 2
MILP model.

Constraint	Expression	Support	
Service req.	$\hat{c}_{as}(t) \le u_s(t)$	$\forall a \in A, s \in S, t \in T$	(1)
Download	$r_{as}(t) \le M \cdot c_{as}(t)$	$\forall a \in A, s \in S, t \in T$	(2)
Connecting	$\sum_{t'=t_1}^{t_1+D_a^{(s)}} \hat{c}_{as}(t') \ge D_a^{(s)} + 1$	$\forall a \in A, s \in S, t' \in T$	(3)
Wireless	$\sum_{a',s':\{(a,s),(a',s')\}\in I(t)} \left(\frac{r_{a's'}(t)}{p_{a's'}(t)} + \frac{sr \cdot sign_{a's'}(t)}{sp_{a's'}(t)} \right) + \frac{r_{as}(t)}{p_{as}(t)} +$	$\forall a \in A, s \in S, t \in T$	(4)
Capacity	$+ \frac{sr \cdot sign_{as}(t)}{sp_{as}(t)} \le \eta$		
Wired cap.	$\Sigma_{s \in S} r_{as}(t) \le b_a$	$\forall a \in A, t \in T$	(5)
Single conn.	$\begin{cases} \sum_{a\in A} \hat{c}_{as}(t) \le 1 \\ \sum_{a\in A} c_{as}(t) \le 1 \\ \sum_{a\in A} sign_{as}(t) \le 1 \end{cases}$	$\forall a \in A, s \in S, t \in T$	(6)
Signaling	$sign_{as}(t) = \hat{c}_{as}(t) - c_{as}(t)$	$\forall a \in A, s \in S, t \in T$	(7)
Object. Func.	**max** $(1-\lambda) \cdot (\alpha + \kappa \sum_{s\in S} q(s)) - \lambda(\sum_{s\in S} sis(s))$		(8)
	$q(s) = \frac{\sum_{t\in T} r_{as}(t)}{\sum_{t\in T} u_s(t)}, \quad -q(s) + \alpha \le 0$	$\forall a \in A, s \in S, t \in T$	(9)
Signal. cost	$sis(s) = \sum_{a\in A} \sum_{t\in T} c_{HO}(a) \cdot sign_{as}(t)$	$\forall a \in A, s \in S, t \in T$	(10)
Cost of one HO	$c_{HO}(a) = \frac{4.06\ ms}{D_a^{(s)} \cdot t_{slot}} \cdot 54$ Mbps,	$\forall a \in A$	(11)

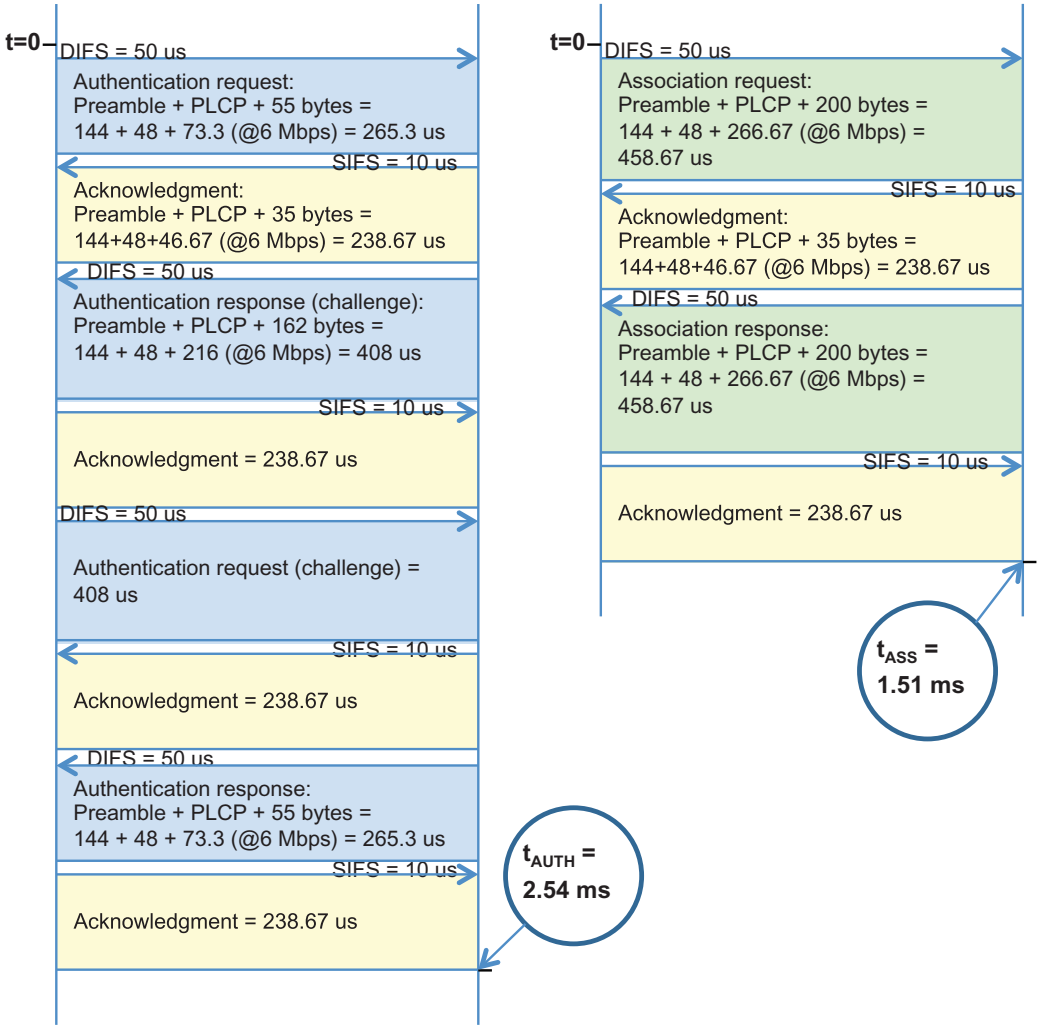


Fig. 1. Management frames exchanged between a STA and the AP during association in IEEE 802.11g WLAN.

could have been used by a STA for downloading data. This STA would transmit at a maximum of 54 Mbps (i.e., best channel conditions). Thus, the opportunity cost of a HO, in the worst-case scenario, can be expressed as in Eq. 11; please note that, whether on the one hand, the maximum transmission rate (54 Mbps) is considered, on the other hand the possibility of authentication and/or association failure is not taken into account in this paper as the authentication and association time (i.e., 4.06 ms) is only counted once. A thorough analysis of the value of the opportunity cost is out of the scope of this paper.

3.4. Small example

In order to make the model easier to understand, we provide a small example. Assume that 10 STAs (*NumS*=10) are located in the overlapping area of two APs (*NumAP*=2) and do not move

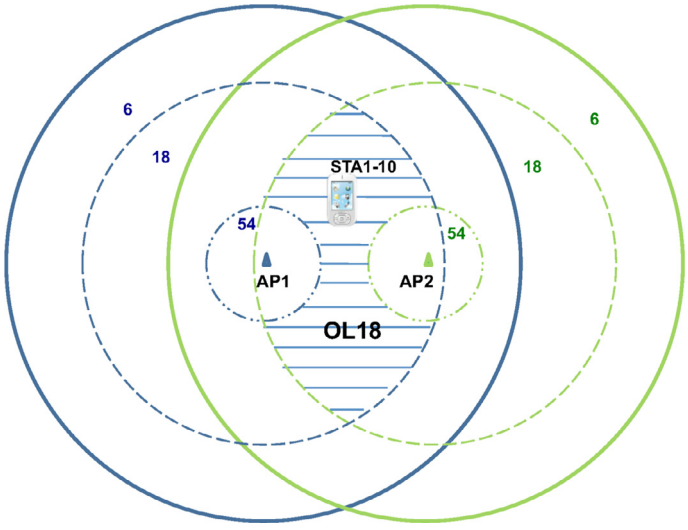


Fig. 2. Layout of the small example. Ten STAs are placed in the overlapping area of two APs, getting 18 Mbps from each.

Table 3
Optimal association pattern for the small example with ten STAs and two APs. *Connecting* (X) or *connected* (AP number) state for each STA and time slot are given. $D_a^{(s)}$ is set to 2, and sr=0.

Connecting or connection states in each slot t			
STA	Proposed model (sr=0)	STA	Proposed model (sr=0)
1	{X1,X1, 1, 1, 1, 1, 1,**X2,X2**,2}	6	{X1,X2,X2, 2, 2, 2, 2, 2, 2, 2}
2	{X1,X1, 1, 1,**X2,X2**, 2, 2, 2, 2}	7	{X1,X2,X2, 2, 2, 2, 2, 2, 2, 2}
3	{X1,X2,X2, 2, 2, 2, 2, 2, 2, 2}	8	{X1,X2,X2, 2, 2, 2, 2, 2, 2, 2}
4	{X1,X1, 1, 1, 1, 1, 1, 1, 1, 1}	9	{X1,X1, 1, 1, 1, 1, 1, 1, 1, 1}
5	{X1,X1, 1,**X2,X2**, 2, 2, 2, 2, 2}	10	{X1,X1, 1, 1, 1, 1, 1, 1, 1, 1}
Num HO: 3 $\sum_{a,t} r_{as}(t)$: 27 Mbps ($\forall s \in S$)			

during the simulation. Based on their positions and the signal strength they experience from each AP, all the STAs get the same bit rate from the two APs (i.e., $p_{as}(t)$= 18 Mbps); we call this area OL18 in Fig. 2, where for simplicity only one device is represented. Assume that all the STAs are active since the beginning and request to download in each time slot (i.e., $u_s(t)$=1 $\forall t \in T$). We will consider T=10 in this small example (i.e., the system is optimized over 10 s).

According to the standard driver, at the beginning (i.e., timeslot t=1), the AP with the best RSSI (thus the highest bit rate) will be selected; since both APs provide the same bit rate in this example, the one with the lowest ID (i.e., AP1) is assumed to be selected. According to the service request constraint in Eq. 1, all the STAs may start connecting with AP1 in the first timeslot (i.e., $\hat{c}_{AP1,s}(t=1) \leq 1\ \forall s \in S$). The number of time slots needed to connect to a given AP a is set to 2 in this example for all the STAs ($D_a^{(s)}$=2). According to Eq. 3, in the first timeslot, none of the STAs can be connected to any AP (i.e., $c_{a,s}(t=1)$= 0 $\forall s \in S$ and $\forall a \in A$); according to the download constraint in Eq. 2, no STA can download data (i.e., $r_{as}(t=1)$=0 $\forall s \in S$ and $\forall a \in A$). Also, according to the signaling constraint in Eq. 7, all the STAs will be in signaling state with AP1 (i.e., $sign_{AP1,s}(t=1)$=1, $\forall s \in S$). Table 3 provides the connecting (X followed by the AP number) and connected states (AP number) for all the STAs and timeslots, after the MILP is solved over 10 s (i.e., $T = \{1, 2, ..., 10\}$).

STAs 1, 2, 4, 5, 9 and 10, are in connecting and signaling states with AP1 during the first two timeslots, and will be connected to AP1 in t=3. As all the STAs start connecting to AP1 in t=1, a HO to a neighbor AP may be forced in order to equally share the BW among all the users, as ensured through the fairness parameter κ. In this example, STAs 3, 6, 7 and 8 start connecting to AP2 in t=2; at t=4, they are connected to AP2 and can start downloading data. At some point, the algorithm may force a STA to perform a HO to the other AP in order to balance the load between the two APs and to guarantee a fair download rate among all the STAs. For example, at t=8, STA 1 will start connecting to AP2 (i.e., $\hat{c}_{AP2,STA1}(t=8,9)$= 1 and $sign_{AP2,STA1}(t=8,9)$=1) and will start downloading data at t=10 from AP2. Bold letters represent a HO in Table 3; as $D_a^{(s)}$=2, then two timeslots are always needed to perform an association to a given AP.

This example was already shown in [10], where we demonstrated that the proposed model avoids the ping-pong effect and reduces the number of HOs, while maintaining the same average download rate obtained with the model in [11].

3.5. Practical implementation and complexity

Regarding practical implementation issues, the system could be implemented using software defined networking (SDN). Download volume may be provided to the SDN controller who may have information about load from all users and link quality estimates to each AP. The latter is already implemented by CloudMAC [22] and it may be extended by an optimization module that runs our problem.

In order to illustrate the complexity of our problem, Table 4 shows the average execution time and its standard deviation for different case studies run in different machines. Scenario 1 refers to the static scenario with 20 users, while scenario 2 to the one where 20 users move at 1 m/s; in both scenarios, $D_a^{(s)}$ is set to 4 and sr to 0. As can be seen, the execution time increases when the number of users increases. Note, that execution time can be significantly reduced if CPLEX precision (*epgap* parameter) for optimality is reduced; however, this may have a negative impact on the results as the true optimum may not be found by the solver. As can be seen, the runtime to solve for large instances is prohibitive. As a consequence, solving the given model with e.g., CPLEX is not suited for online optimization. Rather, it is seen as the benchmark against which any fast heuristic can be compared with. However, the design of such heuristics is outside the scope of this paper.

4. Evaluation settings

This section provides the details on the parameters used for evaluation in the following sections. Provided we know the physical rates, the collision domains and the user activity for the whole run-time, we can solve the MILP problem with MILP solvers such as CPLEX [23] and compute the optimal download rates $r_{as}(t)$ for each STA in each slot in the whole system. Moreover, the HO pattern is also obtained, which maximizes the total download volume regardless of the extra amount of signaling that must be exchanged.

A set of custom-made simulation scripts have been used for the evaluation. These scripts are the same that were used in [11] and contain:

1. the maximum number of STAs *NumS*;
2. the user activity (i.e., when each user wants to download data and for how long). In our setup, each user randomly generates one traffic request within the first 30 s and aims to download for at least 50 s, as in [11];
3. the PHY rates $p_{as}(t)$ that each STA achieves from each AP and at each timeslot;
4. the set of interfering links $I(t)$;
5. the number of slots for HO $D_a^{(s)}$;

Table 4
Average execution times (and standard deviations) for different case studies and machines. Results for each scenario are also given.

Case study	Precision	λ	Machine	EXE Time [s]		Results	
				Mean	Std.	Num HO	Tot down vol. [Mbits]
Scenario 1	10^{-30}	10^{-50}	PC1[a]	8.51	16.64	19.10	31,663
		10^{-10}		10.06	20.96	18.17	31,659
		10^{-6}		13.06	23.28	2.73	31,659
		10^{-4}		11.06	22.34	2.5	31,659
		10^{-2}		10.35	20.81	2.2	31,659
		10^{-1}		11.06	23.48	2.2	31,659
Scenario 1	10^{-1}	10^{-50}	PC2[b]	1.83	1.85	23.3	30,886
		10^{-10}		1.89	1.85	25.6	30,527
		10^{-6}		1.37	1.39	7.6	25,719
		10^{-4}		1.88	1.91	4.4	25,900
		10^{-2}		1.86	2.16	2.7	25,889
		10^{-1}		2.05	2.40	2.63	26,150
Scenario 2	10^{-30}	10^{-50}	PC2[b]	54.55	82.09	53.58	43,617
		10^{-10}		47.87	67.44	53.43	43,617
		10^{-6}		186.04	479.49	11.07	43,617
		10^{-4}		162.75	270.97	7.31	43,617
		10^{-2}		125.83	207.60	7.03	43,617
Scenario 2	10^{-1}	10^{-50}	PC2[b]	9.44	8.28	73.4	40,971
		10^{-10}		9.26	7.24	71.4	41,179
		10^{-6}		27.94	113.16	9.93	31,909
		10^{-4}		9.18	8.86	11.5	31,648
		10^{-2}		35.19	107.31	7.57	32,049

[a] PC1 - Windows XP, Intel®CoreTM2 Duo CPU, T9600, 2.80 GHz, 4 GB RAM, 64-Bit OS.
[b] PC2 - Windows 7, Intel®CoreTM2 Quad CPU, Q9400, 2.66GHz, 4 GB RAM, 64-Bit OS.

6. the signaling download rate *sr*. This parameter is new with respect to [11].

Each simulation was repeated thirty times with different random STA positions, mobility patterns and download schedule, in order to obtain an average estimation of the system performance. The mobility traces were created by randomly placing STAs in the Computer Science Department of Karlstad University, as detailed in [11]. A total of 13 APs were positioned according to the real deployment and assumed to have Fast Ethernet connections to the Internet (i.e., $b_a = 100$ Mbps).

5. Exploring the tradeoff between the two objectives

Our problem is a multi-objective MILP, where two conflicting objective functions should be optimized simultaneously – number of HOs and achievable rate. The answer to such a problem is typically not a single point but a set of points for which locally there exists no other feasible solution that would improve some objective without causing a decrease in at least another objective. In this paper we address such a multi-objective problem by assigning weights to the individual objectives, turning the problem into a single objective one. The major limitation of such an approach is that a local solver will find just a single point in the set of points. Therefore, we will vary the weight λ in order to explore the tradeoff between the two objectives, which is unfortunately computationally expensive.

This section focuses on the impact of λ on the total download volume and on the number of HOs. To this end, the precision in CPLEX is set so to make sure that the solver finds the optimum. We first analyze the behavior in the case of 20 static users (scenario 1). In static scenarios, HOs are only due to new arrivals (i.e., new users that want to download data). As the aim of this section is to show how the number of HOs decreases without having a negative impact on the total system download rate, the signaling download rate *sr* is set to zero here. Also, $D_a^{(s)}$ is set to four for all the APs and STAs (i.e., $D_a^{(s)} = D$ from now on) in this section. These two latter parameters will be varied in Section 6 which focuses more on the impact of HO parameters on the system performance. Table 1 provides the numerical values used in the evaluation.

Results obtained in the static scenario are then compared to the case of mobile STAs (i.e., 20 STAs move at 1 m/s - scenario 2) in Section 5.2 and to the case of 40 static users (scenario 3) in Section 5.3.

5.1. Scenario 1 - 20 static users

In this section, we evaluate the impact of the weighting factor on download volume and number of HOs for the static user case. The plot on the left side of Fig. 3 is a scatter plot of the total download volume versus the total number of HOs for each of the 30 repetitions in the case of 20 static users. Different colors and markers represent the results for different values of λ. The trend given by the average values is also shown with a black dash-dot line. For example, when $\lambda = 0$ (i.e., black asterisks) the number of HOs varies between 0 and 90 due to the random initial position of the users; the total download volume varies between 22.61 and 40.54 Gbits. With λ set to 0.01 (i.e., red stars) the number of HOs varies between 0 and 7, while the total download volume varies between 22.61 and 40.54 Gbits as before.

Average and 95% confidence intervals (CI) are depicted in Fig. 3 (right plot) as error bars for each λ, for both the number of HOs and the download volume (i.e., solid line). The marker for the average values for each λ are filled with the corresponding color. In order to help the reader to follow the trend, a black dash-dot line connects those average values. As can be seen from the figure, the average total download volume remains constant for any $\lambda \preceq 0.990$ (i.e., it is 31.66 Gbits). For $\lambda \succ 0.990$ we can see that the total download volume decreases, as the average is below the 95% CI shown for $\lambda \preceq 0.990$. We can thus conclude that the total download volume is insensible to the weight λ within the optimization function, unless it is bigger than 0.990. On the other hand, the average number of HOs is 21.93 for $\lambda = 0$. As soon as $\lambda \succ 1e^{-10}$, the average number of HOs drops drastically (i.e., 2.73 HOs with $\lambda = 9e^{-10}$). The high 95% CIs (i.e., around 100% in the number of HO, and 14.72% of the average download rate) are due

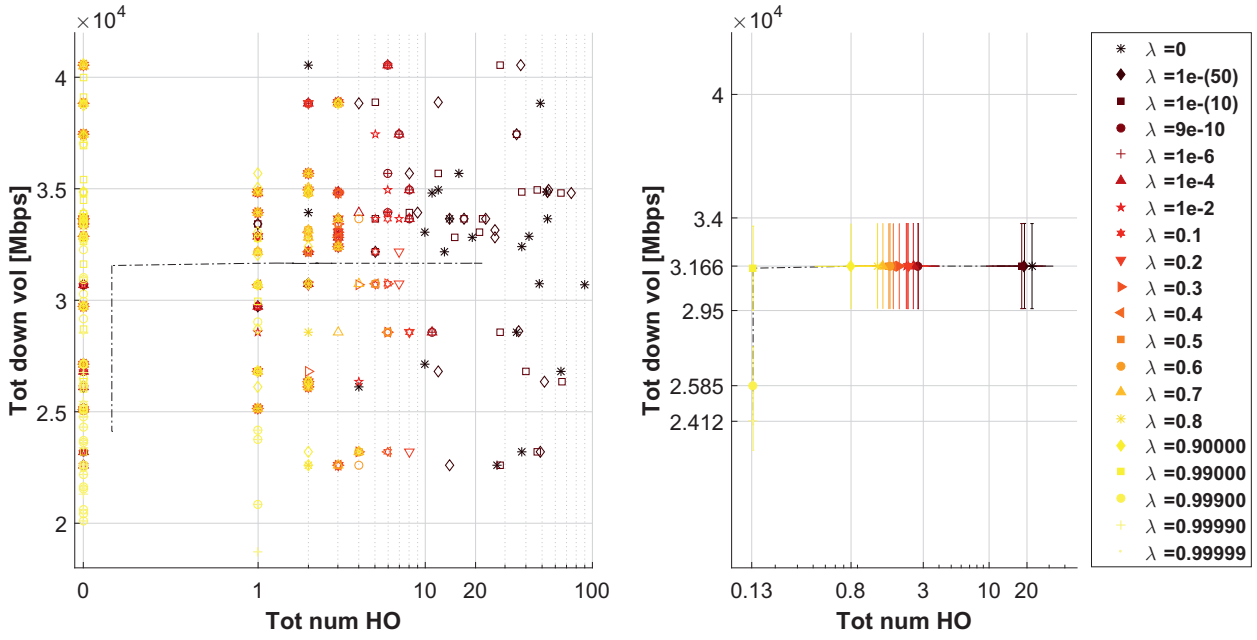


Fig. 3. Total download volume vs total number of HOs for different λ in scenario 1. A black dash-dot line connects average values. Error bars denote 95% CI. X axis is in logarithmic scale. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

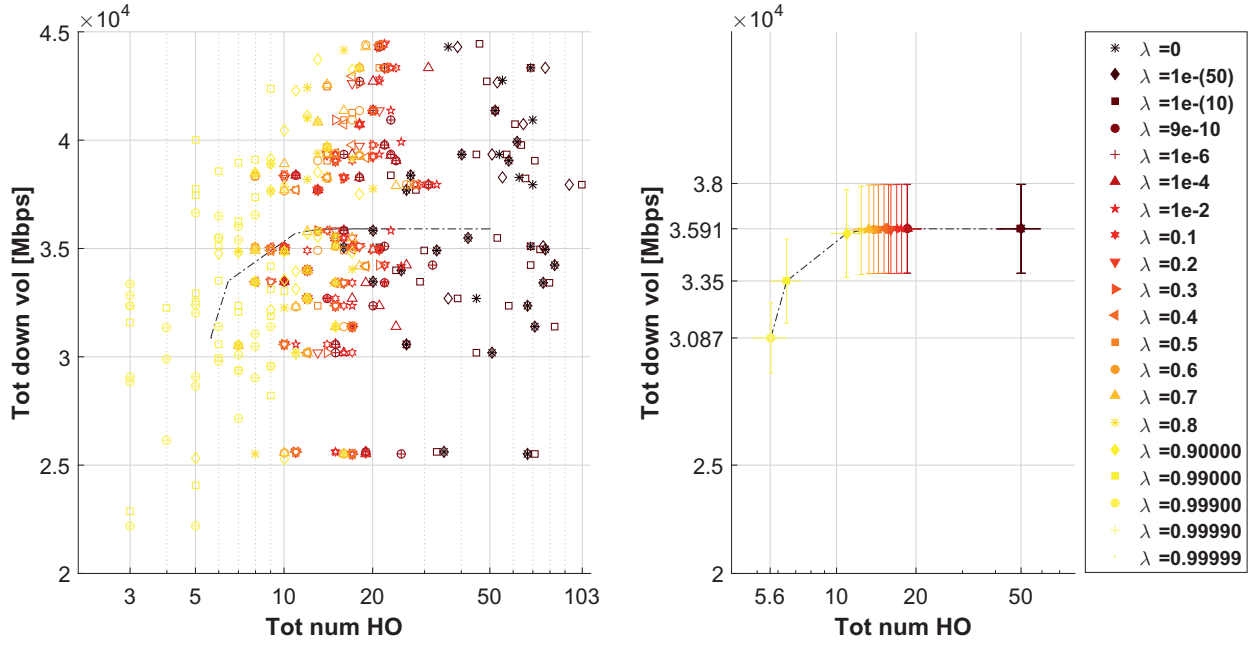


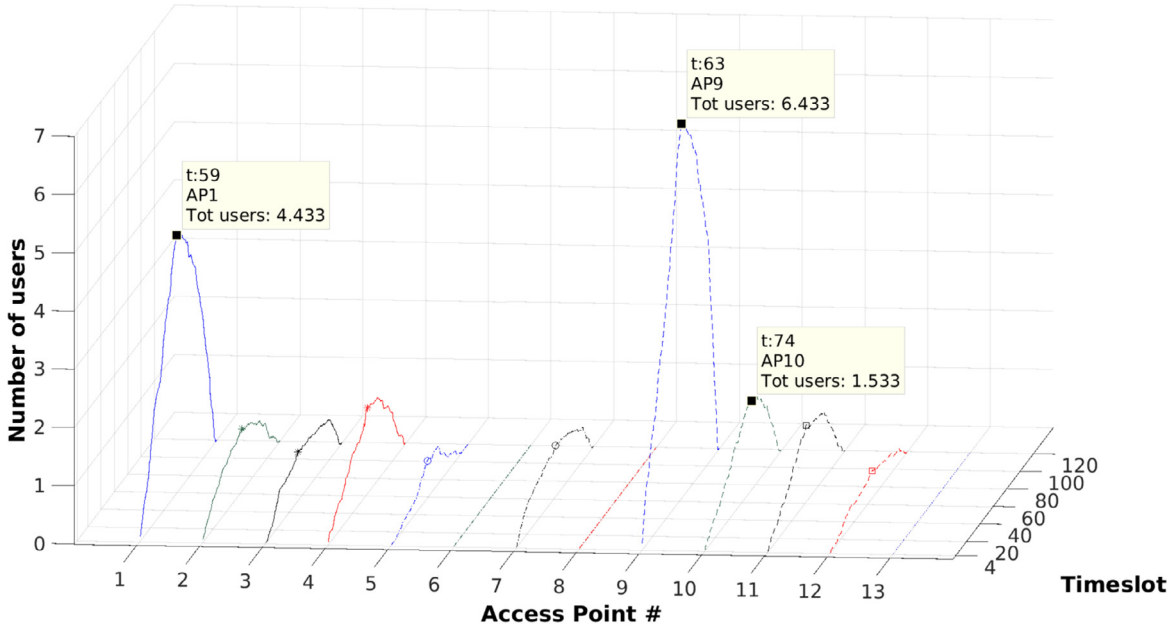
Fig. 4. Total download volume vs total number of HOs for different λ in scenario 2. A black dash-dot line connects average values. Error bars denote 95% CI. X axis is in logarithmic scale. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

to the random initial positions of the users over the 30 repetitions which may result in clustering users at some random APs.

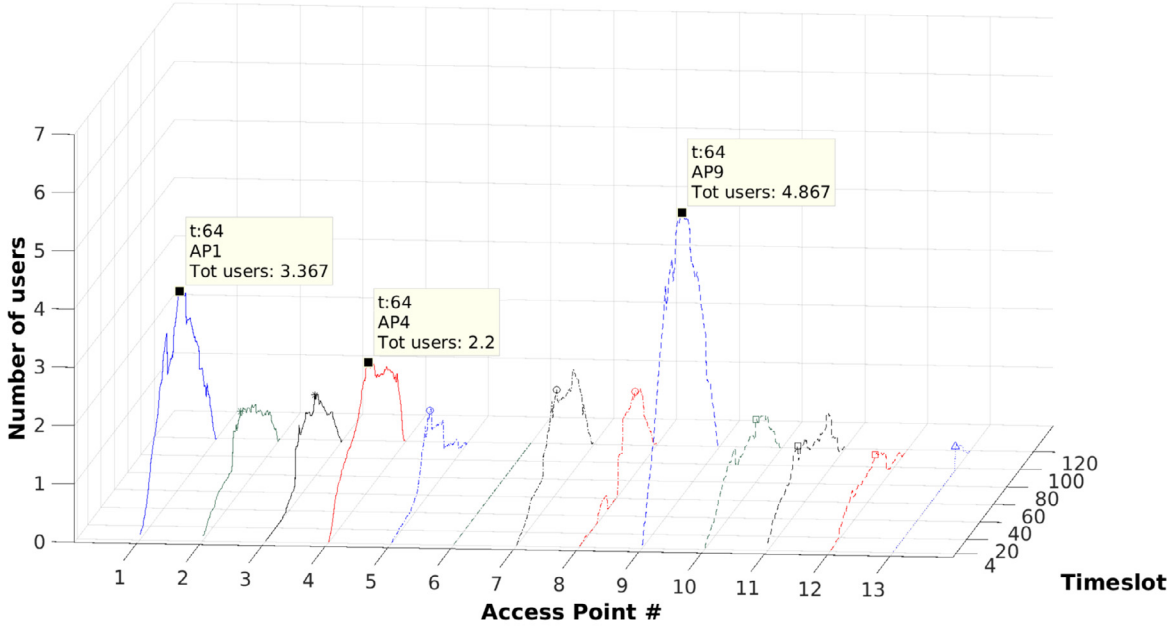
In multi-objective optimization, it is up to the decision maker to choose appropriate values for the weighting factor in order to balance the different conflicting objectives. As we can see from our scenario, one may set λ to any value between 0 and 0.990 and the total download volume is not influenced. Clearly $\lambda = 0.990$ results also in the minimum number of HOs. Even with very small λ (i.e., $\succ 1e^{-10}$ in this case study), one can effectively minimize the number of HOs in the system thus minimizing its negative impact on the network performance (i.e., increased probability of failed HO or potential collisions with other users' data traffic, increased probability of dropping the ongoing service, increased probability of TCP timeouts, increased power consumption, etc.).

5.2. Impact of user mobility

We have studied other scenarios in order to confirm the trends shown above. First, a mobile scenario in which 20 users walk around the building with a speed of 1 m/s is considered (scenario 2). Fig. 4 shows, on the left, the scatter plot of the total download volume and the total number of HOs for different values of λ. Again, high variability can be observed due to the randomness of the initial position and of the moving pattern followed by each user. For example, with $\lambda = 0$ (black asterisks), the total number of



(a) Static scenario



(b) Mobile scenario

Fig. 5. Total number of users for each AP and timeslot. Results are averaged over the 30 repetitions. No user ever connects to: (a) AP6, AP8 and AP13; (b) AP6.

HOs varies between 20 and 83, while the total download volume varies between 25.52 and 44.32 Gbits; when $\lambda = 0.01$ (red stars), the total number of HOs varies between 10 and 33, and the total download volume among 25.52 and 44.43 Gbits.

The plot on the right of Fig. 4 shows error bars and 95% CI. The average total download volume is higher compared to the static scenario (i.e., 35.90 Gbits compared to 31.66 of the static case) and, again, it remains stable for $\lambda \preceq 0.90$. When $\lambda \succ 0.90$, the average total download volume considerably decreases (i.e., it is 30.87 Gbits when $\lambda \succ 0.9990$). The average number of HOs in the system is higher in the mobile scenario , as one can expect. Again, it drastically drops from 50.17 to 18.57 when λ is increased from 0 to $9e^{-10}$, thus confirming the trend shown for the static case.

The total download volume for the mobile scenario is higher on average compared to the static one due to a more even distribution of the users among the APs over time caused by people roaming around in the building. Fig. 5 shows the total number of users per AP for each timeslot, averaged over the 30 random repetitions in scenario 1 and 2. While in the static scenario (Fig. 5(a)) there are two APs (i.e., AP1 and AP9) that on average have more than four users associated in the same timeslot, in the mobile scenario (Fig. 5(b)) only AP9 has more than four users associated in

Please cite this article as: E. Zola, A.J. Kassler, Minimizing the impact of the handover for mobile users in WLAN: A study on performance optimization, Computer Networks (2016), http://dx.doi.org/10.1016/j.comnet.2016.03.014

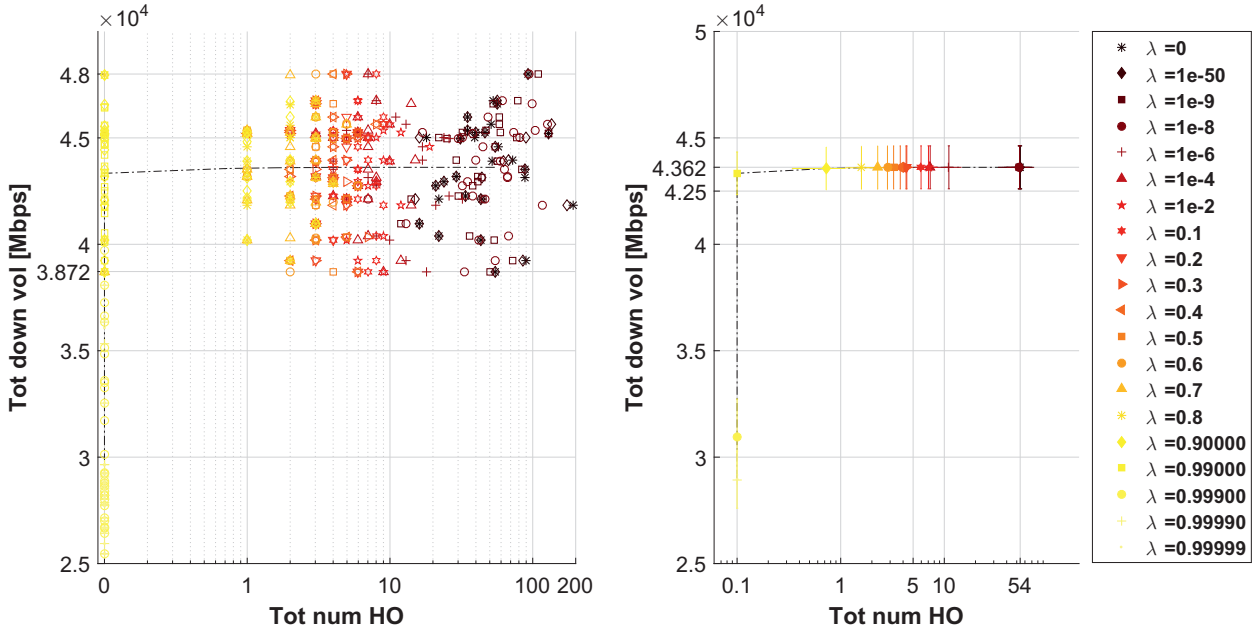


Fig. 6. Total download volume vs total number of HOs for different λ in scenario 3. A black dash-dot line connects average values. Error bars denote 95% CI. X axis is in logarithmic scale. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

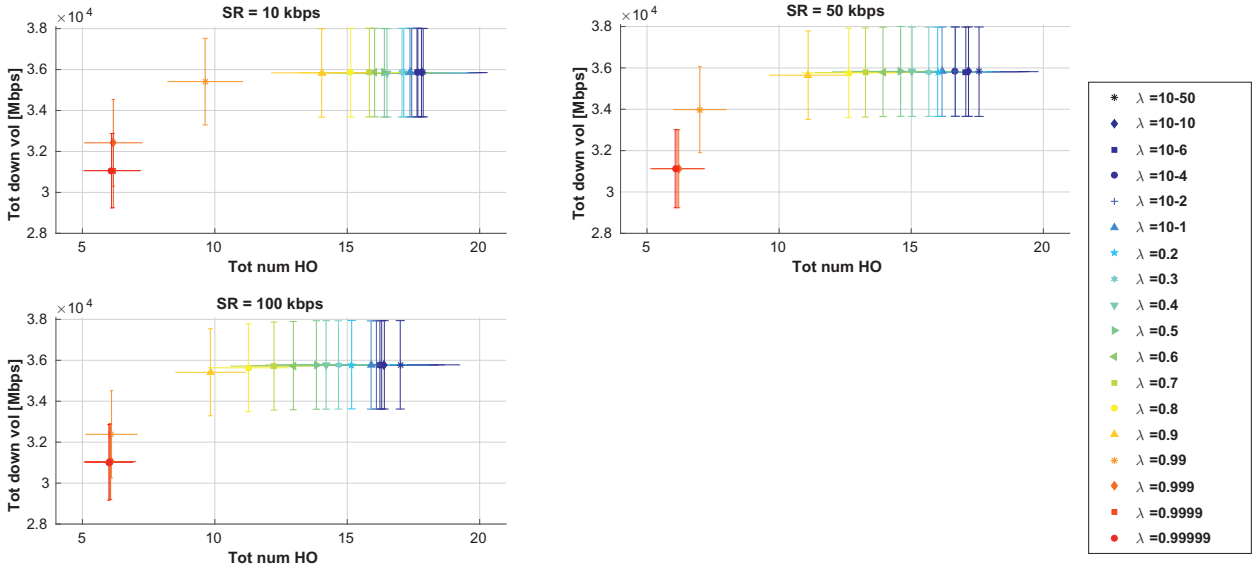


Fig. 7. Impact of *sr* on the system performance for different λ in scenario 2. Error bars denote 95% CI.

a given timeslot. Similarly, in the static scenario there are 3 APs with no users connected (i.e., APs 6, 8 and 13) while in the mobile scenario only AP6 has no users connected throughout the whole simulation.

5.3. Impact of users' density

When the number of users in the system increases, we expect the total download volume to increase. This is because in our simulation setup the system is not saturated, as shown in Fig. 6, where the scatter plot and error bars are displayed in the case of 40 static users (scenario 3). The average number of HOs is expected to increase too, as more users are appearing in the system and more reallocations (i.e., HOs) may be needed so as to provide high throughput. Again, the total download volume in the scenario with 40 users does not change for $\lambda \preceq 0.990$ (i.e., on average it is 43.62 Gbits), while the average number of HOs drops from 54.24 to 11.07 when λ is increased from 0 to $1e^{-6}$, and further decreases to 5.97 when $\lambda = 0.1$. We can also conclude that, as the CI reduces for both the download volume and the number of HOs, the scenario with more users exhibits higher stability.

In summary, we have seen that there is a wide range of the weighting parameter λ that provides a reduction in the number of HOs while maintaining the download volume constant. The range may slightly vary based on the scenario under study (i.e., number of users, user speed, etc.). However, $0.1 \leq \lambda \leq 0.9$ suggest a decrease from 2.5 to 0.8 in scenario 1 , from 16.9 to 10.9 in scenario 2 , and from 5.96 to 0.72 in scenario 3 . Although a bigger λ obviously implies a smaller amount of HO (on average), the difference is not as big as compared to the case with $\lambda = 0$.

Please cite this article as: E. Zola, A.J. Kassler, Minimizing the impact of the handover for mobile users in WLAN: A study on performance optimization, Computer Networks (2016), http://dx.doi.org/10.1016/j.comnet.2016.03.014

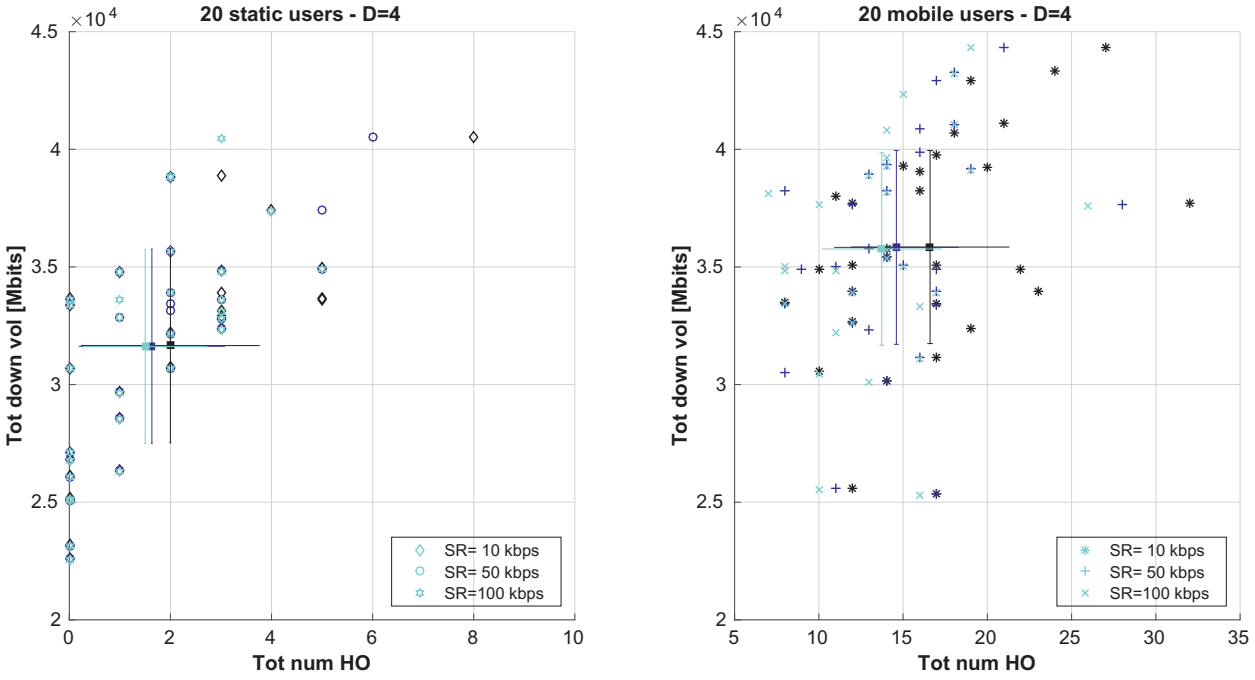


Fig. 8. Impact of *sr* on the system performance. Static (on the left) and mobile (on the right) scenarios with 20 users are displayed. Error bars denote CI.

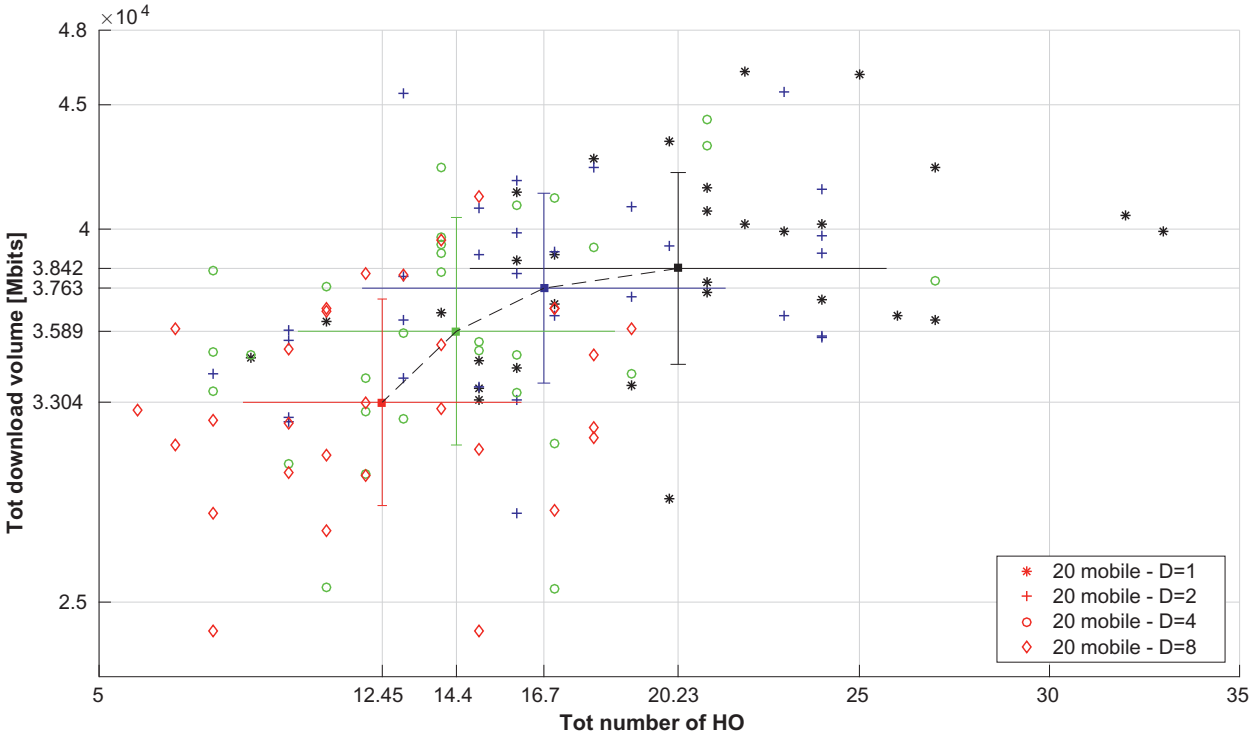


Fig. 9. Impact of the number of HO slots *D* on the system performance, in the case of 20 users moving at 1m/s. Error bars denote CI.

6. Impact of handover parameters

In this section, we study the impact of HO parameters (i.e., *sr* and $D_a^{(s)}$) on the network performance. First, the impact of the signaling rate in scenario 2 is evaluated. To this end, *sr* changes from 10 kbps, to 50 and to 100 kbps. Fig. 7 shows the average number of HO in the system and the average total download volume for different λ; 95% CI for both parameters are also drawn. The trend is very similar to the one depicted in Fig. 4; that is, the total download volume is constant when $\lambda \preceq 0.90$, while the number of HOs decreases as λ increases. When increasing the signaling rate, we observe that the number of HOs decreases. Also, we need to select a smaller λ if we want to avoid a decrease in the total download volume; that is, with $\lambda = 0.990$, when *sr* increases from 10 to 100 kbps, the average total download volume decreases by 3.03 Gbits, while it decreases only by 0.43 Gbits with $\lambda = 0.90$, 0.11 Gbits with λ=0.6, 0.08 Gbits with λ=0.5, and 0.07 Gbits with $\lambda = 10^{-50}$; the decrease in the number of HOs is 3.5, 4.2, 3.1, 2.6 and 0.8,

Please cite this article as: E. Zola, A.J. Kassler, Minimizing the impact of the handover for mobile users in WLAN: A study on performance optimization, Computer Networks (2016), http://dx.doi.org/10.1016/j.comnet.2016.03.014

respectively. As a conclusion for this case study, for higher *sr* a very small λ (i.e., 10^{-50}) will provide the lowest decrease in the total download volume at the expenses of a lower decrease in the number of HO, while a bigger λ (i.e., 0.9) will provide the highest decrease in the number of HO at the expenses of a higher decrease in the total download volume.

We now compare the impact of *sr* in the static and in the mobile scenarios with 20 users when $\lambda = 0.5$. Fig. 8 shows a scatter plot of the total download volume versus the total number of HOs for different *sr*; the average and 95% CIs are also shown (filled in marks plus bars, respectively). Again, the average number of HOs decreases when *sr* increases, while the total download volume remains constant. In the mobile case study, both the number of HOs and the total download volume increase with respect to the static case.

Finally, we evaluate the impact of a higher HO duration on the system performance, by increasing the number of HO slots *D*. The mobile scenario with 20 users (scenario 2) is considered, as we expect a higher amount of HOs compared to the static scenario as already previously shown. When *D* is increased (from right to left in Fig. 9), the average number of HOs decreases, as expected. The average total download volume decreases, too, due to the increased time needed for performing the authentication and association process to the new AP. This results in less download opportunities and thus reduced total download volume. For example, for fast HO ($D = 1$), the average number of HOs in the system is 20.23 at an average total download volume of 38.42 Gbits. When the HO duration increases much ($D = 8$), the average number of HOs reduces to 12.45 while at the same time the total download volume reduces to 33.04 Gbits.

As a conclusion, in case the HO would be very fast and the signaling cost negligible, then there would be no reason for restricting the number of handovers, as in that case the probability for a handover failure would be negligible. The algorithm proposed here is able to take into account the handover cost and to provide an association pattern that avoids handovers when they are potentially harmful.

7. Conclusion

IEEE 802.11 based WLANs are an important part of the future wireless Internet. Understanding their performance is crucial in order to gain an insight into the different system tradeoffs involved. In this paper, we have built a mathematical model that allows to study the tradeoff between the number of HOs and the total system rate for a set of IEEE 802.11 based access points, that are connected to the Internet. Our model is built on a MILP with the objective of both maximizing the total download rate while at the same time minimizing the number of HOs with a configurable HO signaling rate. Because the objective function is composed of two conflicting terms, we have studied the tradeoff between the two objectives in the optimization function, which allowed us to find a good compromise between download rate and number of HOs. An important aspect of our model is the total knowledge of all system parameters over the whole simulation duration, which is typically not known in advance.

As future work, we intend to apply robust optimization techniques on our model in order to cope with unknown or erroneous predicted demands, mobility patterns and download rates. By applying robust optimization techniques we expect to get better insight into the tradeoffs involved when the exact values of several parameters of the system are not precisely known. Also, we intend to study fast heuristics that can be applied when some parameters of the system changes in order to react fast during critical situations.

Acknowledgment

This research was partially supported by the Spanish Government and ERDF through CICYT project TEC2013-48099-C2-1-P and by the WINEMO IC0906 “Wireless Networking for Moving Objects” COST Action.

References

[1] P. Dely, A. Kassler, L. Chow, N. Bambos, N. Bayer, H. Einsiedler, C. Peylo, BEST-AP: Non-intrusive estimation of available bandwidth and its application for dynamic access point selection, Comput. Commun. 39 (2014) 78–91.
[2] S. Thajchayapong, J.M. Peha, Mobility patterns in microcellular wireless networks, IEEE Trans. Mobile Comput. 5 (1) (2006) 52–63.
[3] W.-J. Hsu, A. Helmy, On modeling user associations in wireless LAN traces on university campuses, in: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, 2006 4th International Symposium on, 2006, pp. 1–9.
[4] E. Zola, F. Barcelo-Arroyo, Characterizing user behavior in a european academic WiFi network, Int. J. Handheld Comput. Res. 4 (2) (2013) 55–68.
[5] A. Mishra, M. Shin, W. Arbaugh, An empirical analysis of the IEEE 802.11 MAC layer handoff process, SIGCOMM Comput. Commun. Rev. 33 (2) (2003) 93–102.
[6] H. Velayos, G. Karlsson, Techniques to reduce the IEEE 802.11b handoff time, in: Communications, 2004 IEEE International Conference on, 7, 2004, pp. 3844–3848.
[7] S. Jin, M. Choi, S. Choi, Multiple WNIC-based handoff in IEEE 802.11 WLANs, Commun. Lett. IEEE 13 (10) (2009) 752–754.
[8] K. Munasinghe, A. Jamalipour, Analysis of signaling cost for a roaming user in a heterogeneous mobile data network, in: Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE, 2008, pp. 1–5.
[9] S. Johnson, S. Nath, T. Velmurugan, An optimized algorithm for vertical handoff in heterogeneous wireless networks, in: Information Communication Technologies (ICT), 2013 IEEE Conference on, 2013, pp. 1206–1210.
[10] E. Zola, F. Barcelo-Arroyo, A. Kassler, Multi-objective optimization of WLAN associations with improved handover costs, Commun. Lett. IEEE 18 (11) (2014) 2007–2010.
[11] P. Dely, A. Kassler, N. Bayer, H.J. Einsiedler, C. Peylo, Optimization of WLAN associations considering handover costs, EURASIP J. Wireless Comm. Netw. 255 (2012) 1–14.
[12] E. Zola, F. Barcelo-Arroyo, Impact of mobility models on the cell residence time in WLAN networks, in: Sarnoff Symposium, 2009. SARNOFF '09. IEEE, 2009, pp. 1–5.
[13] M. Abusubaih, A. Wolisz, An optimal station association policy for multi-rate IEEE 802.11 wireless LANs, in: Proceedings of the 10th ACM Symposium on Modeling, Analysis, and Simulation of Wireless and Mobile Systems, in: MSWiM '07, ACM, 2007, pp. 117–123.
[14] A. Baid, M. Schapira, I. Seskar, J. Rexford, D. Raychaudhuri, Network cooperation for client-AP association optimization, in: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2012 10th International Symposium on, 2012, pp. 431–436.
[15] W. Li, Y. Cui, S. Wang, X. Cheng, Approximate optimization for proportional fair AP association in multi-rate WLANs, in: G. Pandurangan, V.S. Anil Kumar, G. Ming, Y. Liu, Y. Li (Eds.), Wireless Algorithms, Systems, and Applications: 5th International Conference, Proceedings (WASA), 2010, pp. 36–46.
[16] Y. Bejerano, S.-J. Han, L. Li, Fairness and load balancing in wireless LANs using association control, IEEE/ACM Trans. Netw. 15 (3) (2007) 560–573.
[17] H. Gong, K. Nahm, J.W. Kim, Distributed fair access point selection for multi-rate IEEE 802.11 WLANs, in: Consumer Communications and Networking Conference (CCNC), 5th IEEE, 2008, pp. 528–532.
[18] G. Kasbekar, P. Nuggehalli, J. Kuri, Online client-AP association in WLANs, in: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, 2006 4th International Symposium on, 2006, pp. 1–8.
[19] R.V. Sathya, V. Venkatesh, R. Ramji, A. Ramamurthy, B.R. Tamma, Handover and SINR optimized deployment of LTE femto base stations in enterprise environments, Wireless Pers. Commun. (2016) 1–25 http://link.springer.com/journal/11277/onlineFirst/page/4.
[20] A. Roy, J. Shin, N. Saxena, Multi-objective handover in LTE macro/femto-cell networks, J. Commun. Netw. 14 (5) (2012) 578–587.
[21] G. Athanasiou, P.C. Weeraddana, C. Fischione, L. Tassiulas, Optimizing client association for load balancing and fairness in millimeter-wave wireless networks, IEEE/ACM Trans. Netw. 23 (3) (2015) 836–850.
[22] P. Dely, J. Vestin, A. Kassler, N. Bayer, H. Einsiedler, C. Peylo, CloudMAC - An openFlow based architecture for 802.11 MAC layer processing in the cloud, in: Globecom Workshops, 2012 IEEE, 2012, pp. 186–191.
[23] IBM, IBM ILOG CPLEX, (http://www.ibm.com/developerworks/downloads/ws/ilogcplex/ (last accessed on Sept. 2014)).

Please cite this article as: E. Zola, A.J. Kassler, Minimizing the impact of the handover for mobile users in WLAN: A study on performance optimization, Computer Networks (2016), http://dx.doi.org/10.1016/j.comnet.2016.03.014

Enrica Zola received the double M.Sc. degree in Telecommunications Engineering from both Politecnico di Torino (Italy) and Universitat Politécnica de Catalunya (UPC, Spain), in 2002 and 2003, respectively. In 2011, she earned a Ph.D. from the UPC. From September 2001 to August 2002, she collaborated with the Radio Department of the Spanish teleoperator Amena. From March 2003 to February 2006, she has been working at UPC as a full-time Lecturer. From March 2006, she serves as an Assistant Professor at the Department of Telematics Engineering at UPC. She has been teaching design and planning of communication networks and wireless networks. Dr. Zola has been involved in a number of research projects supported by the Spanish Government and the European Commission on performance modeling of wireless systems and networks (IST Emily, RUBI, IST Liaison, COST Winemo, COST290). Her research interest areas encompass wireless networking in general, with special attention to mobility management and radio resource management. Recently, her interest has focused on performance optimization modeling and robust optimization techniques.

Andreas J. Kassler is Professor of Computer Science at Karlstad University, Karlstad, Sweden, that he joined in 2005. From 2003 to 2004, he was Assistant Professor at the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests are in the area of Wireless Meshed Networks, Ad-Hoc Networks, Quality of Service, and Software Defined Networking. He has published over 100 conference and journal papers, and several book chapters. He received the Docent title in Computer Science from Karlstad University in 2006, the Ph.D. degree in Computer Science from Universität Ulm, Germany, in 2002 and an M.S. degree from Universität Augsburg, Germany. Prof. Kassler is a Member of the IEEE and IEEE Communication Society.

Please cite this article as: E. Zola, A.J. Kassler, Minimizing the impact of the handover for mobile users in WLAN: A study on performance optimization, Computer Networks (2016), http://dx.doi.org/10.1016/j.comnet.2016.03.014