Optimizing Client Association in 60 GHz Wireless Access Networks

George Athanasiou, *Member, IEEE,* Pradeep Chathuranga Weeraddana, *Member, IEEE,*
Carlo Fischione, *Member, IEEE,* and Leandros Tassiulas, *Fellow, IEEE,*

Abstract—MillimeterWave wireless communications in the 60 GHz band are considered as the key technology for enabling multi-gigabit wireless access. However, the high propagation loss in such a band poses major obstacles to the optimal utilization and fairness of the wireless services, where the problem of efficient association of clients to access points (APs) is of vital importance. In this paper, such a fundamental problem of optimizing the assignment of the clients to the available APs in 60 GHz wireless access networks is studied. The AP utilization and the supported transmission rates over the rapidly vanishing millimeterWave communication links are the main parameters in the optimization. Because of the tricky non-convex and combinatorial nature of the client association optimization problem, new fundamental solution methods are proposed to guarantee balanced and fair resource allocation. A new distributed, lightweight, and easy to implement association algorithm is established based on an iterative procedure that uses Lagrangian duality theory and subgradient methods. It is shown that the algorithm is asymptotically optimal as the relative duality gap diminishes to zero when the number of clients increases. Both theoretical and numerical results evince numerous useful properties, such as fast convergence, scalability, time efficiency, and fair execution in comparison to existing association approaches. It is concluded that the proposed solution can be applied it the forthcoming 60 GHz wireless networks and can support new bandwidth-demanding applications.

Index Terms—60 GHz Wireless Access Networks, Resource Allocation, Association Control

I. INTRODUCTION

MillimeterWave (mmW) communications have attracted the interest of academia, industry, and standardization bodies, although the technology was invented and used many decades ago, especially in the context of military applications [1], [2]. MmW communications utilizes the part of the electromagnetic spectrum between 30 and 300 GHz, which corresponds to wavelengths from 10 mm to 1 mm [3]. The continuous development of Complementary Metal Oxide Semiconductor technology [4], with low cost and low power consumption, has enabled the use of such a mmW spectrum in the communication of the user devices and in the provisioning of QoS-sensitive applications.

Due to the 60 GHz wireless networks great commercial potential, multiple industry-led efforts and international organizations have emerged for the standardization activities [5]. Examples include IEEE 802.15.3, IEEE 802.15.3 Task Group 3c (TG3c) [6], IEEE 802.11 standardization

group, the task group IEEE 802.11ad [7], IEEE 802.11ad, the WirelessHD [8] consortium, the Wireless Gigabit Alliance (WiGig) [9], and many others. More than 5 GHz of continuous bandwidth is available in many countries worldwide, which makes 60 GHz systems particularly attractive for gigabit wireless applications such as gigabyte file transfer, wireless docking station, wireless gigabit ethernet, wireless gaming, and uncompressed high definition video transmission. Thus, 60 GHz radio technology is considered as one of the most promising candidates to boost wireless communication data rates to the order of multi-Gbps.

The 60 GHz huge bandwidth represents great potential in terms of capacity and flexibility. For example, even with a low spectral efficiency such as 0.4 bps/Hz, 60 GHz communication systems can provide a very high data rate of 1 Gbps, making it an ideal candidate to support simple modulation techniques. In contrast, ultra-wideband (UWB) systems require roughly 2 bps/Hz to achieve 1 Gbps, but its actual deployment is limited to 400 Mbps at 1 m operating range. The estimated spectral efficiency for IEEE 802.11n [10] communication systems is 25 bps/Hz for achieving 1 Gbps, which makes their application to high bandwidth services unacceptable in terms of cost and simple implementation. Moreover, the 60 GHz communication systems are less restricted in terms of power limits, as opposed to UWB systems [11]. For example, the regulation related to 60 GHz band allows much higher transmission power, compared to other existing wireless local area and personal systems.

However, there are several challenging aspects and potentialities that must still be addressed in mmW communications. On the negative side, there are severe channel attenuations. Higher power levels are necessary to overcome the high path loss. The equivalent isotropic radiated power in the 60 GHz system is approximately 10 times larger than the IEEE 802.11n [10] and 30000 times larger than the UWB systems. The typical operation range of a 60 GHz communications is 10 to 20 meters [6]. For example, a path loss of 2 meters results in attenuation of approximately 74 dB. Thus directionality and blockage problems are present. On the positive sides, the interference levels for 60 GHz are much lower compared to the congested 2.4 GHz and 5 GHz bands. The compact design of 60 GHz radio allows the use of multiple-antenna solutions at the user device. Compared to 5 GHz systems, the form factor is approximately 140 times smaller and thus can be conveniently integrated into consumer electronic products. Therefore, the exploitation of these unique characteristics is essential for efficient radio resource management algorithms.

In this paper we address the fundamental resource allocation problem of the client association in 60 GHz networks. We

G. Athanasiou, P. C. Weeraddana and C. Fischione are with Electrical Engineering School, Access Linnaeus Center, KTH Royal Institute of Technology, Stockholm, Sweden. E-mail: {georgioa, chatw, carlofi}@kth.se

L. Tassiulas is with Computer and Communication Engineering Department, University of Thessaly, Volos, Greece. E-mail: leandros@uth.gr

consider the natural situation where each client has to be associated to one of the wireless APs present in the system, and the load among the APs must be balanced. The main goal is to *minimize the maximum AP utilization* in the network. This gives a mixed integer linear optimization problem, which is combinatorial and non-convex in general and thus hard to solve efficiently. Nevertheless, based on Lagrangian duality theory [12, § 5] and on subgradient methods, we establish a new fundamental solution approach and derive a distributed and iterative algorithm for client association (*DAA*). We derive an analytical bound for the duality gap and show the asymptotic optimality of the proposed solution method. However, sufficient conditions under which strong duality holds are also characterized. The sensitivity of the convergence speed of the *DAA* to the variation of the numbers of clients and APs is analytically investigated. Numerical simulations compare *DAA* to benchmark algorithms, including the existing standard association approaches, and some others from the literature.

Unlike the client association approaches in more traditional access networks [13], we take into account the load, the channel quality, and the special communication characteristics of 60 GHz radio channel. Because the client association problem is combinatorial, global methods [14] are necessary to find the solution. However, global approaches have the drawbacks of 1) the prohibitive computational complexity, even in the case of problems with few variables and 2) they are inherently centralized, which makes their distributed implementation very difficult if not impossible. In contrast, our proposed method is *fast* and can be implemented in a *distributed* manner, which is crucial for many applications. Last but not least, the model and the lightweight algorithm proposed in this work are general and can be applied with different existing MAC mechanisms for 60 GHz networks, such as IEEE 802.15.3c and IEEE 802.1ad. We discuss in detail the applicability of the association algorithm on top of the current standardized medium access control (MAC) mechanisms.

The rest of the paper is organized as follows. In § II we present related approaches in literature. A description of the system model and the problem formulation in presented in § III. In § IV, we provide the general solution approach to the client association problem by using duality theory and subgradient method. In § V, we describe extensively the important properties of the proposed algorithm. In § VI numerical results are presented. Lastly, § VII concludes the paper.

II. RELATED WORK

During the last decade, resource allocation and in particular association/handoff control for wireless local area networks (WLANs) have been the focus of intense research. Several studies analyzed the performance of the basic association policy that IEEE 802.11 standard defines based on the received signal strength indicator (RSSI). These studies showed that this basic association policy can lead to inefficient use of the network resources [13], [15], [16]. Therefore, there has been increasing interest in designing better client association policies. In what follows we summarize some representative approaches in literature.

The authors in [16] study a client association policy that ensures network-wide max-min fair bandwidth allocation to the users. The work in [17] presents self-configuring algorithms that provide improved client association and fair resource sharing in the wireless network. In [18] the problem of optimal user association to the available APs is formulated as a utility maximization problem. In [19], neighbor and non-overlap graphs are utilized in order to reduce the probing latency. Moreover, the "multi-homing" scenario is introduced in [20], where the traffic is split among the available APs. In this approach the throughput is maximized by constructing a fluid model of user population that is multi-homed by the available APs in the network. The system presented in [13] ensures fairness and quality of service (QoS) provisioning in WLANs with multiple APs. The authors compare their system with 802.11 and prove that 802.11 cannot support both fairness and QoS guarantees.

Several approaches in literature jointly consider the association and handoff problems in WLANs. The authors in [21] propose an association/handoff mechanism that is based on monitoring of the performance of wireless links. In [15], there is a detailed analysis of the 802.11 association/handoff process. The presented approach optimizes the probe phase of the association process to reduce the probe latency. In [22], the authors propose protocols for an architecture called *SMesh*, which supports fast handoff introducing collaboration between the APs in a mesh network. The work in [23] presents a pre-authentication mechanism that speeds up the handoff process in wireless mesh networks.

Another line of research considers user service requests by readjusting the load across all APs [24]. Therein, a network-directed roaming approach suggests to the clients the most appropriate network location to satisfy their requests. In [25] and [26], a dual-association approach in wireless mesh networks is presented, where the APs for unicast traffic and the APs for broadcast traffic are independently chosen by exploiting overlapping coverage and optimizing the overall network load. Lastly, in [27], [28] dynamic association and reassociation procedures are introduced with the use of the notion of the *airtime cost*. This metric reflects the uplink/downlink channel conditions and the traffic load in the network. The cross-layer extension of this mechanism considers the routing-based information from the mesh backbone. Accordingly, the clients optimize their association/handoff decision.

The mentioned approaches are hard to be applied in 60 GHz local area wireless networks due to the special characteristics of the 60 GHz channel and the obvious differences with the rest wireless access technologies that we have previously mentioned. It follows that novel mechanisms must be designed to provide optimal resource allocation. These mechanisms must take into account the characteristics of these networks such as increased path loss, short range, fragile links, etc. Unfortunately, there is no much research in this field for 60 GHz wireless networks.

Some recent interesting approaches on 60 GHz wireless personal and local area networks have appeared in the literature. The authors in [29], study the throughput of mmWave networks. A combination of carrier sense multiple access (CSMA/CA) and time division multiple access (TDMA) is used. A technique for optimizing the throughput of the network is proposed, where collisions are reduced in the CSMA/CA time-slots by a private channel release time. In

[30], a MAC protocol that employs memory and learning to address deafness, while exploiting the reduction of interference between simultaneous transmissions, is proposed with reactive interference management. The directionality and blockage problems of mmW networks are studied in [31]. A cross-layer approach is presented, where a single hop transmission is preferred when line of sight (LOS) is available and a relay node is randomly selected as alternative. In [32], a resource management mechanism is proposed based on the exclusive region (ER) to exploit the spatial reuse of mmW networks. The authors in [33] describe an interference analysis framework that enables a quantitative evaluation of collision loss probability for a mmW mesh network with uncoordinated transmissions, as a function of the antenna patterns and spatial density of simultaneously transmitting nodes. Concurrent transmissions are studied in [34] by exploiting the spatial reuse and time division multiplexing gain. It is shown that the network throughput and flow throughput are significantly improved compared with single hop transmission schemes.

The current 60 GHz standardization bodies, such as IEEE 802.11ad, IEEE 802.15.3c, adopt the RSSI-based mechanism as the basic association/handoff functionality. However, RSSI is not an efficient decision metric for user association for several reasons. High RSSI values cannot univocally indicate high throughput. This is because RSSI not only depends on the distance from the APs, but also on the transmission powers of the APs. The accuracy of the RSSI-based technique is significantly affected by the high path loss, dispersion and directionality of the 60 GHz wireless channel. Moreover, since the wireless channel is a shared medium, throughput depends on the population of the cell served by the AP. An AP may become overloaded if large number of stations are associated with it. Therefore, other metrics must be used that will reflect the channel characteristics and the utilization of the AP.

In contrast to the existing work in literature, in this paper we consider the special characteristics of the 60 GHz channel in an optimization problem where the objective is to minimize the maximum AP utilization in the network. Moreover, we design a lightweight distributed algorithm that balances the AP utilization by optimizing the client association process. We believe that this paper is the first to study such a fundamental resource allocation problem in 60 GHz wireless access networks. We propose a simple yet efficient solution that is compared to basic association policies, already in use in the present mmW technologies under standardization (802.15.3c, 802.11ad). The work that is presented and evaluated in the forthcoming sections is complementary to the aforementioned resource management and scheduling approaches in 60 GHz wireless access networks, while the clients must first be associated to the available APs and then the scheduling of the transmission can be handled under the established association state.

III. System Model and Problem Formulation

A mmW wireless network consisting of N APs and M clients is considered. We denote by $\mathcal{N}$ the set of APs and by $\mathcal{M}$ the set of clients, where $\mathcal{N} = \{1, \ldots, N\}$ and $\mathcal{M} = \{1, \ldots, M\}$. The set of clients that can be associated to AP i is denoted by $\mathcal{M}_i$. We assume that there are no isolated clients, i.e., $\mathcal{M}_1 \cup \cdots \cup \mathcal{M}_N = \mathcal{M}$. We denote by $\mathcal{N}_j$ the set
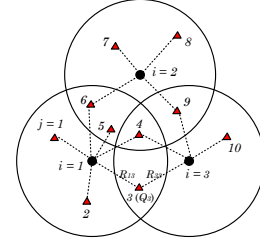

Figure 1. Wireless access network: $\mathcal{N} = \{1, 2, 3\}$, $\mathcal{M} = \{1, \ldots, 10\}$, $\mathcal{M}_1 = \{1, \ldots, 6\}$, $\mathcal{M}_2 = \{4, \ldots, 9\}$, $\mathcal{M}_3 = \{3, 4, 9, 10\}$, $\mathcal{N}_j = \{1\}$ for $j \in \{1, 2\}$, $\mathcal{N}_j = \{2\}$ for $j \in \{7, 8\}$, $\mathcal{N}_j = \{3\}$ for $j = 10$, $\mathcal{N}_j = \{1, 2\}$ for $j \in \{5, 6\}$, $\mathcal{N}_j = \{1, 3\}$ for $j = 3$, $\mathcal{N}_j = \{2, 3\}$ for $j = 9$, $\mathcal{N}_j = \{1, 2, 3\}$ for $j = 4$. The area inside solid-lined circles around APs $1, 2$, and 3 represent the transmission regions of each AP. Demanded data rate for client 3 is Q_3, offered transmission rate from AP 1 to client 3 is R_{13}, and offered transmission rate from AP 3 to client 3 is R_{33}.

of candidate APs that client j could be associated with. An example of access network is shown in Figure 1, where the clients positioned inside a disc with radius r centered at the location of AP i can be associated with that AP.

Each nodes is equipped with steerable directional antennas and it can direct its beams to transmit or to receive. We assume that AP i can support all the clients in $\mathcal{M}_i$ with a separate transmit beam. We consider the case where all receiver nodes are using single-user detection (i.e., a receiver decodes each of its intended signals by treating all other interfering signals as noise) and assume that the achievable rate from AP i to client $j \in \mathcal{M}_i$ is

$$R_{ij} = W \log_2 \left(1 + \frac{P_{ij} G_{ij}}{(N_0 + I_j)W} \right) , \qquad (1)$$

where W is the system bandwidth, P_{ij} is the transmission power of AP i to client j, G_{ij} is the power gain from AP i to client j, N_0 is the power spectral density of the noise at each receiver, and I_j is the interference spectral density at client j. The power gain G_{ij} is modeled as in [34]. In particular, we use the Friis transmission equation together with the *flat-top* transmit/recieve antenna gain model [35], where a fixed gain is considered within the beamwidth and zero gain is considered outside the beamwidth of the antenna. In addition, we consider Rayleigh small-scale fading. Thus we have

$$G_{ij} = \frac{G_{ij}^{\mathrm{Tx}} G_{ij}^{\mathrm{Rx}} \lambda^2 \alpha_{ij}}{16\pi^2 (d_{ij}/d_0)^\eta} , \ i \in \mathcal{N}, j \in \mathcal{M}_i , \qquad (2)$$

where G_{ij}^{Tx} is the transmit antenna gain from AP i to client j, G_{ij}^{Rx} is the receive antenna gain from AP i to client j, λ is the wavelength, α_{ij} is the *fading coefficient* which is an exponentially distributed random variables with unit mean to models the Rayleigh small-scale fading, d_{ij} is the distance between AP i and client j, d_0 is the *far field reference distance* [36], η is the path loss exponent,[1] and I_j is the communication interference at client j. We capitalize on the well studied 60 GHz propagation characteristics [35], such as highly directional transmissions with very narrow beamwidths and increased path losses due to the oxygen absorption, in order to assume that the communication interference I_j is very small and does not affect significantly the achievable commu-

---

[1] $\eta \in [2, 6]$ in IEEE 802.11ad networks [37].

nication rates in the network.[2] The achievable communication rates given in (1) are hypothetical and used to define the AP utilizations as described in the sequel.

We denote by Q_j the *demanded data rate* of client j. The channel utilization of AP i and client j is denoted by β_{ij} and is given by the ratio of Q_j and R_{ij}, i.e.,

$$\beta_{ij} = \frac{Q_j}{R_{ij}} \ . \tag{3}$$

Intuitively, the channel utilization β_{ij} gives an indication of the communications performance, in terms of the *potential* loading of the communication channel between AP i and client j. Thus, the sum of channel utilizations of AP i (or AP i utilization) is given by $\sum_{j\in\mathcal{M}_i}\beta_{ij}x_{ij}$, where $(x_{ij})_{j\in\mathcal{M}_i}$ are binary decision variables, which indicate the client association. In particular, for all $i \in \mathcal{N}$ and $j \in \mathcal{M}_i$

$$x_{ij} = \begin{cases} 1 & \text{if client } j \text{ is associated to AP } i \\ 0 & \text{otherwise .} \end{cases} \tag{4}$$

AP i utilization is a metric reflecting the load of the AP. Our goal is to minimize the maximum AP utilization. Specifically, the problem can be formally express as

$$\text{minimize} \quad \max_{i\in\mathcal{N}} \sum_{j\in\mathcal{M}_i} \beta_{ij}\, x_{ij} \tag{5a}$$

$$\text{subject to} \quad Q_j x_{ij} \leq R_{ij}, \ i \in \mathcal{N}, j \in \mathcal{M}_i \tag{5b}$$

$$\sum_{i\in\mathcal{N}_j} x_{ij} = 1, \ j \in \mathcal{M} \tag{5c}$$

$$x_{ij} \in \{0,1\}, \ j \in \mathcal{M}, i \in \mathcal{N}_j \ , \tag{5d}$$

where the variable is $(x_{ij})_{i\in\mathcal{N},\ j\in\mathcal{M}_i}$. The main problem parameters are $(\beta_{ij})_{i\in\mathcal{N},j\in\mathcal{M}_i}$, $(Q_j)_{j\in\mathcal{M}}$, and $(R_{ij})_{i\in\mathcal{N},j\in\mathcal{M}_i}$. Constraint (5b) assure that the demand of client j is less or equal to the achievable rate from AP i to client j. This constraint can usually be satisfied due to the huge available bandwidth of mmW communications. Constraint (5c) ensures that client j is assigned to only one AP. Constraint (5d) indicates that the decision variables are binary.

Note that problem (5) is combinatorial, and therefore we have to rely on exponentially complex global methods [14] to *solve* the problem. Nevertheless, in the sequel, we present an efficient, yet nonoptimal solution approach, followed by its asymptotic optimality results.

IV. Solution via dual problem

Without loss of generality, we assume that $\beta_{ij} \leq 1$ for all $i \in \mathcal{N}$ and $j \in \mathcal{M}_i$.[3] As a result, the constrain (5b) is redundant and can be dropped. Thus, the standard equivalent

[2] In particular, a probabilistic analysis of the interference incurred due to uncoordinated transmissions is presented in [35]. It is shown that even uncoordinated transmission for different transmit-receive pairs leads to small collision probabilities and therefore, the links in the network can be thought of as *pseudo-wired*. That is, interference can essentially be ignored in MAC or higher layers design. Similar assumptions are also supported by using efficient channel allocation in the network [38] and efficient scheduling algorithms that supports concurrent transmissions [34].

[3] $\beta_{ij} > 1$ means that one demands a data rate more than what the wireless channel can provide. In this case, we can collect all the i,j pairs for which $\beta_{ij} > 1$ and modify the corresponding sets $\mathcal{N}_j$ and $\mathcal{M}_i$, without affecting the optimal value of the original problem (5). In particular, if $\beta_{ij} > 1$, then we set $\mathcal{N}_j := \mathcal{N}_j \setminus \{i\}$ and $\mathcal{M}_i := \mathcal{M}_i \setminus \{j\}$. For example, suppose that $\beta_{12} > 1$, and the corresponding sets $\mathcal{N}_2 = \{1,2\}$ and $\mathcal{M}_1 = \{1,2,3\}$. Then, we simply modify $\mathcal{N}_j$ and $\mathcal{M}_i$ as follows: $\mathcal{N}_2 = \{2\}$, $\mathcal{M}_1 = \{1,3\}$.

epigraph form [12, § 4.1.3] of problem (5) is given by,

$$\text{minimize} \quad t \tag{6a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{M}_i} \beta_{ij}\, x_{ij} \leq t, \ i \in \mathcal{N} \tag{6b}$$

$$\sum_{i\in\mathcal{N}_j} x_{ij} = 1, \ j \in \mathcal{M} \tag{6c}$$

$$x_{ij} \in \{0,1\}, \ j \in \mathcal{M}, i \in \mathcal{N}_j \ , \tag{6d}$$

where the variables are t and $\mathbf{x} = (x_{ij})_{i\in\mathcal{N},\ j\in\mathcal{M}_i}$. We denote by $p^\star$ the optimal value of the problem (6). Note that in problem (6) is a mixed integer linear program (MILP) and is feasible. The existing MILP solvers are, of course, centralized and are typically based on global branch and bound algorithms, where the worst-case complexity grows exponentially with the problem sizes [12, § 1.4.2]. Even small problems, with a few tens of variables, can take a very long time. In the sequel, we apply Lagrangian duality to problem (6) to develop a novel solution method that is distributed and fast.

A. Dual problem

Let us first form the partial Lagrangian by *dualizing* the first constraints of problem (6). To do this we introduce multipliers $\boldsymbol{\lambda} = (\lambda_i)_{i\in\mathcal{N}}$ for the first set of inequality constraints. Thus, the partial Lagrangian is given by

$$
L\big(t,\mathbf{x},\boldsymbol{\lambda}\big) = t + \sum_{i\in\mathcal{N}} \lambda_i \bigg(\sum_{j\in\mathcal{M}_i} \beta_{ij}x_{ij} - t \bigg)
$$
$$
= t\bigg(1 - \sum_{i\in\mathcal{N}} \lambda_i\bigg) + \sum_{j\in\mathcal{M}} \sum_{i\in\mathcal{N}_j} \beta_{ij}\lambda_i x_{ij} \ , \tag{7}
$$

where the second equality follows from rearranging the terms and from the equivalence of the following two sets:[4]

$$\{(i,j)\,|\,i \in \mathcal{N}, j \in \mathcal{M}_i\} \equiv \{(n,m)\,|\,m \in \mathcal{M}, n \in \mathcal{N}_m\} \ . \tag{8}$$

Let $g(\boldsymbol{\lambda})$ denote the dual function obtained by minimizing the partial Lagrangian (7) with respect to t and $\mathbf{x}$. For notational simplicity, let us further denote by $\mathcal{X}$ the set of vectors $\mathbf{x}$ that satisfy the constraints (6c)-(6d) of problem (6). In particular, $\mathcal{X}$ can be expressed as a Cartesian product of some sets $\mathcal{X}_j \subset \mathbb{R}^{n_j}$, $j \in \mathcal{M}$, i.e.,

$$\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_M, \tag{9}$$

where $\mathcal{X}_j$ is given by

$$\mathcal{X}_j = \left\{ \mathbf{x}_j = (x_{ij})_{i\in\mathcal{N}_j} \,\middle|\, \sum_{i\in\mathcal{N}_j} x_{ij} = 1, \ x_{ij} \in \{0,1\}, \ i \in \mathcal{N}_j \right\}. \tag{10}$$

Thus, the dual function is

$$g(\boldsymbol{\lambda}) = \inf_{\substack{t\in\mathbb{R} \\ \mathbf{x}\in\mathcal{X}}} L\big(t,\mathbf{x},\boldsymbol{\lambda}\big) \tag{11a}$$

$$
= \begin{cases} \displaystyle\inf_{\mathbf{x}\in\mathcal{X}} \sum_{j\in\mathcal{M}} \sum_{i\in\mathcal{N}_j} \beta_{ij}\lambda_i x_{ij} & \sum_{i\in\mathcal{N}} \lambda_i = 1 \\ -\infty & \text{otherwise} \end{cases} \tag{11b}
$$

$$
= \begin{cases} \displaystyle\sum_{j\in\mathcal{M}} \inf_{\mathbf{x}_j\in\mathcal{X}_j} \bigg(\sum_{i\in\mathcal{N}_j} \beta_{ij}\lambda_i x_{ij} \bigg) & \sum_{i\in\mathcal{N}} \lambda_i = 1 \\ -\infty & \text{otherwise} \end{cases} \tag{11c}
$$

[4] This equivalence can be visualized by using a bipartite graph, where the nodes are the elements of two disjoint sets, the set of APs (i.e., $\mathcal{N}$) and the set of clients (i.e., $\mathcal{M}$), and the edges are the potential AP-client associations.

$$
= \begin{cases} \sum\limits_{j \in \mathcal{M}} g_j(\boldsymbol{\lambda}) & \sum\limits_{i \in \mathcal{N}} \lambda_i = 1 \\ -\infty & \text{otherwise} , \end{cases} \tag{11d}
$$

where the equality (11b) follows from that the linear function $t(1 - \sum_{i \in \mathcal{N}} \lambda_i)$ is bounded below only when it is identically zero, the equality (11c) follows from (9)-(10), and $g_j(\boldsymbol{\lambda})$ in (11d) is the optimal value of the problem

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i \in \mathcal{N}_j} \beta_{ij} \lambda_i x_{ij} \\
\text{subject to} \quad & \mathbf{x}_j \in \mathcal{X}_j ,
\end{aligned} \tag{12}
$$

with the variable $\mathbf{x}_j$. Even though problem (12) is combinatorial, it has a closed-form solution given by

$$
x_{ij}^\star = \begin{cases} 1 & i = \arg\min\limits_{n \in \mathcal{N}_j} \beta_{nj} \lambda_n \\ 0 & \text{otherwise} . \end{cases} \tag{13}
$$

and is computable very fast[5]. Thus, the Lagrange dual problem is given by

$$
\begin{aligned}
\text{maximize} \quad & g(\boldsymbol{\lambda}) = \sum_{j \in \mathcal{M}} g_j(\boldsymbol{\lambda}) \tag{14a} \\
\text{subject to} \quad & \sum_{i \in \mathcal{N}} \lambda_i = 1 \tag{14b} \\
& \lambda_i \geq 0, \ i \in \mathcal{N} , \tag{14c}
\end{aligned}
$$

where the variables is $\boldsymbol{\lambda}$. We denote by $d^\star$ the optimal value of the problem (14) that will be useful later. Note that the Lagrange dual problem (14) is a convex optimization problem, even though the primal problem (6) is *not* convex (see [12, § 5.2]). Let us next focus on the dual problem (14) and its solution method, which allows us to find a good feasible solution to the original problem (6).

B. Solving the dual problem via projected subgradient method

The objective function $g(\boldsymbol{\lambda})$ is, in general, a non-smooth (therefore non-differentiable) concave function. A common approach to handle such non-differentiable functions is the subgradient method [39], because gradient based algorithms cannot be applied. Therefore, the projected subgradient method [39], [40] is used to solve the dual problem (14).

First, we denote by $\mathbf{u}$ a subgradient of $-g$ at a feasible $\boldsymbol{\lambda}$, where $\mathbf{u} = (u_i)_{i \in \mathcal{N}}$. Specifically,

$$
u_i = -\sum_{j \in \mathcal{M}_i} \beta_{ij} x_{ij}^\star , \tag{15}
$$

where $x_{ij}^\star$ for all $j \in \mathcal{M}$ and $i \in \mathcal{N}_j$ is obtained as the solution of problem (12) for all $j \in \mathcal{M}$. Thus the projected subgradient method is given by

$$
\boldsymbol{\lambda}^{(k+1)} = P\big(\boldsymbol{\lambda}^{(k)} - \alpha_k \mathbf{u}^{(k)}\big) , \tag{16}
$$

where k is the current iteration index of the subgradient method, $\alpha_k > 0$ is the kth step size[6], and P is Euclidean projection onto the unit simplex $\Pi = \{\boldsymbol{\lambda} \,|\, \sum_{i \in \mathcal{N}} \lambda_i = 1, \lambda_i \geq 0\}$ (see [40, Excercise 2.1.12]). By employing (16) in an iterative manner, we can *solve* the dual problem (14). However, recovering a primal feasible solution is nontrivial because the original problem (6) is noncovex. A discussion of these nontrivial issues and how to find a good feasible solution is

---

[5]If $\mathcal{I} = \arg\min_{n \in \mathcal{N}_j} \beta_{nj}(\lambda_n + \mu_{nj})$ is not a singleton, then an arbitrary $i \in \mathcal{I}$ is chosen.

[6]We chose *square summable but not summable* step size (e.g., $\alpha_k = a/k$, where $0 < a < \infty$), that guarantees the asymptotic convergence of the subgradient method [39].

DAA: Distributed algorithm for client association

1. Initialization: At every AP i, the local channel utilizations, i.e., $(\beta_{ij})_{j \in \mathcal{M}_i}$ are given. Set subgradient iteration index $k = 1$. Each AP i broadcasts the initial *feasible* prices $\lambda_i^{(k)}$ to its local clients $j \in \mathcal{M}_i$.
2. Every client j sets $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(k)}$ and locally determines its association by solving problem (12). Denote by AP i_j the AP for which $x_{ij} = 1$.
3. Client j ($\in \mathcal{M}$) signals *only* to AP i_j an do not send any signalling to other AP i s, where $i \in \mathcal{N}_j \setminus \{i_j\}$.
4. Every AP i computes u_i by summing β_{ij} over the clients $j \in \mathcal{M}_i$, who had signalled in step 3, see (15).
5. Subgradient iteration: APs communicate and form $\mathbf{u}^{(k)}$ by combining each u_i and by performing (16) to compute $\boldsymbol{\lambda}^{(k+1)}$.
6. Stopping criterion: if the stopping criterion is satisfied, STOP. Otherwise, set $k = k + 1$, each AP i broadcasts the *feasible* prices $\lambda_i^{(k)}$ to its local clients $j \in \mathcal{M}_i$ and go to step 2.

---

deferred to § V-B, in order to maintain a cohesive presentation. Let us next describe how the computation of the solution of problem (14) is performed in a distributed manner.

C. Distributed algorithm for client association (DAA)

Recall that the dual function $g(\boldsymbol{\lambda})$ is separable among the clients $j \in \mathcal{M}$, see the objective function (14a) of problem (14). Therefore, the subgradient components (15) for the subgradient method (16) can be computed by coordinating the problem (12) for all $j \in \mathcal{M}$. This suggests the distributed algorithm for client association (*DAA*) presented at the top of this page.

The first step initializes *DAA*. Steps 2 represents the optimization performed in a decentralized fashion by each client for fixed $\boldsymbol{\lambda}$. The optimization at each client j is a very simple operation and is to find the AP i_j, where, $i_j = \arg\min_{i \in \mathcal{N}_j} \beta_{ij} \lambda_i$ (see (13)). Step 3 requires signaling between clients and APs. In particular, each client j signals only to AP i_j. This signalling process can be performed very efficiently, e.g., binary signaling. As a result, we have a light protocol between clients and APs. In Step 4, each AP i locally computes u_i (see 15), which is the summation of β_{ij} over the client j s who signalled the AP. Step 5 requires AP coordination. In particular, APs coordinate to perform (16), which is the projection of a point onto the unit simplex. The result of this operation is given by the solution to a convex optimization problem, which can be carried out efficiently. Step 6, is the stopping criterion for the algorithm. If the stopping criterion is satisfied, *DAA* terminates. Otherwise, the algorithm continues in an iterative manner. In practice, a natural stopping criterion for the algorithm would be running it for a fixed number of iterations.

D. Distributed implementation over existing standards

Let us discuss now how the actual implementation of the proposed *DAA* algorithm could be achieved on top of the existing standards, IEEE 802.15.3c and IEEE 802.11ad. The distributed algorithm is performed periodically in the system

to ensure the balanced operation of the network. The period of the execution is given by the control messages established by the medium access control protocol, as we see in detail below.

The iterative association algorithm does not have to be executed every time a client initiates an association or a handoff process. We assume that the *newcomer* client follows the association mechanism that IEEE 802.15.3c and IEEE 802.11ad define, based on the RSSI. Then, our algorithm is periodically executed to *correct* possible suboptimal client associations in the network by reallocating the available resources. As mentioned before the distributed nature of the association algorithm is crucial, in the direction of offloading the APs and make good use of the small computational resources that the clients may provide. Both 802.15.3c and 802.11ad define control frames (denoted as beacon frames) that are periodically broadcasted by the APs in the network. The APs can utilize these frames to trigger the initialization of *DAA* and carry the required information to the clients. The APs inform their clients about the initialization of *DAA* by setting a special bit into the beacon frame. Thus, the clients are ready to cooperate towards the optimal resource allocation in the network. The information required by the algorithm can be carried in the control frames or piggy-backed to the data frames that the APs send to the clients [38]. Moreover, the clients are piggy-backing the information in the data frames that they send to the APs. Thus, the algorithm is executed in perfect harmony with the networking protocols, without interrupting the actual network operation (data communication) and without causing extra delays. Therefore, we conclude that the proposed approach is well supported by the existing standards and does not require modifications to the way they operate.

V. ALGORITHM PROPERTIES

In this section, we first show the convergence performance for the proposed algorithm. Then we show how to recover a *good* primal feasible solution. Next, we highlight some sufficient conditions under which strong duality holds for the MILP (6) followed by a couple of examples. Finally, we show analytically the *asymptotic optimality* of the algorithm, where the *relative duality gap* diminishes to zero as the number of clients in the system grows.

Recall that $p^\star$ is the optimal value of the original MILP (6) and $d^\star$ is the optimal value of the associated dual problem (14). We refer to $p^\star$ as the *primal optimal value*, $d^\star$ as the *dual optimal value*, $(p^\star - d^\star)$ as the *optimal duality gap*, and $(p^\star - d^\star)/p^\star$ as the *optimal relative duality gap*, which are useful in the rest of the paper.

A. Convergence

DAA essentially solves the dual problem (14) by using the projected subgradients method and the convergence of the algorithm is established by the following proposition:

Proposition 1: Let $g_{\text{best}}^{(k)}$ denote the *best* dual objective value found after k subgradient iterations, i.e., $g_{\text{best}}^{(k)} = \max\{g(\boldsymbol{\lambda}^{(1)}), \ldots, g(\boldsymbol{\lambda}^{(k)})\}$. Then, $\forall \epsilon > 0 \ \exists n \geq 1$ such that $\forall k \ k \geq n \Rightarrow \left(d^\star - g_{\text{best}}^{(k)}\right) < \epsilon$.

Proof: The proof is built on the material presented in [39, § 3.2] [40]. Let us denote by $\boldsymbol{\lambda}^\star$ the optimal solution of dual problem (14). Thus, we have

$$||\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^\star||_2^2 = ||P\big(\boldsymbol{\lambda}^{(k)} - \alpha_k \mathbf{u}^{(k)}\big) - \boldsymbol{\lambda}^\star||_2^2 \quad (17\text{a})$$

$$\leq ||\big(\boldsymbol{\lambda}^{(k)} - \alpha_k \mathbf{u}^{(k)}\big) - \boldsymbol{\lambda}^\star||_2^2 \quad (17\text{b})$$

$$= ||\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^\star||_2^2 - 2\alpha_k \mathbf{u}^{(k)\text{T}}(\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^\star)$$
$$+ \alpha_k^2 ||\mathbf{u}^{(k)}||_2^2 \quad (17\text{c})$$

$$\leq ||\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^\star||_2^2 - 2\alpha_k(g(\boldsymbol{\lambda}^\star) - g(\boldsymbol{\lambda}^{(k)}))$$
$$+ \alpha_k^2 ||\mathbf{u}^{(k)}||_2^2 \quad (17\text{d})$$

$$= ||\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^\star||_2^2 - 2\alpha_k(d^\star - g(\boldsymbol{\lambda}^{(k)}))$$
$$+ \alpha_k^2 ||\mathbf{u}^{(k)}||_2^2 \, , \quad (17\text{e})$$

where (17a) follows from (16), (17b) follows from that the projection onto unit simplex Π always decrease the distance of a point to every point in Π and in particular to the optimal point $\boldsymbol{\lambda}^\star$, (17d) follows from the definition of subgradient, i.e., $-g(\boldsymbol{\lambda}^\star) \geq -g(\boldsymbol{\lambda}^{(k)}) + \mathbf{u}^{(k)\text{T}}(\boldsymbol{\lambda}^\star - \boldsymbol{\lambda}^{(k)})$, and (17e) follows from that $d^\star = g(\boldsymbol{\lambda}^\star)$. Recursively applying (17e) and rearranging the terms, we get

$$2\sum_{l=1}^{k} \alpha_l(d^\star - g(\boldsymbol{\lambda}^{(l)})) = -||\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^\star||_2^2 + ||\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^\star||_2^2$$
$$+ \sum_{l=1}^{k} \alpha_l^2 ||\mathbf{u}^{(k)}||_2^2 \quad (18\text{a})$$

$$\leq R^2 + G^2 \sum_{l=1}^{k} \alpha_l^2 \, , \quad (18\text{b})$$

where (18b) follows from that $||\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^\star||_2 \geq 0$, $||\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^\star||_2 \leq R = \sqrt{2}$ for any $\tilde{\boldsymbol{\lambda}} \in \Pi$, and the norm of any subgradient $\mathbf{u}$ of $-g$ at any $\tilde{\boldsymbol{\lambda}} \in \Pi$ is bounded, i.e.,[7]

$$||\mathbf{u}||_2 \leq G = \sqrt{\sum_{i \in \mathcal{N}} \big(\sum_{j \in \mathcal{M}_i} \beta_{ij}\big)^2}. \quad (19)$$

Moreover, clearly we have

$$d^\star - g_{\text{best}}^{(k)} \leq d^\star - g(\boldsymbol{\lambda}^{(l)}), \ l = 1, \ldots, k \, . \quad (20)$$

Thus, from (18b), (20), and by noting that step size $\alpha_l = a/l$, $0 < a < \infty$ is *square summable* $\big(\text{i.e., } \sum_{l=1}^{\infty} \alpha_l^2 = a^2\pi/6\big)$, we obtain an upper bound on $d^\star - g_{\text{best}}^{(k)}$ as

$$d^\star - g_{\text{best}}^{(k)} \leq (R^2 + G^2 \sum_{l=1}^{k} \alpha_l^2)/(2\sum_{l=1}^{k} \alpha_l) \quad (21\text{a})$$

$$\leq \big(R^2/2 + a^2 G^2 \pi/12\big)/\big(\sum_{l=1}^{k} \alpha_l\big) \, . \quad (21\text{b})$$

Since the denominator $\sum_{l=1}^{k} \alpha_l$ is strictly monotonically increasing in k (it grows without bound as $k \to \infty$), for any $\epsilon > 0$ we can always find a integer $n \geq 1$ such that $\sum_{l=1}^{k} \alpha_l > \epsilon \, (R^2/2 + a^2 G^2 \pi/12)$ if $k \geq n$, which concludes the proof. ∎

The bound derived in (21b) allows us to predict some key behaviors of the convergence of the proposed algorithm. To see this, we note from (19) that the numerator of the bound depends on $(\beta_{ij})_{i \in \mathcal{N}, \ j \in \mathcal{M}_i}$ for fixed a. Now suppose that the number of clients increases. This forces G to increase as well. As a result, the corresponding total iterations to reach the given accuracy ϵ will also grow. Roughly speaking, this means that, for fixed number of APs, the larger the number of clients, the larger the total number of iterations required for the convergence of *DAA*. On the other hand, suppose that the

---

[7]Compare to (15).

user distribution is such that $\sum_{j\in\mathcal{M}_i}\beta_{ij}$ is roughly the same for each AP i. Thus, if the total number of APs is increased, then G will become larger and as a result, the total number of iterations to convergence is increased as well. These algorithm behaviors are numerically illustrated in § VI.

B. Recovering a feasible primal point

As we discussed in § V-A, we can *solve* the dual problem to any given accuracy to yield the dual optimal value $d^\star$ and the dual optimal solution $\boldsymbol{\lambda}^\star$. If the primal problem is convex, from $d^\star$ and $\boldsymbol{\lambda}^\star$, we can usually obtain the primal optimal value $p^\star$ and primal optimal solution $(t^\star, \mathbf{x}^\star)$ [12, § 5.5.5]. However, recall that the original MILP (6) is a *nonconvex* problem. Therefore, unlike convex problems, there is *no guarantee* that from $d^\star$ and $\boldsymbol{\lambda}^\star$, we obtain $p^\star$ and $(t^\star, \mathbf{x}^\star)$.[8] Nevertheless, in the case of MILP (6), a *primal feasible point* is obtained during each iteration k of the algorithm (see step 2 of the algorithm). Thus, it is natural to go for the best choice, among all the primal feasible points obtained so far. For example, a good approximation for the primal optimal value would be

$$p_{\text{best}}^{(k)} = \min\{t^{(1)}, \ldots, t^{(k)}\}, \tag{22}$$

where $(t^{(k)}, \mathbf{x}^{(k)})$ is the primal feasible point in iteration k.[9] A good feasible point would be $(t_{\text{best}}^{(k)}, \mathbf{x}_{\text{best}}^{(k)})$, which is the primal feasible point that corresponds to $p_{\text{best}}^{(k)}$.

Even though the value $p_{\text{best}}^{(k)}$ is not usually as good as the primal optimal value $p^\star$, monte Carlo simulations show that it is a good approximate value with k on the order of hundreds or more, e.g., $k \geq 100$ (see § VI). There is no clear analytical explanations of these fortuitous encounters, especially because the original problem (6) is nonconvex [40, § 6.3].

C. Duality gap

The *duality gap* $(p^\star - d^\star)$ is one of the important metric that can be used to quantify the performance of the proposed *DAA*. Note that, in general, the duality gap for MILP (6) is not zero, because the problem is nonconvex. Therefore, it is not surprising that deriving general conditions under which the strong duality for MILP (6) is very difficult. Nevertheless, we first provide some examples to highlight sufficient conditions for strong duality for MILP (6). The latter part of this section derives an analytical bound on the duality gap. Moreover, we show the *asymptotic optimality* of the algorithm, where the *relative duality gap* $(p^\star - d^\star)/p^\star$ diminishes to zero as the number of clients in the system grows. Such asymptotic results are indeed important from a theoretical, as well as from a practical perspective, see for example the duality results associated with the well known Knapsack problem [40, Excercise 5.1.7].

The following proposition establishes a simple result, which is instrumental to study zero duality.

---

[8]The first component $t^\star$ of the primal optimal solution of MILP (6) and the primal optimal value $p^\star$ are clearly the same.

[9]APs can compute $(t^{(k)}, \mathbf{x}^{(k)})$ easily by using the client signaling they received at step 2 and the AP coordination at step 5. For example, $t^{(k)} = ||-\mathbf{u}^{(k)}||_\infty$.

Proposition 2: Let $p_{\text{relax}}^\star$ denote the optimal value of the linear programming (LP) relaxation of problem (6), i.e.,

$$\text{minimize} \quad t \tag{23a}$$
$$\text{subject to} \quad \sum_{j\in\mathcal{M}_i}\beta_{ij}\, x_{ij} \leq t,\ i\in\mathcal{N} \tag{23b}$$
$$\sum_{i\in\mathcal{N}_j} x_{ij} = 1,\ j\in\mathcal{M} \tag{23c}$$
$$0 \leq x_{ij} \leq 1,\ j\in\mathcal{M}, i\in\mathcal{N}_j, \tag{23d}$$

with variables t and $\mathbf{x} = (x_{ij})_{i\in\mathcal{N},\ j\in\mathcal{M}_i}$. Then $d^\star = p_{\text{relax}}^\star$.

Proof: Note that problem (6) is always feasible. The proof is based on two key results:

(a) for problem (23), we always have strong duality, i.e., $p_{\text{ralax}}^\star = d_{\text{ralax}}^\star$, and

(b) the dual of problem (23) is identical to problem (14), and therefore $d_{\text{ralax}}^\star = d^\star$, where $d_{\text{ralax}}^\star$ denotes the dual optimal value of problem (23).

In particular, the item (a) above is guaranteed from strong duality results for linear programs, see [12, § 5.2.3]. To show the part (b) mentioned above, we first note the following: the partial Lagrangian obtained by dualizing first constraint of problem (23) is identical to (7), where we consider same notations for the dual variables. This is a slight abuse of notation, but it helps in the clarity of the exposition. Let $h(\boldsymbol{\lambda})$ denote the dual function obtained by minimizing the partial Lagrangian as given by [compare with (11)]

$$h(\boldsymbol{\lambda}) = \begin{cases} \sum_{j\in\mathcal{M}} h_j(\boldsymbol{\lambda}) & \sum_{i\in\mathcal{N}}\lambda_i = 1 \\ -\infty & \text{otherwise}, \end{cases} \tag{24}$$

where $h_j(\boldsymbol{\lambda})$ is the optimal solution of a problem very similar to (12), except that the constraint $\mathbf{x}_j \in \mathcal{X}_j$ is replaced by $\mathbf{x}_j \in \text{conv}(\mathcal{X}_j)$. Note that $\text{conv}(\mathcal{X}_j) \in \mathbb{R}^{n_j}$ is a *unit simplex*, and therefore the optimal value $h_j(\boldsymbol{\lambda})$ is attained at one of the vertexes $\mathbf{x}_j = (x_{ij})_{i\in\mathcal{N}_j}$ of the unit simplex [41, Corollary 32.3.4]. Specifically, the components of the vertex $\mathbf{x}_j$ is identically given by (13). As a result, $h_j(\boldsymbol{\lambda}) = g_j(\boldsymbol{\lambda})$ for all $j \in \mathcal{M}$ and $h(\boldsymbol{\lambda}) = g(\boldsymbol{\lambda})$. ∎

Note that if problem (23), the LP relaxation of problem (6), has integer solutions, then we can easily show that the optimal value $p^\star$ of the original MILP (6) is equal to the optimal value $p_{\text{relax}}^\star$, i.e., $p^\star = p_{\text{relax}}^\star$. Therefore, from *Proposition 2*, we have $p^\star = d^\star$. In other words, if problem (23) has integer solutions, then *strong duality* holds for the original MILP (6). Thus, it is natural to investigate the conditions, under which the LP (23) has integer solutions. Roughly speaking, there are not many results that establish conditions on LPs, beyond *total unimodularity* [42, § 9] of the associated problem matrices, under which they have integer solutions. Unfortunately, particularized to LP (23), the related matrices are not total unimodular, and therefore the theoretical implications of total unimodularity does not applies [42, § 9]. Nevertheless, *Proposition 2* allows us to imagine special cases of MILP (6), where we have strong duality. We present two such examples in the following.

Example 1: Suppose that the initial clients and APs communication is as shown in Figure 2(a). In particular, there are M clients and $N = M$ APs. Each client $j \in \{2, 3, \ldots, M\}$ can be associated to either AP $j-1$ or j. However, the first client can be associated only to the AP 1. Moreover, suppose that $\beta_{jj} = \beta \in (0, 1]$ for all $j \in \{1, \ldots, M\}$ and the remaining β_{ij}s can be arbitrary values from the range $(0, 1]$.
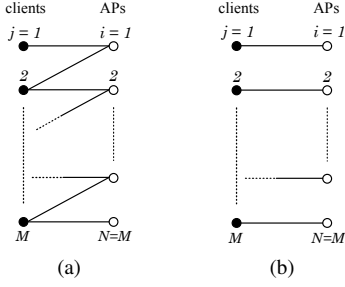
Figure 2. Examples of network topology 1 (a) The initial communication graph; (b) The optimal association
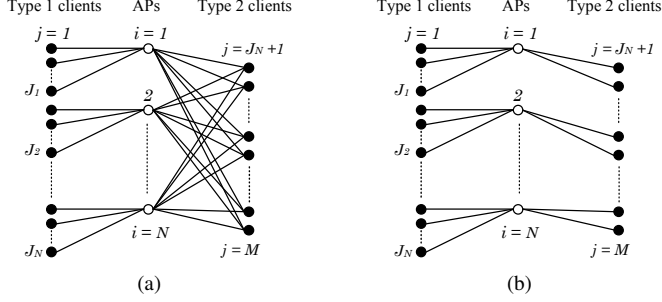

Figure 3. Examples of network topology 2 (a) The initial communication graph; (b) The optimal association

We can easily show that the optimal client association (i.e., the optimal solution of MILP (6)) corresponds to Figure 2(b). Moreover, we have $p^\star = \beta$. Let us now focus to the LP relaxation (23) applied to the network in Figure 2(a). In this case, every client j, except the client 1, can be associated to both AP $j-1$ *and* AP j. However, we can prove by contradiction: the solution of problem (23) corresponds to the optimal association depicted in Figure 2(b). Thus, we have $p^\star = p^\star_{\text{relax}}$. From *Proposition 2*, we have strong duality for MILP (6), (i.e., $p^\star = d^\star$) and the proposed algorithm achieves $d^\star$.

Example 2: Consider the network shown in Figure 3(a). There are N APs and two types of clients connected to APs. The Type 1 clients can communicate only with a *single* AP. For example each client $j \in \{1, \ldots, J_1\}$ can communicate only with AP 1, and therefore, they must be assigned to AP 1. On the other hand, the Type 2 clients can communicate with all the APs, e.g, clients $J_N+1, \ldots, M$. As a result, Type 2 clients can be associated to any AP. Now suppose that β_{ij} values associated with Type 1 clients are such that $\sum_{j=1}^{J_1} \beta_{1j} = \sum_{j=J_1+1}^{J_2} \beta_{2j} \cdots = \sum_{j=J_{N-1}+1}^{J_N} \beta_{Nj} = B \in \mathbb{R}_+$ and β_{ij} values associated with Type 2 clients are all equal to $\beta \in (0, 1]$. Moreover, suppose that the number of Type 2 clients is a multiple of N, i.e., $M - J_N = mN$ for some $m \in \mathbb{Z}_+$.

The optimal client association (i.e., the optimal solution of MILP (6)) corresponds to Figure 3(b), where the Type 2 clients are equally distributed among the APs. In particular, each AP is associated with m Type 2 clients and we have $p^\star = B + m\beta$. Let us now consider solution given by the LP relaxation (23). note that there is no choice for Type 1 clients, other than associating them to the only AP they can communicate. From the symmetry, we can easily see that associating each Type 2 client j among all the APs $i = 1, \ldots, N$ with equal shares is a particular solution to the LP relaxation (23), i.e., for every Type 2 client j, $x_{ij} = (1/N)$, $i = 1, \ldots, N$. Thus, at each AP i, the utilization corresponds to Type 1 clients is B and the utilization corresponds to Type 2 clients becomes $(M - J_N)(1/N)$, which is identical to m (recall $M - J_N = mN$). Therefore, we have $p^\star_{\text{relax}} = (B + m) = p^\star$ and strong duality holds for MILP (6).

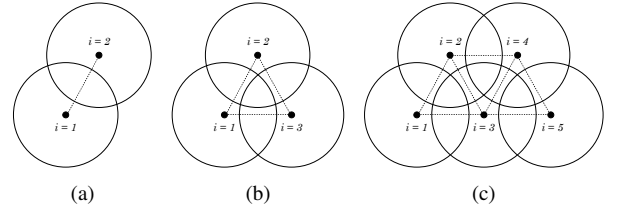The examples above give insights into the algorithm's

Figure 4. Example simulation topologies (a) 2APs; (b) 3APs; (c) 5APs

behavior in spacial cases. However, they can be used to build intuitive ideas of general networks. It is indeed important to analyze the proposed algorithm's behavior in general as well. In the sequel, we provide theoretical substantiation that allows us to predict the general algorithm properties in terms of the optimal duality gap and the relative duality gap.

The following theorem formally establishes a bound on the duality gap and the asymptotic optimality of *DAA*.

Theorem 1: The optimal duality gap of mixed integer linear program (6) is bounded as follows:

$$p^\star - d^\star \le (N+1)(\varrho + \max_{j \in \mathcal{M}} \varrho_j) , \qquad (25)$$

where $\varrho = \max_{i \in \mathcal{N}, j \in \mathcal{M}_i} \beta_{ij}$ and $\varrho_j = \min_{i \in \mathcal{N}_j} \beta_{ij}$. Moreover, the *relative* duality gap $(p^\star - d^\star)/p^\star$ diminishes to 0 as $M \to \infty$.

Proof: See Appendix A. ∎

The theorem above suggests that the duality gap is always bounded by a *constant* that does not depend on the number of clients in the system. Note that the bound grows like N, and therefore we can expect an increase in the duality gap for larger N. The theorem states that for larger M, the relative duality gap become almost zero. These are very important to get insight into the behavior of the proposed algorithm in general networks, see § VI for numerical examples.

VI. NUMERICAL EXAMPLES

In this section we present the numerical evaluation of the proposed algorithm in a multi-user multi-cell environment. We compare *DAA* with: a) random association policy, b) RSSI-based policy, which is the standard association mechanism used in 802.11, 802.15.3 and 802.11ad networks, and c) optimal policy given by solving the optimization problem (6) by using IBM CPLEX optimizer [43].

We define the SNR operating point at a distance d [distance units] form any AP as

$$\text{SNR}(d) = \begin{cases} P_0 \lambda^2/(16\pi^2 N_0 W) & d \le d_0 \\ P_0 \lambda^2/(16\pi^2 N_0 W) \cdot (d/d_0)^{-\eta} & \text{otherwise} . \end{cases}$$

Circular cells as depicted in Figure 1 are considered, where the radius of each cell r is chosen such that $\text{SNR}(r) = 10$ dB. The APs are located such that the distance between any consecutive APs is $D = 1.1r$. For example, Figure 4(a) shows the case for $N = 2$ APs, Figure 4(b) shows the case for $N = 3$, and Figure 4(c) shows the case for $N = 5$. The clients are uniformly distributed among the circular cells, and the potential AP-client association (i.e., $\mathcal{M}_i$ and $\mathcal{N}_j$) is found as pointed out in Figure 1.

We set $\lambda = 5$ mm, $N_0 = -134$ dBm/MHz, $W = 1200$ MHz, and $d_0 = 1$ m, see (1) and (2). Moreover, for all $j \in \mathcal{M}$, we set $I_j = 0$ and for all $i \in \mathcal{N}, j \in \mathcal{M}_i$, we set
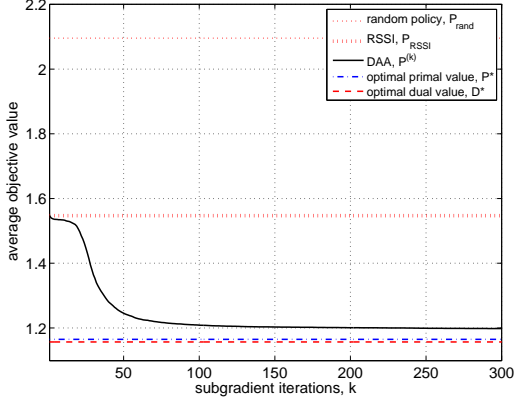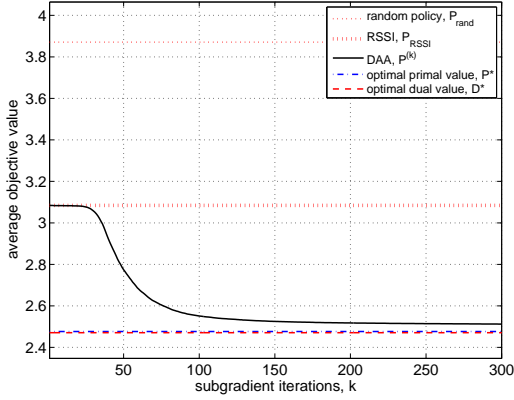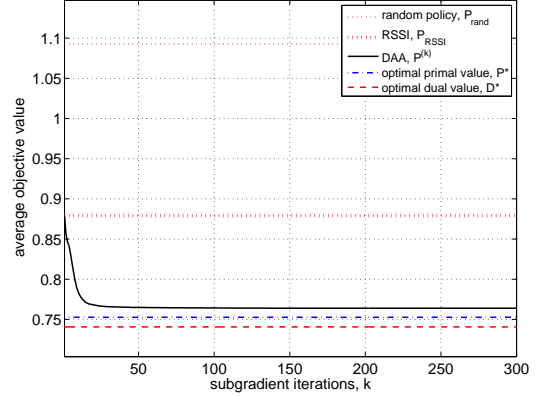
Figure 5. Influence of the number of clients on the convergence (a) Average objective value $\mathrm{P}^{(k)}$ vs. iterations k, 5 APs, 100 clients; (b) Average objective value $\mathrm{P}^{(k)}$ vs. iterations k, 5 APs, 200 clients

Figure 6. Influence of the number of APs on the convergence. (a) Average objective value $\mathrm{P}^{(k)}$ vs. iterations k, 3 APs, 30 clients; (b) Average objective value $\mathrm{P}^{(k)}$ vs. iterations k, 10 APs, 100 clients

$P_{ij} = P_0 = 0.1$ mW and $G_{ij}^{\mathrm{Tx}} = G_{ij}^{\mathrm{Rx}} = 1$. In order to check the average performance of the algorithms, we consider $\bar{T} = 1000$ time slots, where the fading coefficients α_{ij} for all $i \in \mathcal{N}$, $j \in \mathcal{M}_i$ and the demanded data rates Q_j for all $j \in \mathcal{M}$ are constant during each time slot $T \in \{1, \ldots, \bar{T}\}$ and independently change from one slot to another. In particular, the exponential random variables α_{ij}'s with unit mean are independent and identically distributed over the time slots. Moreover, we assume that Q_js are uniformly distributed on $[0, 400]$Mbits/s and independent and identically distributed over the time slots.
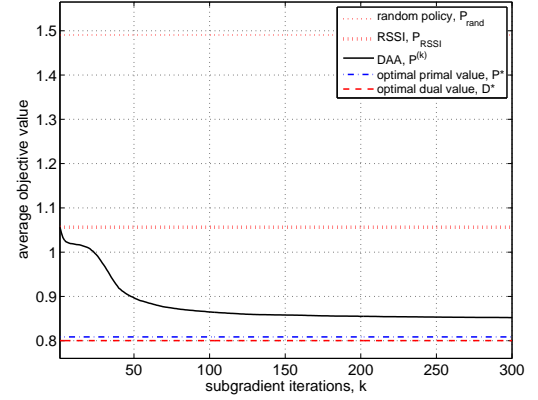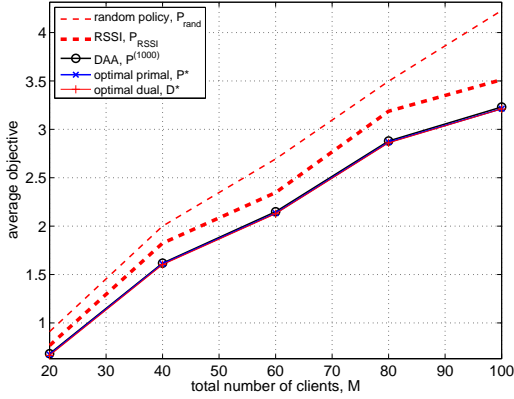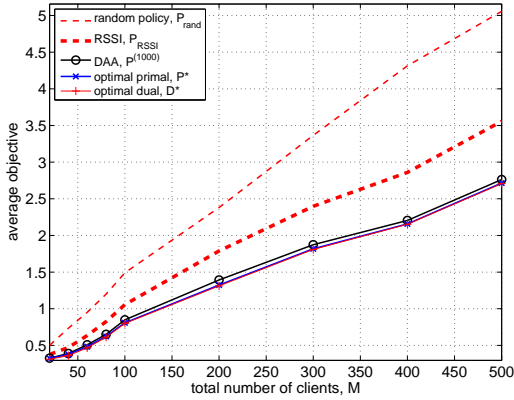
To see the average convergence behavior of the proposed algorithm, we consider the average *primal* objective value of problem (6) obtained by *DAA*. In particular, the average primal objective value from *DAA* after k subgradient iterations is defined as $\mathrm{P}^{(k)} = (1/\bar{T}) \sum_{T=1}^{\bar{T}} p_{\mathrm{best}}^{(k)}(T)$, where $p_{\mathrm{best}}^{(k)}(T)$ is the best *primal feasible* objective value of problem (6) after k iterations at time slot T, see (22).[10] The average objective values from benchmark algorithms, random association policy, RSSI policy, and the optimal policy are defined in a similar manner, and are denoted by $\mathrm{P}_{\mathrm{rand}}$, $\mathrm{P}_{\mathrm{RSSI}}$, and $\mathrm{P}^\star$, respectively.[11] Moreover, the average dual optimal value obtained

by *DAA* is defined as $\mathrm{D}^\star = (1/\bar{T}) \sum_{T=1}^{\bar{T}} d^\star(T)$, where $d^\star(T)$ is the optimal objective value of dual problem (14) at time slot T.

Figure 5 shows $\mathrm{P}^{(k)}$ versus subgradient iteration k, for the cases $N = 5$, $M = 100$ (Figure 5(a)) and $N = 5$, $M = 200$ (Figure 5(b)). Results show that there is a noticeable effect of varying M (the number of clients in the system) on convergence time. In particular, the convergence is faster for smaller M. This observation is consistent with our analytical study presented in § V-A. Proposed *DAA* clearly outperforms the RSSI policy used in 802.11, 802.15.3c and 802.11ad, as well as the random policy. For example, *DAA*, yields a performance improvement of about 20% compared to RSSI policy in both considered cases. The gap between $\mathrm{P}^\star$ and $\mathrm{P}^{(k)}$, even after relatively larger the number of subgradient iterations (e.g., $k = 300$) is indeed expected due to the nonconvexity of the original problem (5), see § V-B. Nevertheless, average dual optimal value $\mathrm{D}^\star$ from *DAA* is almost equal to the optimal $\mathrm{P}^\star$.

Figure 6 shows $\mathrm{P}^{(k)}$ versus subgradient iteration k, for the cases $N = 3$, $M = 30$ (Figure 6(a)) and $N = 10$, $M = 100$ (Figure 6(b)). Here, the clients density or the number of clients per AP is roughly the same (i.e., 10). Results show that there is a clear effect of varying N (while keeping client density fixed), on convergence. In particular, the convergence is faster for smaller N. This observation also agrees with our analytical study presented in § V-A. The performance of other benchmark
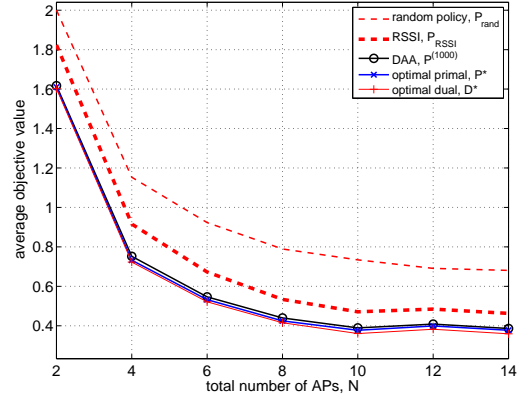
---

[10]Due to the nonconvexity of the original problem (5), $p_{\mathrm{best}}^{(k)}(T)$ is not necessarily achieves the optimal value even when $k \to \infty$.

[11]Since the benchmark algorithms do not depend on subgradient iteration k, like $\mathrm{P}^{(k)}$, there is no superscript (k) required for $\mathrm{P}_{\mathrm{rand}}$, $\mathrm{P}_{\mathrm{RSSI}}$, and $\mathrm{P}^\star$.

Figure 7. Average objective values $\mathtt{P_{rand}}, \mathtt{P_{RSSI}}, \mathtt{P}^\star, \mathtt{P}^{(1000)}$, and average dual optimal $\mathtt{D}^\star$ vs. the number of users M. (a) 2 APs; (b) 10 APs

Figure 8. Average objective value $\mathtt{P_{rand}}, \mathtt{P_{RSSI}}, \mathtt{P}^\star, \mathtt{P}^{(1000)}$, and average dual optimal $\mathtt{D}^\star$ vs. the number of APs N. (a) 40 users; (b) 100 users

algorithms are very similar to those for Figure 5.

Figure 7 shows the average objective from *DAA* after $K = 1000$ subgradient iterations, $\mathtt{P}^{(K)}$ versus the number of clients M for the cases $N = 2$ (Figure 7(a)) and $N = 10$ (Figure 7(b)). Plots for the benchmark algorithms are also depicted. Results show that the average objective values associated with each algorithms increase as M increases. This is intuitively expected because APs become more loaded as the number of clients in the network grows. Results further show that $\mathtt{P}^{(K)}$ from *DAA* is very close to the optimal $\mathtt{P}^\star$ in both cases and the performance gap is not sensitive to changes in M. Note that the performance of the random and the RSSI policies are substantially low and their performance degradation become even noticeable for larger M, see Figure 7(b). As expected $\mathtt{D}^\star$ provides a global lower bound on the performance, and is hardly distinguishable from $\mathtt{P}^\star$.

Figure 8 shows the average objective value versus the number of APs N for the cases $M = 40$ (Figure 8(a)) and $M = 100$ (Figure 8(b)). The performance ranking of the algorithms is very similar to that of previous plots Figure 7. Results show that the average objective values from the algorithms decrease as M increases. This is intuitively explained by noting that, the larger the N, the smaller the client density, and therefore the smaller the average objective values of each algorithm. Results again shows that the proposed *DAA* performs very close to the optimal and outperforms the random and RSSI policies noticeably.

To see the effect of increasing clients M on the *relative duality gap* (see § V-C), we define the metric *average relative duality gap*, $\mathtt{Ave\text{-}RDG}$. In particular, $\mathtt{Ave\text{-}RDG} = (1/\bar{T}) \sum_{T=1}^{\bar{T}} (p^\star(T) - d^\star(T))/p^\star(T)$, where $p^\star(T)$ is the optimal value of primal problem (6) and $d^\star(T)$ is the optimal value of dual problem (14), at time slot T. Moreover, we denote by $\mathtt{Ave\text{-}RDG\text{-}best\text{-}achieved}$, a related metric defined very similar to $\mathtt{Ave\text{-}RDG}$, except that $p^\star(T)$ is replaced with $p^{(K)}_{\text{best}}(T)$, i.e., the best *primal feasible* objective value achieved from *DAA* after K iterations at time slot T.

Figure 9(a) shows percentage $\mathtt{Ave\text{-}RDG}$ versus M for different Ns. For all considered $\mathtt{Ave\text{-}RDG}$, N approaches zero as M is increased. This is consistent with our analytical results established by *Theorem 1*, see § V-C. Figure 9(b) shows that the plots of percentage $\mathtt{Ave\text{-}RDG\text{-}best\text{-}achieved}$ versus M are very similar to those in Figure 9(a). This behavior is not surprising, because the performance of *DAA* is very close to the performance of optimal method, see Figure 7 and Figure 8.

Figures 10 depicts the dependence of the average duality gap on M. In particular, we define the *optimal average duality gap* $\mathtt{Ave\text{-}DG}$ as $\mathtt{P}^\star - \mathtt{D}^\star$ and plots of $\mathtt{Ave\text{-}DG}$ versus M are shown in Figures 10(a). Moreover, we define the *best achieved average duality gap* $\mathtt{Ave\text{-}DG\text{-}best\text{-}achieved}$ as $\mathtt{P}^{(K)} - \mathtt{D}^\star$ and the corresponding plots are shown in Figures 10(b). In both cases, results show that there is no apparent effect of changing M on duality gaps. Nevertheless, as we discussed in § V-C (see (25)), the average duality gap
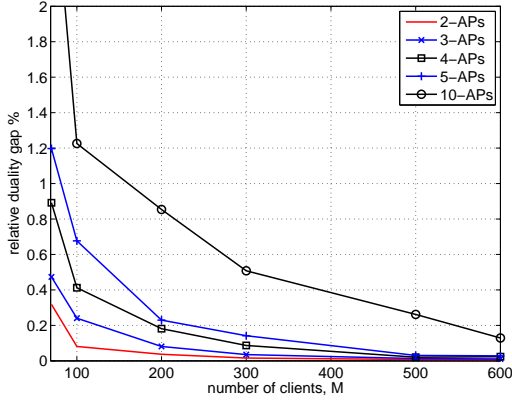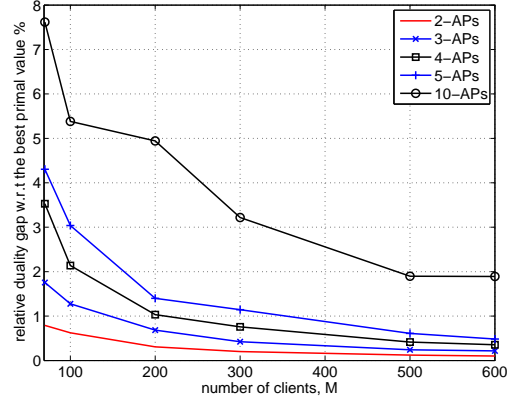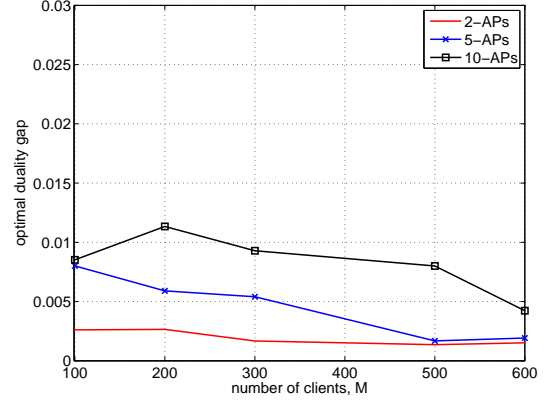
Figure 9. (a) Average relative duality gap `Ave-RDG` vs. the number of clients M; (b) Average best achieved relative duality gap `Ave-RDG-best-achieved` vs. the number of clients M

Figure 10. (a) Average optimal duality gap `Ave-DG` vs. the number of clients M; (b) Best achieved average duality gap `Ave-DG-best-achieved` vs. the number of clients M
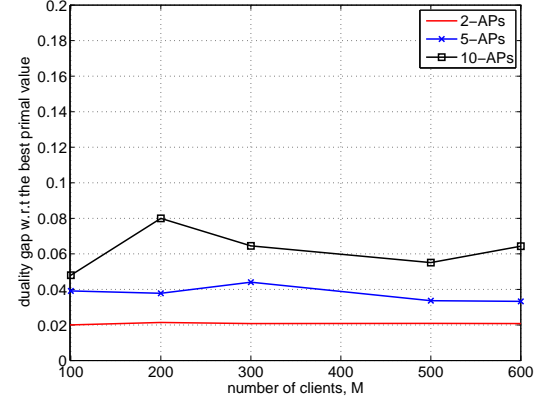
grows when N increases.

To examine the *fairness* of the final client association among the APs, we consider the well known Jain's fairness index [44] as the fairness metric. We denote by $J^{(k)}(T)$ the fairness resulted by the proposed *DAA* at time slot T and after k iterations. In particular, we define it as $J^{(k)}(T) = \left(\sum_{i \in \mathcal{N}} Y_i^{(k)}(T) \right)^2 / (N \sum_{i \in \mathcal{N}} Y_i^{(k)}(T)^2)$, where $Y_i^{(k)}(T) = \sum_{j \in \mathcal{M}_i} \beta_{ij} x_{ij}^{(k)}(T)$ with $x_{ij}^{(k)}(T)$ being the best feasible solution from *DAA* at time slot T and after k iterations. The average fairness index $J^{(k)}$ from *DAA* after k iterations is simply defined as $J^{(k)} = (1/\bar{T}) \sum_{T=1}^{\bar{T}} J^{(k)}(T)$. The average fairness indexes from benchmark algorithms, random association policy, RSSI policy, and the optimal policy are defined in a similar manner, and are denoted by $\mathtt{J}_{\text{rand}}$, $\mathtt{J}_{\text{RSSI}}$, and $\mathtt{J}^\star$, respectively.[12]

Figure 11 shows $J^{(k)}$ versus k together with the benchmark fairness indexes $\mathtt{J}_{\text{rand}}$, $\mathtt{J}_{\text{RSSI}}$, and $\mathtt{J}^\star$ for the case $N = 5$ and $M = 100$. Note that the fairness index ranges from $1/N$ (worst performance) to 1 (best performance). As expected, the optimal association provides the best performance. Results show that within a few hundreds of iterations, *DAA* achieves a fairness very close to the optimal, e.g., $J^{(500)}$ is 99.3% of $\mathtt{J}^\star$. Results further show that *DAA* significantly outperforms the random and the RSSI policies. For example, there is a 8.3%

---

[12]Benchmark algorithms do not depend on subgradient iteration k. Therefore, like $\mathtt{J}^{(k)}$, there is no superscript (k) required for $\mathtt{J}_{\text{rand}}$, $\mathtt{J}_{\text{RSSI}}$, and $\mathtt{J}^\star$.

performance gain from *DAA* compared to RSSI.

In order to provide a statistical description for the speed of the proposed algorithm, we consider empirical cumulative distribution function (CDF) plots. Specifically, for each time slot $T \in \{1, \ldots, \bar{T}\}$, we store the total CPU time required to find $p_{\text{best}}^{(K)}(T)$ from proposed *DAA*. For comparison, we use the total CPU time required to find the *optimal* value $p^\star(T)$. Figure 12(a) shows the empirical CDF plots of the number of iterations for $M = 100, 200, 300$, with $N = 10$. In the case of *DAA*, the effect of changing the problem size by increasing M on the CDF plots are almost indistinguishable. However, in the case of optimal method, there is a prominent increase in the time required to compute $p^\star(T)$. Figure 12(b) depicts the average time required by *DAA* and the optimal method versus M. Results show that the average time required by *DAA* to find possibly a suboptimal solution is not sensitive to the changes in M, and is almost zero. However, the average time required by the optimal method to find the optimal solution grows approximately exponentially with M. This is certainly expected because problem (6) is combinatorial, and therefore the worst-case complexity of the global method (CPLEX) grows exponentially with the problem size [12, § 1.4.2]. Thus, there is naturally a tradeoff between the optimality and the efficiency of the algorithms. Nevertheless, Figure 12 together with Figure 7, Figure 8 and the asymptotic results depicted in Figure 9 shows that the proposed *DAA* yields a good tradeoff
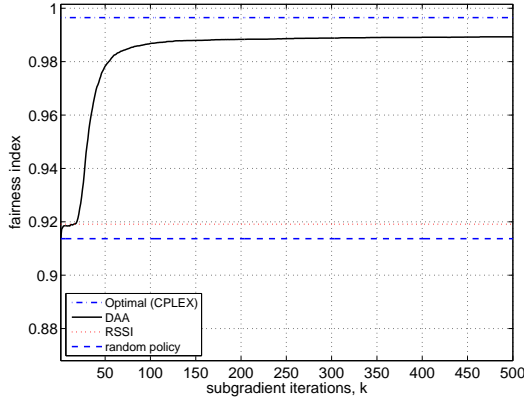
Figure 11. Average fairness index vs. the iterations k, 5 APs, 100 clients

between the optimality and the efficiency, which are favorable for the practical implementation.

VII. CONCLUSIONS

In this paper we considered the problem of optimizing the allocation of the clients to the available APs in 60 GHz wireless access networks. The objective in our problem formulation was to minimize the maximum AP utilization in the network. The optimization problem was combinatorial. Thus, we proposed a distributed association algorithm (*DAA*), based on Langragian duality theory and subgradient methods. *DAA* is fully compliant with the existing WIFI and 60 GHz protocols/standards (such as 802.11, 802.15.3 and 802.11ad) and it can be easily implemented on top of the MAC mechanisms that they define. We studied the behavior of *DAA* through theoretical analysis, where we proved its asymptotic optimality properties. Moreover, we presented a numerical analysis, where *DAA* was compared to other association policies in realistic scenarios. We tested convergence, scalability, time efficiency, and fairness. Our results indicate that the proposed solution could be well applied in the forthcoming 60 GHz communication networks.

APPENDIX A
PROOF OF THEOREM 1

The proof is based on an important proposition from [45], which we restate here for clarity and for simplifying the presentation.

Proposition 3: Consider the following possibly *nonconvex* problem:

$$\text{minimize} \quad \sum_{j\in\mathcal{J}} f_j(\mathbf{y}_j) \tag{26a}$$
$$\text{subject to} \quad \mathbf{y}_j \in \mathcal{Y}_j \ , \ j \in \mathcal{J} \tag{26b}$$
$$\sum_{j\in\mathcal{J}} \mathbf{h}_j(\mathbf{y}_j) \leq \mathbf{b} \ , \tag{26c}$$

where the variables are $\mathbf{y}_j \in \mathbb{R}^{y_j}$. The problem parameters $\mathcal{J} = \{1,\ldots,J\}$, $\mathbf{b}$ is a given vector in $\mathbb{R}^Q$, $\mathcal{Y}_j$ is a subset of $\mathbb{R}^{y_j}$, and $f_j : \text{conv}(\mathcal{Y}_j) \to \mathbb{R}$ and $\mathbf{h}_j : \text{conv}(\mathcal{Y}_j) \to \mathbb{R}^Q$ are functions defined on the convex hull of $\mathcal{Y}_j$. The following assumptions hold for the primal problem (26):
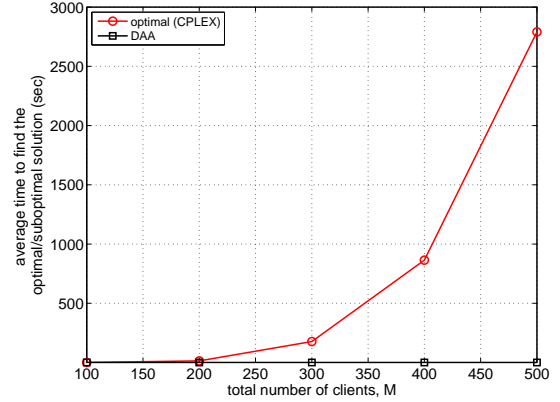
Assumption 1: There exist at least one feasible solution of problem (26).

Assumption 2: For each j, the subset of $\mathbb{R}^{y_j+Q+1}$

$$\{(\mathbf{y}_j, \mathbf{h}_j(\mathbf{y}_j), f_j(\mathbf{y}_j)) \mid \mathbf{y}_j \in \mathcal{Y}_j\} \tag{27}$$
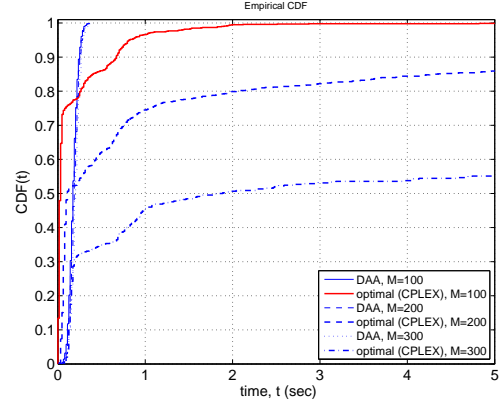
(a)

(b)

Figure 12. (a) Empirical CDF plots of Total CPU time, $N = 10$; (b) Average time to find optimal/suboptimal solution, $N = 10$

is compact.

Assumption 3: For each j, given any vector $\tilde{\mathbf{y}}$ in $\text{conv}(\mathcal{Y}_j)$, there exists $\mathbf{y} \in \mathcal{Y}_j$ such that $\mathbf{h}_j(\mathbf{y}) \leq \tilde{\mathbf{h}}_j(\tilde{\mathbf{y}})$, where $\tilde{\mathbf{h}}_j : \text{conv}(\mathcal{Y}_j) \to \mathbb{R}^Q$ is the *convexified* version of $\mathbf{h}_j$ on $\text{conv}(\mathcal{Y}_j)$ and the notation "$\leq$" here means the componentwise inequality. In particular, for all $\tilde{\mathbf{y}} \in \text{conv}(\mathcal{Y}_j)$

$$\tilde{\mathbf{h}}_j(\tilde{\mathbf{y}}) = \inf\left\{ \sum_{k=1}^{y_j+1} \alpha^k \mathbf{h}_j(\mathbf{y}^k) \,\Big|\, \tilde{\mathbf{y}} = \sum_{k=1}^{y_j+1} \alpha^k \mathbf{y}^k, \right.$$
$$\left. \mathbf{y}^k \in \mathcal{Y}_j, \sum_{k=1}^{y_j+1} \alpha^k = 1, \alpha^k \geq 0 \right\} . \tag{28}$$

Moreover, consider the dual problem of (26), i.e.,

$$\text{maxmize} \quad d(\boldsymbol{\nu}) = \inf_{\substack{\mathbf{y}_j \in \mathcal{Y}_j \\ j \in \mathcal{J}}} \left\{ \sum_{j\in\mathcal{J}} [f_j(\mathbf{y}_j) + \boldsymbol{\nu}^{\mathrm{T}} \mathbf{h}_j(\mathbf{y}_j)] - \boldsymbol{\nu}^{\mathrm{T}} \mathbf{b} \right\}$$

subject to $\quad \boldsymbol{\nu} \geq \mathbf{0}$,

with variables $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_Q) \in \mathbb{R}^Q$. Then we have

$$P^\star - D^\star \leq (Q+1) \max_{j\in\mathcal{J}} \rho_j \ , \tag{29}$$

where $P^\star$ denotes the optimal value of problem (26), $D^\star$ denotes the optimal value of the dual problem (29), and ρ_j is a nonnegative scalar such that

$$\rho_j \leq \sup_{\mathbf{y}_i \in \mathcal{Y}_j} f_j(\mathbf{y}_j) - \inf_{\mathbf{y}_i \in \mathcal{Y}_j} f_j(\mathbf{y}_j) \ . \tag{30}$$

Proof: We do not reproduce the proof here, but refer the interested reader to [45, § 5.6.1, pp. 371-376] for a rigorous proof and to [40, § 5.1.6] for an intuitive explanation. ∎

Now we rely on *Proposition 3* above to prove *Theorem 1*. The key steps of the proof can be summarized as follows: (a) We equivalently reformulate MILP (6) in the form (26) and (b) We show that the **Assumptions 1-3** of *Proposition 3* hold for this equivalent problem.

Let us start by considering the following problem which is closely related to MILP (6):

$$
\begin{aligned}
\text{minimize} \quad & \sum_{j \in \mathcal{M}} t_j \\
\text{subject to} \quad & \sum_{j \in \mathcal{M}} \bar{\beta}_{ij}\, x_{ij} \leq \sum_{j \in \mathcal{M}} t_j,\ i \in \mathcal{N} \\
& \sum_{i \in \mathcal{N}} \gamma_{ij} x_{ij} = 1,\ j \in \mathcal{M} \\
& x_{ij} \in \{0,1\},\ j \in \mathcal{M}, i \in \mathcal{N} \\
& 0 \leq t_j \leq t^{\max} + \bar{\beta}_{n_j j} x_{n_j j},\ j \in \mathcal{M}\ ,
\end{aligned}
\tag{31}
$$

where the variables are $\mathbf{t} = (t_1, \ldots, t_M)$ and $\mathbf{x} = (x_{ij})_{i \in \mathcal{N},\ j \in \mathcal{M}}$. The problem $\bar{\beta}_{ij}$ and γ_{ij} are defined as

$$
\bar{\beta}_{ij} = \begin{cases} \beta_{ij} & i \in \mathcal{N},\ j \in \mathcal{M}_i \\ 0 & \text{otherwise} \end{cases}, \quad \gamma_{ij} = \begin{cases} 1 & j \in \mathcal{M},\ i \in \mathcal{N}_j \\ 0 & \text{otherwise} \end{cases},
$$

$n_j = \arg\min_{i \in \mathcal{N}_j} \bar{\beta}_{ij}$, and $t^{\max} < \infty$ is an upper bound on $t_j^\star$, the optimal solution component of problem (31) that corresponds to t_j. For example, we use $t^{\max} = \max_{i \in \mathcal{N}, \bar{j} \in \mathcal{M}} \bar{\beta}_{i\bar{j}}$, throughout this paper. We can easily show that problem (31) is equivalent to original MILP (6) and the optimal value $P^\star$ is equal to the optimal value $p^\star$ of MILP (6), i.e.,

$$
P^\star = p^\star\ .
\tag{32}
$$

We refer to problem (31) as the *modified MILP*, which is in the form (26), where

1. $\mathcal{J} = \mathcal{M}$ and $J = M$,
2. $\mathbf{y}_j = (\mathbf{z}_j, t_j) \in \mathbb{R}^{y_i}$, with $\mathbf{z}_j = (x_{ij})_{i \in \mathcal{N}}$ and $y_j = N + 1$,
3. $f_j(\mathbf{y}_i) = t_j$,
4. $\mathcal{Y}_j = \{((x_{ij})_{i \in \mathcal{N}}, t_j) \mid \sum_{i \in \mathcal{N}} \gamma_{ij} x_{ij} = 1,\ x_{ij} \in \{0,1\},\ i \in \mathcal{N},\ t_j \in [0, t^{\max} + \bar{\beta}_{n_j j} x_{n_j j}]\}$,
5. $\mathbf{h}_j(\mathbf{y}_j) = ((\bar{\beta}_{1j} x_{1j} - t_j), \ldots, (\bar{\beta}_{Nj} x_{Nj} - t_j)) \in \mathbb{R}^Q$, with $Q = N$, and
6. $\mathbf{b} = \mathbf{0}$.

Let us now show that the **Assumptions 1-3** of *Proposition 3* hold for the modified MILP (31). It is straightforward to see

that **Assumptions 1-2** hold. Checking whether **Assumption 3** holds is less trivial as we show next.

Let $\tilde{\mathbf{y}}$ be any given vector in $\text{conv}(\mathcal{Y}_j)$. From the definition of $\tilde{\mathbf{h}}_j(\tilde{\mathbf{y}})$ (see (28)), we can express it as

$$
\tilde{\mathbf{h}}_j(\tilde{\mathbf{y}}) = \sum_{k=1}^{y_j+1} \alpha^k \mathbf{h}_j(\mathbf{y}^k)
\tag{33}
$$

for some $\mathbf{y}^k \in \mathcal{Y}_j$ and α^k such that $\tilde{\mathbf{y}} = \sum_{k=1}^{y_j+1} \alpha^k \mathbf{y}^k$, $\sum_{k=1}^{y_j+1} \alpha^k = 1$, $\alpha^k \geq 0$. Particularized to the modified MILP (6), we can write the set of equations (34a)-(34g), where $t_c^{\max} = t^{\max} + \bar{\beta}_{nj}$ and $\mathbf{y} = (0, 0, \ldots, 1, 0, \ldots, 0, t_c^{\max})$, which is feasible, i.e., $\mathbf{y} \in \mathcal{Y}_j$ and **Assumption 3** holds. Note that the first three equalities (34a)-(34c) follows from straightforward manipulations, inequality (34d) follows from that $t_j^k \leq t^{\max} + \bar{\beta}_{n_j j} x_{n_j j}^k$ for all k, inequality (34e) follows from that $\sum_k (\alpha^k x_{1j}^k) \geq 0$ for all k, and last two equalities (34f)-(34g) follow from straightforward manipulations.

Finally, let us show that $D^\star = d^\star$, where $D^\star$ is the dual optimal value of the associated dual problem [compare with (29)] of the modified MILP (31). We denote by $P_{\text{relax}}^\star$ the optimal value of the LP relaxation of problem (31). Thus, we have

$$
D^\star = P_{\text{relax}}^\star = p_{\text{relax}}^\star = d^\star\ ,
\tag{35}
$$

where the first equality follows from a similar approach as described in the proof of *Proposition 2*, the second equality follows from that the LP relaxations of problem (31) and of problem (6) (see problem (23)) have the same optimal value, and the last equality follows from *Proposition 2*.

By using (32), (34), and that **Assumptions 1-3** hold for the MILP (6) together with *Proposition 3*, we have

$$
\begin{aligned}
p^\star - d^\star &\leq (N + 1) \max_{j \in \mathcal{M}} \rho_j \\
&\leq (N + 1) \max_{j \in \mathcal{M}} \left(t^{\max} + \bar{\beta}_{n_j j} \right) \\
&= (N + 1) \left(\varrho + \max_{j \in \mathcal{M}} \varrho_j \right)\ ,
\end{aligned}
\tag{36}
$$

where the first inequality follows from (29), the second inequality follows from (30), $\sup t_j = t^{\max} + \bar{\beta}_{n_j j}$, and $\inf t_j = 0$, and the last equality follows from that $t^{\max} = \varrho$ and $\bar{\beta}_{n_j j} = \varrho_j$, which yields (25).

$$
\tilde{\mathbf{h}}_j(\tilde{\mathbf{y}}) = \sum_k \alpha^k (\bar{\beta}_{1j} x_{1j}^k - t_j^k, \ldots, \bar{\beta}_{n_j j} x_{nj}^k - t_j^k, \ldots, \bar{\beta}_{Nj} x_{Nj}^k - t_j^k)
\tag{34a}
$$

$$
= \sum_k (\bar{\beta}_{1j} \alpha^k x_{1j}^k - \alpha^k t_j^k, \ldots, \bar{\beta}_{n_j j} \alpha^k x_{nj}^k - \alpha^k t_j^k, \ldots, \bar{\beta}_{Nj} \alpha^k x_{Nj}^k - \alpha^k t_j^k)
\tag{34b}
$$

$$
= \left(\bar{\beta}_{1j} \sum_k (\alpha^k x_{1j}^k) - \sum_k (\alpha^k t_j^k), \ldots, \bar{\beta}_{n_j j} \sum_k (\alpha^k x_{nj}^k) - \sum_k (\alpha^k t_j^k), \ldots, \bar{\beta}_{Nj} \sum_k (\alpha^k x_{Nj}^k) - \sum_k (\alpha^k t_j^k) \right)
\tag{34c}
$$

$$
\geq \left(\bar{\beta}_{1j} \sum_k (\alpha^k x_{1j}^k) - \sum_k (\alpha^k t_j^k), \ldots, \bar{\beta}_{n_j j} \sum_k (\alpha^k x_{nj}^k) - \sum_k \left(\alpha^k (t^{\max} + \bar{\beta}_{n_j j} x_{nj}^k) \right), \ldots, \bar{\beta}_{Nj} \sum_k (\alpha^k x_{Nj}^k) - \sum_k (\alpha^k t_j^k) \right)
\tag{34d}
$$

$$
\geq (\bar{\beta}_{1j} 0 - t_c^{\max}, \ldots, -t^{\max}, \ldots, \bar{\beta}_{Nj} 0 - t_c^{\max})
\tag{34e}
$$

$$
= (-t_c^{\max}, \ldots, -t^{\max}, \ldots, -t_c^{\max})
\tag{34f}
$$

$$
= \mathbf{h}_j(\mathbf{y})\ ,
\tag{34g}
$$

We note from (36) that the duality gap *does not* depend on M, i.e., the total number of clients. Moreover, we note that $p^\star \to \infty$ as $m \to \infty$ (see the objective function of original problem (5)). Thus, we conclude that, the *relative* duality gap $(p^\star - d^\star)/p^\star$ diminishes to zero as $M \to \infty$.

REFERENCES

[1] F. Giannetti, M. Luise, and R. Reggiannini, "Mobile and personal communications in the 60 GHz band: A survey," *Springer Wireless Personal Communications*, vol. 10, no. 2, pp. 207–243, July 1999.

[2] P. Smulders, H. Yang, and I. Akkermans, "On the design of low-cost 60GHz radios for multigigabit-per-second transmission over short distances," *IEEE Communications Magazine*, vol. 45, no. 12, pp. 44–51, Dec. 2007.

[3] A. D. Oliver, "Millimeter wave systems past, present and future," *IEEE Proceedings*, vol. 136, no. 1, pp. 35–52, Feb. 1989.

[4] C. H. Doan, S. Emami, D. A. Sobel, A. M. Niknejad, and R. W. Brodersen, "Design considerations for 60 GHz cmos radios," *IEEE Communications Magazine*, vol. 42, no. 12, pp. 132–140, Dec. 2004.

[5] Y. P. Zhang and D. Liu, "Antenna-on-chip and antenna-in-package solutions to highly integrated millimeter-wave devices for wireless communications," *IEEE Transactions on Antennas and Propagation*, vol. 57, no. 10, pp. 2830–2841, Oct. 2009.

[6] "IEEE 802.15.3c Part 15.3: Wireless medium access control (MAC) and physical layer (PHY) specifications for high rate wireless personal area networks (WPANs) amendment 2: Millimeter-wave-based alternative physical layer extension," 2009.

[7] "IEEE 802.11ad. Part 11: Wireless lan medium access control (MAC) and physical layer (PHY) specifications - amendment 3: Enhancements for very high throughput in the 60 GHz band," 2012.

[8] "WirelessHD: WirelessHD specification overview," 2009.

[9] "Wireless Gigabit Alliance. [online]. Available: http://wirelessgigabitalliance.org/," .

[10] "IEEE P802.11n Part 11: Wireless lan medium access control (mac) and physical layer (phy) specications: Enhancements for higher throughput," 2009.

[11] L. Yang and G. B. Giannakis, "Ultra-wideband communications - an idea whose time has come," *IEEE Signal Processing Magazine*, vol. 21, no. 6, pp. 26–54, Nov. 2004.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[13] Y. Bejerano and R. Bhatia, "Mifi: A framework for fairness and QoS assurance in current IEEE 802.11 networks with multiple access points," *IEEE Transactions on Networking*, vol. 14, no. 4, pp. 849–862, Aug. 2006.

[14] R. Horst, P. Pardolos, and N. Thoai, *Introduction to Global Optimization*, vol. 48, Kluwer Academic Publishers, Dordrecht, Boston, London, second edition, 2000.

[15] W. Arbaugh, A. Mishra, and M. Shin, "An empirical analysis of the IEEE 802.11 mac layer handoff process," *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 2, pp. 93–102, Apr. 2003.

[16] Y. Bejerano, S. Han, and L. Li, "Fairness and load balancing in wireless lans using association control," *IEEE Transactions on Networking*, vol. 15, no. 3, pp. 560–573, June 2007.

[17] B. Kauffmann, F. Baccelli, A. Chaintreau, K. Papagiannaki, and C. Diot, "Measurement-based self organization of interfering 802.11 wireless access networks," in *IEEE INFOCOM*, Anhcorage, Alaska, USA, May 2007, pp. 1451–1459.

[18] A. Kumar and V. Kumar, "Optimal association of stations and aps in an ieee 802.11 wlan," in *National Conference on Communications (NCC)*, IIT Kharagpur, Jan. 2005, pp. 567–573.

[19] M. Shin, A. Mishra, and W. A. Arbaugh, "Improving the latency of 802.11 hand-offs using neighbor graphs," in *ACM MOBISYS*, New York, USA, June 2004, pp. 70–82.

[20] S. Shakkottai, E. Altman, and A. Kumar, "The case for non-cooperative multihoming of users to access points in IEEE 802.11 WLANs," in *IEEE INFOCOM*, Barcelona, Spain, Apr. 2006, pp. 1–12.

[21] V. Mhatre and K. Papagiannaki, "Using smart triggers for improved user performance in 802.11 wireless networks," in *ACM MOBISYS*, Anhcorage, Alaska, USA, June 2006, pp. 246–259.

[22] Y. Amir, C. Danilov, M. Hilsdale, R. Musaloiu-Elefteri, and N. Rivera, "Fast handoff for seamless wireless mesh networks," in *ACM MOBISYS*, Uppsala, Sweden, June 2006, pp. 83–95.

[23] C. Park, J. Hur, C. Kim, Y. Shin, and H. Yoon, "Pre-authentication for fast handoff in wireless mesh networks with mobile aps," in *International Workshop on Information Security Applications*, NY, USA, June 2006, pp. 349–363.

[24] A. Balachandran, P. Bahl, and G. Voelker, "Hot-spot congestion relief in public-area wireless networks," *ACM SIGCOMM Computer Communication Review*, vol. 32, pp. 65–76, Jan. 2002.

[25] D. Lee, G. Chandrasekaran, M. Sridharan, and P. Sinha, "Association management for data dissemination over wireless mesh networks," *Elsevier Computer Networks*, vol. 51, no. 15, pp. 4338–4355, Oct. 2007.

[26] D. Lee, G. Chandrasekaran, and P. Sinha, "Optimizing broadcast load in mesh networks using dual-association," in *EEE Workshop on Wireless Mesh Networks*, Santa Clara, CA, 2005.

[27] G. Athanasiou, T. Korakis, O. Ercetin, and L. Tassiulas, "Dynamic cross-layer association in 802.11-based mesh networks," in *IEEE INFOCOM*, Anhcorage, Alaska, USA, May 2007, pp. 2090–2098.

[28] G. Athanasiou, T. Korakis, O. Ercetin, and L. Tassiulas, "A cross-layer framework for association control in wireless mesh networks," *IEEE Transactions on Mobile Computing*, vol. 8, no. 1, pp. 65–80, Jan. 2009.

[29] C. W. Pyo and H. Harada, "Throughput analysis and improvements of hybrid multiple access in ieee 802.15.3c mmWave-WPAN," *IEEE Journal on Selected Areas on Communications*, vol. 27, no. 8, pp. 1414–1424, Oct. 2009.

[30] S. Singh, R. Mudumbai, and U. Madhow, "Distributed coordination with deaf neighbors: efficient medium access for 60 GHz mesh networks," in *EEE INFOCOM*, Santa Barbara, CA, Mar. 2010, pp. 1–9.

[31] S. Singh, F. Ziliotto, U. Madhow, E. M. Belding, and M. J. W. Rodwell, "Millimeter wave WPAN: cross-layer modeling and multihop architecture," in *EEE INFOCOM*, Anchorage, Alaska, US, May 2007, pp. 2336–2240.

[32] L. X. Cai, L. Cai, X. Shen, and J. W. Mark, "Rex: a randomized exclusive region based scheduling scheme for mmwave WPANs with directional antenna," *IEEE Transactions on Wireless Communications*, vol. 9, no. 1, pp. 113–121, Jan. 2010.

[33] S. Singh, R. Mudumbai, and U. Madhow, "Interference analysis for highly directional 60-GHz mesh networks: The case for rethinking medium access control," *IEEE Transactions on Networking*, vol. 19, no. 5, pp. 1513–1527, Oct. 2011.

[34] J. Qiao, L. X. Cai, X. S. Shen, and J. W. Mark, "Enabling multi-hop concurrent transmissions in 60 GHz wireless personal area networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 11, pp. 3824–3833, Nov. 2011.

[35] R. Mudumbai, S. Singh, and U. Madhow, "Medium access control for 60 GHz outdoor mesh networks with highly directional links," in *EEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009, pp. 2871–2875.

[36] A. Kumar, D. Manjunath, and J. Kuri, *Wireless Networking*, ELSEVIER Inc., Burlington, MA , USA, 2008.

[37] S. Y. Geng, J. Kivinen, X. W. Zhao, and P. Vainikainen, "Millimeter-wave propagation channel characterization for short-range wireless communications," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 1, pp. 3–13, Jan. 2009.

[38] G. Athanasiou, I. Broustis, T. Korakis, and L. Tassiulas, "Lac: Load-aware channel selection in 802.11 wlans," in *EEE PIMRC*, Cannes, France, Sept. 2008, pp. 1–6.

[39] S. Boyd, "Subgradient methods," [Online]. Available: http://www.stanford.edu/class/ee364b/lectures/subgrad_method_slides.pdf, 2007.

[40] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 2nd edition, 1999.

[41] R. T. Rockafellar, *Convex Analysis*, Prinston, N.J., Princeton University Press, 1970.

[42] K. Truemper, *Matroid Decomposition*, Leibniz, Cliffside Drive, Plano, Texas, USA, 1998, revised edition.

[43] "IBM ILOG CPLEX Optimizer. [online]. Available: http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/," .

[44] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE Transactions on Networking*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[45] D. P. Bertsekas, *Constrained optimization and Lagrange Multiplier Method*, Athena Scientific, Belmont, MA, 1998.