THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Maximising the Utility of Enterprise Millimetre-Wave Networks

Citation for published version:
Facchi, N, Gringoli, F & Patras, P 2018, 'Maximising the Utility of Enterprise Millimetre-Wave Networks', *Computer Communications*, vol. 119, pp. 29-42. https://doi.org/10.1016/j.comcom.2018.01.011

Digital Object Identifier (DOI):
10.1016/j.comcom.2018.01.011

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Computer Communications

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 02. May. 2024

Maximising the Utility of Enterprise Millimetre-Wave Networks

N. Facchi[1], F. Gringoli[2], and P. Patras[3]

[1]Deptartment of Information Engineering and Computer Science, University of Trento
[2]Deptartment of Information Engineering, CNIT / University of Brescia
[3]School of Informatics, University of Edinburgh

Abstract

Millimetre-wave (mmWave) technology is a promising candidate for meeting the intensifying demand for ultra fast wireless connectivity, especially in high-end enterprise networks. Very narrow beam forming is mandatory to mitigate the severe attenuation specific to the extremely high frequency (EHF) bands exploited. Simultaneously, this greatly reduces interference, but generates problematic communication blockages. As a consequence, client association control and scheduling in scenarios with densely deployed mmWave access points become particularly challenging, while policies designed for traditional wireless networks remain inappropriate. In this paper we formulate and solve these tasks as utility maximisation problems under different traffic regimes, for the first time in the mmWave context. We specify a set of low-complexity algorithms that capture distinctive terminal deafness and user demand constraints, while providing near-optimal client associations and airtime allocations, despite the problems' inherent NP-completeness. To evaluate our solutions, we develop an NS-3 implementation of the IEEE 802.11ad protocol, which we construct upon preliminary 60GHz channel measurements. Simulation results demonstrate that our schemes provide up to 60% higher throughput as compared to the commonly used signal strength based association policy for mmWave networks, and outperform recently proposed load-balancing oriented solutions, as we accommodate the demand of 33% more clients in both static and mobile scenarios.

1 Introduction

Users' predilection for wireless connectivity is increasingly incompatible with the stringent performance requirements of emerging applications, including uncompressed ultra high definition (HD) video, wire-equivalent docking, virtual reality streaming, and low latency data upload/download [1]. In response, the industry is exploring the use of license exempt extremely high frequencies (mmWave) in the 60GHz band, for short range multi-gigabit per second wireless communications [2]. These efforts have already materialised as new standard amendments, e.g. IEEE 802.11ad [3], recently unveiled WiGig routers [4], and business-oriented laptops [5].

Different to legacy wireless solutions, mmWave technology leverages vast spectral resources (up to 2GHz-wide channels) currently underutilised. Their potential, however, can only be realised through highly directional digital beamforming, since signals attenuate dramatically in this frequency range [6]. Forming narrow beams not only mitigates fading, but also reduces interference between adjoining TX/RX pairs. Consequently, links between stations and access points (APs) can be regarded as pseudo-wired and channel access no longer subject to collisions. The caveat is that *associated clients are shut out whenever an AP communicates with anyone of their neighbours*. To ensure all stations are given opportunities to receive and/or transmit packets, the IEEE 802.11ad standard defines a Service Period (SP) based channel access mechanism, though the task of scheduling SPs is deliberately left open to accommodate proprietary implementations [3].

This problem is further complicated in enterprise wireless networks, including stocks trading offices, broadcasting studios that manipulate raw ultra HD video,[1] and emerging tactile Internet environments.[2]

[1] See BBC IP Studio, http://www.bbc.co.uk/rd/projects/ip-studio

[2] E.g. http://www.huawei.com/minisite/5g/en/touch-internet-5G.html

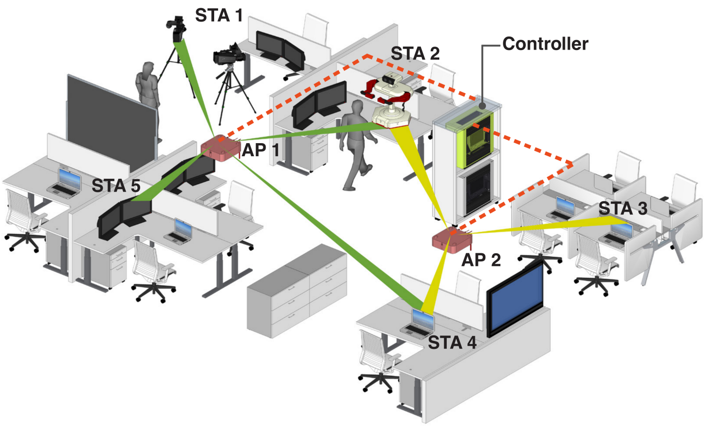


Figure 1: Simple example of the envisioned system, comprising two APs and five active stations. APs are connected to the central controller that runs the algorithms we introduce in this paper to find optimal client associations and airtime allocations that maximise network utility. Clients and APs communicate over directional links (shaded beams). Stations 2 and 4 are within the range of both APs.

There, mmWave clients will often lie within the range of multiple APs that serve different numbers of stations, as exemplified in Fig. 1, possibly having dissimilar traffic demands. In this scenarios, *the challenge is deciding both to which AP to associate clients and what airtime budget to allocate for each.* With an appropriate logic that is yet to be developed, such decisions could be enforced by *central controllers* similar to those widely used in today's enterprise wireless networks to load balance clients over the available APs and bands. [3]

Commonly adopted signal strength based association policies are oblivious to load conditions [7] and thus may lead to inappropriate decisions in mmWave networks. Likewise, client association mechanisms for traditional 802.11 wireless networks [8, 9] or cellular systems (e.g. [10]) are ill suited to mmWave, due to the substantial differences between these technologies. Association control and SP allocation in mmWave networks are largely unexplored; recent solutions focus primarily on load balancing, downplaying airtime budget constraints, and requiring non standard signalling [11]. Without carefully controlling which AP serves each client and for how long on average, we argue that the overall network throughput performance will be sub-optimal and user demand often unsatisfied, even when sufficient resources are available in the network.

Contributions: In this paper we formulate and solve the client association control and SP allocation tasks in high-end mmWave networks as utility maximisation problems, *capturing the severe terminal issues unique to such systems.* We consider general scenarios with both backlogged stations and clients with finite load requirements, which encounter heterogeneous link qualities to the APs within range, and may be either static or mobile. We use the same definition of utility as given by F. Kelly, i.e. the sum of the logarithms of individual station throughputs [12], which strikes a *good trade-off between maximising network throughput and providing airtime fairness.* We envision a centralised network driven architecture (as in Fig. 1) that could be built upon recent advances in software-defined networking (SDN) [13], and IEEE 802.11 protocol amendments for wireless network [14] and radio resource management [15]. These would enable the central controller to collect information from the deployed APs and their clients, and *enforce the computed client–AP associations and airtime allocations in a standard compliant fashion.* Specifically,

1. We show that under backlog conditions, the utility optimisation problem we pose is NP-complete, but its relaxed version is convex. We use well-established Lagrangian tools to solve the relaxed version and give a linear complexity iterative rounding algorithm, to derive solutions to the original problem;

2. For finite load scenarios we introduce an algorithm that captures terminal deafness and traffic

[3] This is the case of commercially available solutions including Cisco WLC (`http://www.cisco.com/c/en/us/products/wireless/wireless-lan-controller/index.html`) and Aruba Mobility Controller (`http://www.arubanetworks.com/products/networking/controllers`)

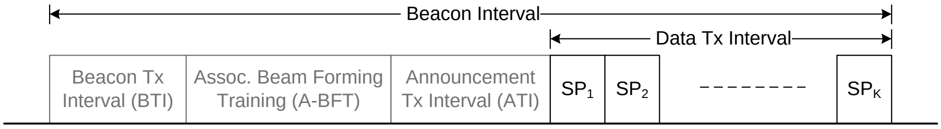


Figure 2: IEEE 802.11ad super-frame. Frame transmissions are performed in a scheduled fashion during the Data Tx Interval using service periods (SPs) [3].

constraints, and combines simulated annealing with airtime water filling techniques, to find near-optimal association matrices and airtime allocation vectors almost in real-time;

3. Using an NS-3 based 802.11ad simulation module we develop, building on preliminary 60GHz channel measurements, we demonstrate that our solutions provide up to 60% higher total throughput as compared to the standard's default signal strength based policy, while satisfying the demand of 33% more clients, in comparison with the recently proposed DAA mechanism [11].

To the best our knowledge, this is the first attempt to cast client association control and airtime allocation as utility maximisation problems in the mmWave context, whilst the solutions we provide are demonstrably effective under a broad range of network conditions.

2 System Architecture

We consider enterprise mmWave wireless network deployments with M access points and N clients. To thwart high signal attenuation inherent to the 60GHz band, each station (AP or client) is equipped with a transceiver that can digitally form and steer beams of very narrow widths, to transmit or to receive packets. Therefore, interference levels can be considered negligible and not impacting on the achievable bit rates. This is in line with recent studies [16] that confirm even uncoordinated packet exchanges between different transmit–receive pairs experience very small collision probabilities. While we acknowledge that interference increases in outdoor deployments as cell density grows [17], this is not applicable to indoor scenarios where ceiling-mounted access points will experience no interference when the angular separation between links is as little as 10-12° [18]. This is feasible in office environments even with consumer-grade equipment whose antenna patterns exhibit side lobes [19]. As such, in our setting mmWave links can be regarded as *pseudo-wired* point-to-point connections.

We focus on emerging enterprise networks with SDN capabilities that enable controllers to manipulate the configuration of APs (and their clients) via protocols such as NETCONF [20]. In addition, APs can request neighbour and link measurement reports from stations, through 802.11k primitives [15], which a centralised controller will use to enforce control association to specific APs, as determined by the algorithms we propose in this work. mmWave channel sounding capable of measuring multi-path delays with 2ns granularity has been recently demonstrated [21] and APs can represent per-client PHY rate measurements as 2-byte *(client_ID, PHY_rate)* tuples whose transmission every beacon interval incurs negligible overhead. Network assisted association requests as we proposed will be underpinned by 802.11v enhancements [14]. Upon bootstrap, we assume clients associate following the de facto signal-to-noise ratio (SNR) based procedure, as stipulated by the standard [3].

Given the carrier-grade requirements specific to business-oriented networks, we focus on 802.11ad mmWave networks where APs operate in a scheduled fashion using the Service Period (SP) channel access scheme. Channel time is divided into fixed-size beacon intervals (∼100ms) that carry at the beginning synchronisation information, beam form training related signalling, and announcement of the subsequent SPs assigned [3]. The duration of these is variable, while stations may aggregate multiple data frames within their SPs, to improve protocol efficiency. We summarise this behaviour in Fig. 2. By this approach, clients associated to the same AP do not contend for channel access and each AP serves only one associated client at a time during non-overlapping scheduled SPs, whilst multiple APs could simultaneously serve different clients. We confine consideration to protocol features already specified by the approved 802.11ad standard [3]. We note however that the association control and airtime allocation solutions we propose herein could be easily adapted to encompass more aggressive modulation schemes, channel bonding, and MU-MIMO enhancements, which are candidates for future standard amendments, e.g. 802.11ay [22].

The potential of mmWave networks can only be realised using highly directional narrow beams and any two communicating stations must know how to configure their antennas for beam alignment. This is achieved through antenna and beamforming training, and the standard specifies a set of protocols for this purpose, but does not provide a precise rule on how to employ these. Importantly the standard does not specify how to select the best antenna configuration, the training process being complex and chip implementation dependant.[4] For these reasons, in this work, we assume all stations are capable of performing reliable beamforming training, while we do not consider the precise details of the actual mechanism. However, our modelling and the NS-3 based simulator we develop take into account the possible overhead due to beamforming protocols. Our focus is on the optimal client association and airtime allocation tasks, which have been largely overlooked by the research community. mmWave channel modelling remains outside the scope of our work.

3 Throughput Analysis

In this section we formalise the throughput performance of mmWave clients in the envisioned enterprise scenario. Recall that we aim to meet strict quality-of-service requirements and therefore medium access control is regulated by the APs using the SP mechanism. We consider realistic multi-rate conditions, i.e. clients employ Single Carrier or OFDM PHY modulation and coding schemes that depend on the signal-to-noise ratio (SNR) on the links to the APs and the channel bandwidth. Note that the SNR may be subject to link blockage due to human movement or other obstructions on the direct propagation path between a client and an AP. Our model captures such circumstances as low achievable rates, or null bit rates when establishing a communications link may not be (temporarily) feasible. We denote by $r_{i,j}$ the bit rate a client i achieves when transmitting to an AP j during an SP.

Consider a binary association vector $\mathbf{x}$, where an element $x_{i,j} \in \{0,1\}$ indicates whether a client i is associated to an AP j ($i = 1, \ldots, N$, and $j = 1, \ldots, M$), i.e.

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

We initially assume that all stations are backlogged (saturation conditions) and later relax this assumption to account for general finite load scenarios. Under these circumstances, maximising network utility has been shown to correspond to allocating equal airtimes to all stations connected to a given AP, irrespective of their bit rates [24]. As such, our goal is to allocate a $t_{i,j}$ fraction of an AP j's total airtime budget to a connected client i. Formally,

$$t_{i,j} = \frac{T_j - O_j}{\sum_{k=1}^{N} x_{k,j}}, \forall i, j, \tag{2}$$

where T_j denotes the duration of a super-frame (beacon interval) as enforced by AP j and O_j is the protocol overhead due to beacon transmission, (optional) beamforming (BF) training, and management operations (see Fig. 2). Given the scheduled nature of the medium access in 802.11ad, hereafter we will use the terms airtime and service period (SP) interchangeably.

With the above, the throughput $S_{i,j}$ obtained by client i when connected to AP j is given by

$$S_{i,j} = \frac{r_{i,j} t_{i,j}}{T_j} = \frac{h_j r_{i,j}}{\sum_{k=1}^{N} x_{k,j}}, \tag{3}$$

where $h_j = (T_j - O_j)/T_j$. In what follows we only consider feasible associations, i.e. those for which a client i falls in the coverage of an AP j and thus $r_{i,j} \neq 0$.

4 Utility Maximisation for Saturation Scenario

Our objective is to find the client *association matrix* $\mathbf{x}$ that *maximises the total utility of the network*, i.e. solve the following optimisation problem:

[4] Mechanisms to enable learning with high accuracy the relative positions of devices in the network, currently under discussion in the IEEE 802.11az task group [23], may alleviate the complexity of beamforming.

$$\max_{\mathbf{x}} U := \sum_{j=1}^{M} \sum_{i=1}^{N} x_{i,j} \log S_{i,j}, \tag{4}$$

$$\text{s.t.} \sum_{j=1}^{M} x_{i,j} = 1, \forall i; \qquad \text{(single AP association)} \tag{5}$$

$$x_{i,j} \in \{0,1\}, \forall i,j. \qquad \text{(function domain)} \tag{6}$$

Lemma 1. *The optimisation problem specified by (4)–(6) is NP-complete.*

Proof. Denote P the problem given in (4)–(6). By (3), $S_{i,j}$ is a function of the inverse sum of some terms $x_{k,j}$, for all j. Therefore the objective of the problem posed is a non-linear function of variables in the $\{0,1\}^N$ set (6). Now, consider a simpler problem P' where a client i attains a small constant throughput θ_j when connected to AP j, irrespective of the number of clients this servers. The objective (4) becomes $\sum_{j=1}^{M} \sum_{i=1}^{N} \theta_j x_{i,j}$ and since $x_{i,j} \in \{0,1\}$, the constraint (5) is equivalent to $\sum_{j=1}^{M} x_{i,j} \leq 1$. It follows that P' is an instance of the 0–1 knapsack problem, which by Theorem 15.8 in [25] is NP-complete. Since a solution to P can be verified, P is NP, and as $P > P'$, while P' is NP-complete, then P is NP-complete. □

Finding a solution to this type of optimisation problems within reasonable time is known to be difficult [26]. Consequently, we proceed with a relaxation of our original problem, replacing the constraint $x_{i,j} \in \{0,1\}$ and allowing $x_{i,j}$ in the $[0,1]$ interval (Fractional User Association). This is similar in nature to the linear programming relaxation of the set cover problem studied by Lovász [27]. We then give a linear complexity iterative rounding algorithm that derives a solution to the original problem from that of the relaxed version.

We express formally the relaxed optimisation problem as:

$$\max_{\mathbf{x}} U := \sum_{j=1}^{M} \sum_{i=1}^{N} x_{i,j} \log S_{i,j}, \tag{7}$$

$$\text{s.t. } S_{i,j} \leq h_j r_{i,j}, \forall i,j; \tag{8}$$

$$\sum_{j=1}^{M} x_{i,j} \leq 1, \forall i; \tag{9}$$

$$-x_{i,j} \leq 0, \forall i,j. \tag{10}$$

The constraint in (9) ensures any client i does not communicate to more than one AP at a given time (single transceiver), while (8) ensures that the throughput allocated to client i when connected to AP j does not exceed the maximum attainable bit rate under the current signal quality conditions, if client i was the only one connected to AP j. Note that in the original problem (4), where $x_{i,j} \in \{0,1\}$, it was note necessary to explicitly impose this constraint, since it is implicitly satisfied by (5).

4.1 Convexity Properties and Problem Solution

Next we analyse the convexity properties of the objective function in the relaxed optimisation problem and give insights into the solution space.

Lemma 2. *The utility U function defined by (7) is concave.*

Proof. The second order partial derivative of the terms $x_{i,j} \log S_{i,j}$ with respect to $x_{i,j} \neq 0$ is

$$\frac{\partial^2 \left(x_{i,j} \log S_{i,j}\right)}{\partial^2 x_{i,j}} = -\frac{1}{\sum_{k=1}^{N} x_{k,j}} - \frac{\sum_{k=1,k\neq i}^{N} x_{k,j}}{\left(\sum_{k=1}^{N} x_{k,j}\right)^2} < 0,$$

and the same with respect to $x_{l,j} \neq 0, l \neq i$ is

$$\frac{\partial^2 \left(x_{i,j} \log S_{i,j}\right)}{\partial^2 x_{l,j}} = -\frac{\sum_{k=1,k\neq i}^{N} x_{k,j}}{\left(\sum_{k=1}^{N} x_{k,j}\right)^2} < 0.$$

Thus the Hessian $\nabla^2 \mathbf{x} \log(\mathbf{S})^T$ is negative semi-definite. By Boyd and Vandenberghe [28], it follows that functions $x_{i,j} \log S_{i,j}$ are concave, and since the utility U is an affine combination of such functions, then it is concave. □

Since we are working with multiple single-hop TDMA-type systems, the capacity region of which is convex [29], constraint (8) is also convex. Further, constraints given by (9) and (10) are convex and thus by Lemma 2 the relaxed optimisation problem defined by (7)–(10) is convex and a solution exists. Slater's sufficient condition is satisfied and thus strong duality holds. The Lagrangian is

$$\begin{aligned} L(\mathbf{x}, \lambda, \mu, \nu) = & - \sum_{j=1}^{M} \sum_{i=1}^{N} x_{i,j} \log S_{i,j} \\ & + \sum_{i=1}^{N} \sum_{j=1}^{M} \lambda_{i,j} \left(S_{i,j} - h_j r_{i,j}\right) \\ & + \sum_{i=1}^{N} \mu_i \left(\sum_{j=1}^{M} x_{i,j} - 1 \right) - \sum_{i=1}^{N} \sum_{j=1}^{M} \nu_{i,j} x_{i,j} \end{aligned}$$

The Karush-Kuhn-Tucker (KKT) condition [30] for $S_{i,j}$ is

$$\frac{\partial L}{\partial S_{i,j}} = 0,$$

which gives

$$\lambda_{i,j} = x_{i,j} \frac{1}{S_{i,j}}.$$

In the above, we distinguish two possible cases: (1) client i associates to AP j and thus $x_{i,j} > 0$, which means $\lambda_{i,j} = 0$ (note that $S_{i,j}$ is non zero with $x_{i,j} > 0$); and (2) $x_{i,j} = 0$ from which it follows that $\lambda_{i,j} = 0$. From complementary slackness it follows that the inequality constraint (8) is not tight and thus the optimum $\mathbf{x}'$ may not be unique. However we can still employ the widely used trust region method (TRM) to solve numerically the relaxed optimisation problem [31].

4.2 Rounding Algorithm

Once we solved the relaxed problem (7), the next step is finding a solution to the original utility maximisation problem (4), which recall is NP-complete. To this end, we design an iterative rounding algorithm that converts the fractional association matrix $\mathbf{x}'$, to an integer association matrix $\mathbf{x}^*$, which is the solution of the original problem (4).

The simplest way to accomplish this would be a maximum likelihood approach, i.e.

$$x_{i,j}^* = \begin{cases} 1, & \text{if } \max_k \{x'_{i,k}\} = x'_{i,j}; \\ 0, & \text{otherwise;} \end{cases} \tag{11}$$

however, this performs poorly in numerous situations, where the solution it returns is identical to that of the SNR-based association. We exemplify this in Fig. 3 for a simple topology with two APs and four stations. Also shown in the figure is the superior performance (87% higher total throughput) attainable with the iterative rounding algorithm we propose, whose pseudo-code is given in Algorithm 1 and detailed next.

The proposed rounding algorithm requires N iterations (only depending on the number of clients) and thus has linear complexity $O(N)$. We work with a set X which maintains a list of (i, j) tuples

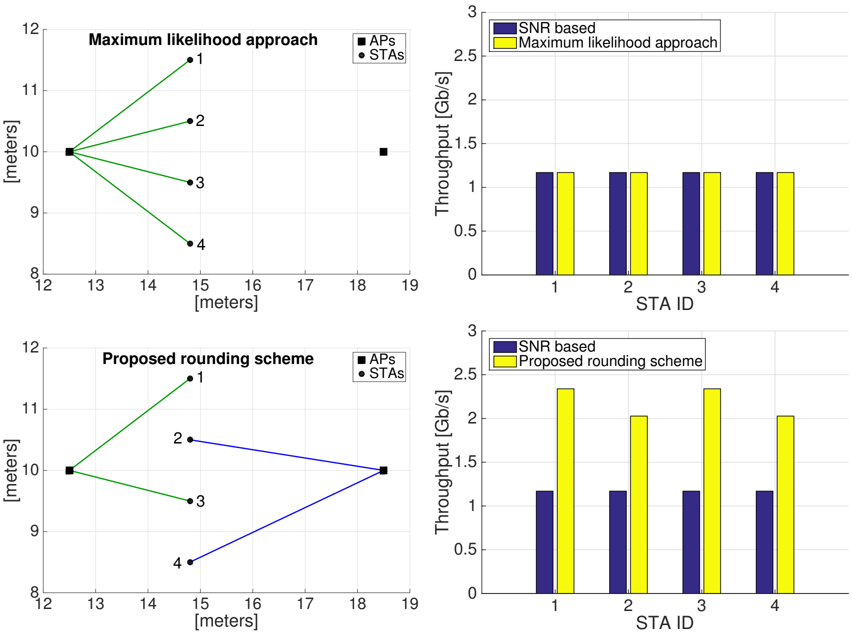


Figure 3: Example of mmWave network with 2 APs and 4 clients. Association enforced by the maximum likelihood approach (top left) and attainable throughput vs that of SNR-based method (top right). Association enforced by the proposed rounding algorithm (bottom left) and corresponding throughput (bottom right). Aggregate throughput gain achieved is 87%. Numerical results.

corresponding to the $x'_{i,j}$ terms that have not been yet subject to rounding. Initially, X contains all the tuples and we remove j of them at each iteration, as one client is assigned to a *single* AP, until X is empty.

Each iteration is composed of a rounding (lines 3, 8–9) and an update operation (lines 4–7). The rounding operation sets to 1 (line 8) the $x'_{\hat{i},\hat{j}}$ element whose value is the largest among all $x'_{i,j}$ for which (i,j) is in X at the current iteration n (line 3). Then, we set to 0 the $x'_{\hat{i},j}$ terms, $\forall j \neq \hat{j}$ (line 9). The key idea is to prioritise the rounding of the $x'_{i,j}$ closest to 1. If there is more than one such $(\hat{i},\hat{j})$ tuple, the algorithm chooses one randomly.

The update operation recomputes the values of all the $x'_{i,j}$ which have not been rounded during iteration n and that are still to be rounded (line 6). The new value is computed by adding to $x'_{i,j}$ the

Algorithm 1 Iterative rounding

```
Require: $\mathbf{x}'$ is a feasible solution of problem (7).
Ensure: $\mathbf{x}^*$ is a feasible solution of problem (4).
 1: $X = \{(i,j) \mid 1 \le i \le N, 1 \le j \le M\}$; $X^0 = X$; $n = 0$.
 2: repeat
 3:     Find $\hat{i}$, $\hat{j}$ s.t. $x'_{\hat{i},\hat{j}} = \max_{(i,j)\in X^n}\left(x'_{i,j}\right)$;      {Find $x'_{i,j}$ closest to 1}
 4:     Build vector $\mathbf{R}$, s.t. $R_j = x'_{\hat{i},j}, \forall j$;      {Frac. assoc. freed by $x'_{\hat{i},\hat{j}}$ rounding}
 5:     Build vector $\mathbf{D}$, s.t.
        $D_j = \{\|x'_{i,j}\| \mid i \neq \hat{i}, (i,j) \in X^n, r_{i,j} > 0\}, \forall j$;
 6:     Set $x'_{i,j} = x'_{i,j} + R_j/D_j, \forall (i,j) \in X^n$,
        s.t. $i \neq \hat{i}, r_{i,j} > 0$;      {Update $x'_{i,j}$ not already rounded}
 7:     Set $X^{n+1} = X^n \setminus \{(\hat{i},j) \mid 1 \le j \le M\}$;      {Remove $(\hat{i},j)$ for rounded $x'_{\hat{i},j}$}
 8:     Set $x'_{\hat{i},\hat{j}} = 1$;      {Round to 1 the selected $x'_{\hat{i},\hat{j}}$}
 9:     Set $x'_{\hat{i},j} = 0, \forall j \neq \hat{j}$;      {Ensure $\hat{i}$ is associated only to AP $\hat{j}$}
10:     Set $n = n + 1$;
11: until $X^n \neq \varnothing$      {Rounding complete}
12: Set $\mathbf{x}^* = \mathbf{x}'$.      {$x^*$ is a solution of problem (4)}
```

value R_j/D_j, where R_j (line 4) is the fractional association freed on AP $j, j \neq \hat{j}$, by the rounding of $x'_{\hat{i},\hat{j}}$, and D_j (line 5) is the number of still-to-be-rounded clients that could associate to AP j. This update is designed in order to satisfy constraint (9), i.e. $\sum_{j=1}^{M} x'_{i,j} = 1, \forall i$.

In Sec. 6 we demonstrate that by solving the relaxed optimisation problem and subsequently applying our rounding algorithm, we achieve substantial improvements as compared to SNR-based association control mechanisms for mmWave under saturation condition. In what follows, we address utility maximisation under finite load circumstances, i.e. when stations do not always have traffic to transmit.

5 Utility Maximisation for Finite Load Scenario

In this section we consider the general finite load scenario where each client i has an offered load λ_i. With the introduction of the parameter λ_i, the definitions of throughput $S_{i,j}$ and airtime $t_{i,j}$ given in (3) and (2) need to be revisited, since the airtime allocated to client i when associated to the AP j is now also a function of the client's offered load (λ_i) and that of the other clients associated to the same AP j. This effectively means the airtime $t_{i,j}$ becomes a variable of the optimisation problem, which we formalise as:

$$\max_{\mathbf{x},\mathbf{t}} U := \sum_{j=1}^{M} \sum_{i=1}^{N} x_{i,j} \log t_{i,j} r_{i,j}, \tag{12}$$

$$\text{s.t.} \sum_{j=1}^{M} x_{i,j} = 1, \forall i; \qquad \text{(single AP association)} \tag{13}$$

$$\sum_{i=1}^{N} x_{i,j} t_{i,j} \leq h_j T_j, \forall j; \qquad \text{(airtime feasibility)} \tag{14}$$

$$x_{i,j} t_{i,j} r_{i,j} \leq \lambda_i, \forall i, j; \qquad \text{(load feasibility)} \tag{15}$$

$$- t_{i,j} \leq 0, \forall i, j; \tag{16}$$

$$x_{i,j} \in \{0, 1\}, \forall i, j. \tag{17}$$

Finding a solution to the above involves solving two different problems in parallel, namely:

1. Finding the best association matrix $\mathbf{x}$ as in the case of the saturation scenario, and
2. Finding the best airtime allocation $t_{i,j}$ that takes into account the load requirements λ_i, while providing some form of fairness.

To accomplish these tasks, we propose an approach that combines simulated annealing and water filling algorithms, and subsequently show that this achieves remarkably higher throughput performance in comparison with the recent DAA scheme [11] and the default SNR-based association policy.

5.1 Simulated Annealing and Water Filling

The underlying principle behind solving the problem defined by (12)–(17) is the following: first we assume saturation conditions and use the method described in Sec. 4 to find an initial integer association matrix $\mathbf{x}^*$. We use $\mathbf{x}^*$ as the starting point ($\mathbf{x0}$) for the simulated annealing algorithm we propose, which we summarise in Algorithm 2 and detail next. Note that choosing the starting point in this way ensures a solution is found significantly faster, as compared to when using the outcome of the SNR-based association instead. This is particularly true when offered loads λ_i are moderate–high, since $\mathbf{x}^*$ is usually close to the best solution found by our heuristic, as revealed by analysing multiple topologies.

After the initialisation steps (lines 1–2), the simulated annealing algorithm enters a loop (lines 3–22) which is executed until the parameter T, called temperature, exceeds a certain minimum $Tmin$.[5] The temperature is decremented at each iteration of the loop (line 20) with a step proportional to a parameter α, which controls the speed of the algorithm and the granularity of the temperature values ($0 < \alpha < 1$).

[5] We discuss the proposed simulated annealing parameters in Sec. 6.

Algorithm 2 SimulatedAnnealing $(\mathbf{x0}, \mathbf{r}, \lambda, \mathbf{h}, T0, Tmin, \alpha, q, p)$

```
1:  Set $\mathbf{x} = \mathbf{x0}, T = T0, v = 1$
2:  Set $\mathbf{t} = WaterFilling(\mathbf{x}, \mathbf{r}, \mathbf{h}, \lambda)$
3:  **repeat**
        # *Stabilisation loop for a given temperature $T$*
4:      **for** $k = 1$ **to** $k = q$ **do**
5:          **if** $x_{i,j} t_{i,j} r_{i,j} = \lambda_i, \forall i, j$ s.t. $x_{i,j} = 1$ **then**
6:              Return $\mathbf{x}$ and $\mathbf{t}$                          {*Optimal solution found*}
7:          **end if**
8:          Set $\mathbf{x}' = Perturbate(\mathbf{x}, \mathbf{t}, \lambda, p)$
9:          Set $\mathbf{t}' = WaterFilling(\mathbf{x}', \mathbf{r}, \mathbf{h}, \lambda)$
10:         Set $\Delta E = U(\mathbf{x}', \mathbf{t}') - U(\mathbf{x}, \mathbf{t})$            {*Compute energy*}
11:         **if** $\Delta E > 0$ **then**
12:             Set $\mathbf{x} = \mathbf{x}'$, $\mathbf{t} = \mathbf{t}'$                    {*Better solution found*}
13:         **else**
14:             Set $y$ to a random value $\in [0, 1)$
15:             **if** $y < e^{\Delta E / T}$ **then**
16:                 Set $\mathbf{x} = \mathbf{x}'$, $\mathbf{t} = \mathbf{t}'$                {*Accept worse solution*}
17:             **end if**
18:         **end if**
19:     **end for**
20:     Set $T = T\alpha^v$                                          {*Update temperature*}
21:     Set $v = v + 1$
22: **until** $T > Tmin$
23: Return $\mathbf{x}$ and $\mathbf{t}$
```

For each temperature value T, the algorithm enters a second, stabilisation loop (lines 4–19), which explores the solution space (including the initial association matrix $\mathbf{x0}$ and the corresponding airtime allocation computed in line 2). In line with standard practice, the inner loop is repeated a number of times q proportional to the size of the problem the algorithm attempts to solve. In our case we set $q = \lceil NM/2 \rceil$.

Then for every iteration of the stabilisation loop, the algorithm checks if the current solution, given by the association matrix $\mathbf{x}$ and the corresponding airtime allocation $\mathbf{t}$, is able to satisfy the offered load λ_i of each client i (line 5). If the condition is satisfied, the algorithm terminates, returning $\mathbf{x}$ and $\mathbf{t}$. This effectively means the solution is an optimum of the problem (12)–(17) and no other solution that does better would be found, given that all offered loads are satisfied.

If the current solution is not optimal, the algorithm calls a perturbation function to generate a new association matrix $\mathbf{x}'$ (line 8). The perturbation function, whose implementation is domain dependent, is a key component of simulated annealing, as it defines the way in which the solution space is explored. In our case, this generates a neighbour association matrix $\mathbf{x}'$ starting from the current one $\mathbf{x}$. To increase the chances of finding a good solution, the perturbation function must be designed to satisfy the *irreducibility property*, i.e. for a number of iterations that tends to infinity, the starting point of the simulated annealing should not influence the final result [32]. As such, the perturbation function must introduce some form of randomness when generating a neighbour of the current solution $\mathbf{x}$. However, it can also employ user-defined rules to prioritise the generation of particular neighbours over other candidates. As we will discuss later, for the association problem at hand, we propose a perturbation function that prioritises the offloading of the bottleneck APs.

Given the new association matrix $\mathbf{x}'$ the algorithm uses the water filling procedure in Algorithm 3 to compute the appropriate airtime allocation $\mathbf{t}'$ (line 9). As we will detail below, this procedure implements an airtime-based water filling algorithm, which returns an airtime allocation,[6] that satisfies the the max-min fairness criterion [33].

Finally, the algorithm computes the energy ΔE (line 10) as the difference between the utility obtained

[6] Note that the water filling procedure returns a vector $\mathbf{t}$, in which each element $t_{i,j}$ represents the fraction of superframe time allocated to client i when associated to AP j. This means the elements of $\mathbf{t}$ returned by Algorithm 2 must be translated into actual airtimes, before being used in our analysis. This is obtained through a simple conversion, i.e. $t_{i,j} \leftarrow t_{i,j}(T_j - O_j)$.

using the new solution $(\mathbf{x}', \mathbf{t}')$ and respectively the previous one, according to (12). If the energy is positive, this means the new solution is better and thus must be kept (line 12). Otherwise, we keep the new solution only with a probability $e^{\Delta E/T}$ that depends on the current energy ΔE and temperature T.

The water filling based airtime allocation procedure invoked at line 2 is outlined in Algorithm 3. First, this computes the airtime $t^{\lambda}_{i,j}$ required to satisfy the load λ_i for each client i and each AP j (line 2), and then loops over all the APs (lines 3–17) as follows. Three sets, A_j, A'_j, and A''_j (line 4) maintain the list of clients that must be associated to AP j (i.e. $x_{i.j} = 1$) and whose corresponding airtimes $t_{i,j}$ have not been set yet; the list of clients with allocated airtimes; and the list of clients who can only be allocated a fraction of the airtime required to satisfy their load. In addition, the residual airtime still available at AP j is maintained in $\hat{h}_j$ (line 5).

Algorithm 3 WaterFilling $(\mathbf{x}, \mathbf{r}, \mathbf{h}, \lambda)$

```
1: Set $t_{i,j} = 0, \forall j$ and $f = 0$
2: Set $t^{\lambda}_{i,j} = h_j \lambda_j / r_{i,j}, \forall i, j$
3: for $i = 1$ to $i = M$ do {Loop on APs}
4:    Define $A_j = \{i \mid x_{i,j} = 1, \forall i\}$, $A'_j = \varnothing$ and $A''_j = \varnothing$
5:    Set $\hat{h}_j = h_j$                                        {AP $j$ residual time}
6:    repeat {Loop on clients associated to AP $j$}
7:       Set $f = \hat{h}_j / (\|A_j\| + \|A''_j\| - \|A'_j\|)$
      {Try to satisfy load requested by the "easier" client $\hat{i}$}
8:       Set $\hat{i} = i$, s.t. $t^{\lambda}_{i,j} = \min_{i \in Aj} t^{\lambda}_{i,j}$
9:       if $t^{\lambda}_{\hat{i},j} < f$ then {Client load can be satisfied}
10:         Set $t_{i,j} = t^{\lambda}_{\hat{i},j}$, $\hat{h}_j = \hat{h}_j - t^{\lambda}_{\hat{i},j}$
11:         Set $A_j = A_j \setminus \{\hat{i}\}$, $A'_j = A'_j \cup \{\hat{i}\}$
12:      else {Client load can not be satisfied}
13:         Set $A_j = A_j \setminus \{\hat{i}\}$, $A''_j = A''_j \cup \{\hat{i}\}$
14:      end if
15:   until $A_j \neq \varnothing$
16:   Set $t_{i,j} = f, \forall i \in A''_j$ {Set time $f$ for unsatisfied clients}
17: end for
18: Return $\mathbf{t}$
```

Then an inner loop (lines 6–15) first computes the fraction of equal airtime f that can be assigned to each client i (line 7) and selects from A_j the index of the client i whose corresponding $t^{\lambda}_{i,j}$ is the minimum among all set members; i.e. it searches the client whose load request is the easiest to satisfy. If the time required to satisfy client i's load $(t^{\lambda}_{i,j})$ is less than the fraction of airtime available to that client (f), it means AP j can completely satisfy that request (lines 9–11). Therefore the airtime allocated to client i associated to AP j is set to $t^{\lambda}_{i,j}$, the residual time $\hat{h}_j$ available at AP j is updated, the current index i is removed from set A_j and inserted in A'_j. If instead the fraction of available airtime (f) is insufficient to satisfy the load request, the current index i is removed from A_j and inserted in A''_j (lines 12–14). Finally, an equal slice of the residual airtime is assigned to each client in A''_j (line 16).

We conclude this section with a brief description of the perturbation function summarised in Algorithm 4, whose key objective is to prioritise the offloading of the network bottlenecks. Given an association matrix $\mathbf{x}$ and an airtime allocation matrix $\mathbf{t}$, we define the bottleneck value B_j for each AP j as:

$$B_j = B^{load}_j - B^{time}_j \tag{18}$$

where

$$B^{load}_j = \sum_{i=1}^{N} x_{i,j} \lambda_i - \sum_{i=1}^{N} x_{i,j} r_{i,j} t_{i,j} \tag{19}$$

$$B^{time}_j = (T_j - O_j) - \sum_{i=1}^{N} x_{i,j} r_{i,j} t_{i,j} \tag{20}$$

$B_j^{load} \geq 0$ is the difference between the total load request of clients i associated to AP j and the amount of requested load AP j is able to satisfy. We have that $B_j^{load} = 0$ when the AP can completely satisfy the load request. Instead, $B_j^{time} \leq 0$ is the difference between the total airtime available in a super–frame at AP j and the airtime consumed by the associated clients. We also have that $B_j^{time} = 0$ when $B_j^{load} > 0$ and that $B_j^{time} \geq 0$ when $B_j^{load} = 0$. From (18)–(20) it is easy to observe that AP j is a network bottleneck when $B_j \geq 0$ and it is not when $B_j < 0$. The perturbation function starts by building the following sets:

$$\begin{aligned} B^- &= \{j \mid B_j < 0\}, \forall j; \\ B^+ &= \{j \mid B_j \geq 0\}, \forall j, \end{aligned} \tag{21}$$

where B^- and B^+ contain indexes j of the APs that are not, and respectively are bottlenecks.

Then the algorithm selects a random client i with probability p, and moves it to a different AP (lines 4–6). This introduces the randomness required for satisfying the *irreducibility property*. On the other hand, the algorithm tries to reduce network bottlenecks with a probability $1 - p$, moving a random client from a bottlenecked AP to one with available resources (lines 8–10). If all the APs are bottlenecks, the algorithm moves a random client from an AP j to a different AP $\hat{j}$ with $B_{\hat{j}} < B_j$.

6 Performance Evaluation

In this section we evaluate the performance of the proposed association control algorithms under different traffic load conditions, considering enterprise environments where clients are within coverage of multiple APs and encounter different link qualities. We compare the performance of our solutions in terms of individual and total throughput, as well as network utility, with that of the 802.11ad standard SNR-based policy, that of a greedy association algorithm whereby APs take turns in associating the nearest clients, and respectively that of the recent DAA scheme [11]. We further compare the performance of all approaches in small topologies, with that of the global optimum we obtain through exhaustive search. Subsequently, we evaluate the average performance attained by the proposed and existing schemes, when stations' position evolve according to a random waypoint mobility model. Lastly, we demonstrate the short runtime of the combined simulated annealing and water filling mechanism we propose.

Algorithm 4 $\mathbf{x}' = \mathit{Perturbate}(\mathbf{x}, \mathbf{t}, \lambda, p)$

```
 1: Let $\mathbf{x}' = \mathbf{x}$
 2: Builds sets $B^-$ and $B^+$ as defined by (21).
 3: Set $y$ to a random value $\in [0, 1)$
 4: if $y < p$ then {irreducibility property}
 5:   Choose random $(i, j)$, s.t. $x_{i,j} = 1$
 6:   Choose random $\hat{j}$, s.t. $\hat{j} \neq j, r_{i,\hat{j}} > 0$
 7: else {Bottlenecks offloading}
 8:   if $B^- \neq \varnothing$ then
 9:     Choose random $(i, j)$, s.t. $x_{i,j} = 1, j \in B^+$
10:     Choose random $\hat{j}$, s.t. $\hat{j} \in B^-, r_{i,\hat{j}} > 0$
11:   else
12:     Choose random $(i, j)$, s.t. $x_{i,j} = 1, B_j \neq \min_{j^*} B_{j^*}$
13:     Choose random $\hat{j}$, s.t. $r_{i,\hat{j}} > 0, B_{\hat{j}} < B_j$
14:   end if
15: end if
16: Set $x'_{i,j} = 0$
17: Set $x'_{i,\hat{j}} = 1$
```

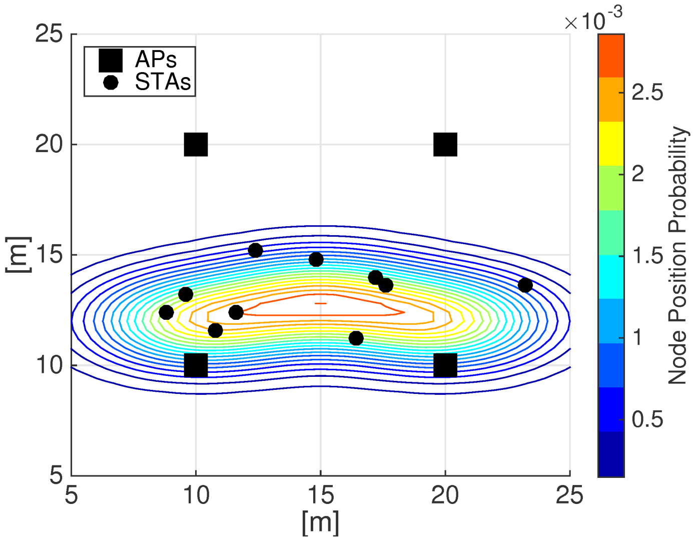


Figure 4: Probability mass function used for extracting the positions of the stations in 4APs/10STAs scenarios used for evaluation. Probability increases from blue ($\approx 0.2 \cdot 10^{-3}$) to red ($\approx 2.9 \cdot 10^{-3}$).

6.1 Simulation Environment

As mmWave platforms suitable for large scale experimentation are yet to appear [34], to evaluate our proposal we develop an NS-3[7] simulation module that implements closely the 802.11ad protocol details [3]. We employ the isosceles cone antenna pattern defined in [35], which can be steered to an arbitrary angle and whose elevation and azimuth are functions of the gain. In our simulations we configure the antenna gain to 15dB, and note this model was used successfully by Halperin *et al.*, who also developed a basic NS-3 implementation of the 802.11ad physical layer [36]. We used their code as a starting point for our own implementation, adding the missing Directional Multi-Gigabit (DMG) PHY capabilities and the scheduled Service Period (SP) based MAC within the Data Transmission Interval (DTI), as illustrated in Fig. 2. Similarly to [36], we compute the SINR for different parts of the frames, combining the power from multiple interferers and noise, and model free space propagation using Friis law. We consider indoor deployments with ceiling mounted mmWave APs, in which simulation results we obtain reveal that the collision rates are below 0.003. This confirms the validity of the pseudo-wired link assumption used. To estimate the Bit Error Rate (BER), we used the receiver sensitivity specified by the standard (table 21-3 in [3]). While advanced channel modelling is outside of this work, the assumptions we make are appropriate for indoor scenarios with finite coverage and number of access points. By allowing for arbitrary SNRs on client–AP links, we decouple the client association and airtime allocation tasks at the core of our contribution, from the environment dependent (e.g. reflections, obstacles, etc.) PHY channel properties already documented [37, 19].

Recall that an SP is allocated for contention-free access between a client and an AP, without carrier sensing. We implement A-MPDU aggregation for efficient transmission (up to 64 frames in a single A-MPDU) and the Block Ack mechanism.

Since the standard does not specify a beamforming training mechanism, we use a conservative 10% overhead for this procedure, noting that performance gains will remain unchanged with other values, and assume negligible beam switching overhead. The standard neither mandates a specific rate control algorithm, therefore we implement a rate controller that selects the best transmission MCS based on the SINR measured at the receiver. We assume that each clients i can estimate the rates $r_{i,j}$ towards each AP j by measuring the SINR of the beacons the APs transmit periodically on each antenna sector. Finally, we extend NS-3 to enable automatic generation of network topologies and rapid configuration of 802.11ad WLANs.[8]

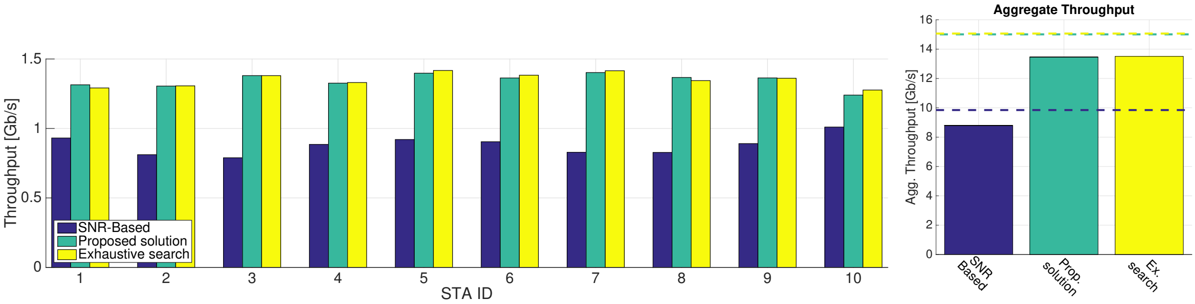
For evaluation purposes we consider two deployment scenarios. The first is an indoor 24m×20m area, where four ceiling mounted mmWave APs are placed in a square layout, as depicted by the black squares

[7] NS-3 discrete-event network simulator, `https://www.nsnam.org/`

[8] The source code of our NS-3 simulation module is publicly available at `https://bitbucket.org/uoeunibs/11ad-for-ns3`

Simulation Environment	
Antenna Model	Cone pattern
Antenna gain	15 dBm
Antenna beamwidth	$\sim 41^{\circ}$
Channel access	Service Periods based
Propagation model	Free space (Friis law)
Bit Error Rate	Receiver sensitivity (table 21-3 in [3])
A-MPDU aggregation	64 frames
DMG PHY	OFDM up to 6.756Gb/s
Beamforming training overhead	10%
Rate controller	SNR-based
Traffic type	1470-byte UDP downlink packets
Evaluated scenarios	Backlogged Traffic Finite Load Conditions User Mobility
Considered metrics	Individual throughput Total throughput Network utility
Evaluated algorithms	Proposed solutions SNR-based (Eq. and Water fill airtime) DAA scheme
Deployments	24m20m area, 4 APs, 10 clients 30m30m area, 9 APs, 30 clients

Table 1: Simulation environment summary



(a) Individual throughputs achieved with the SNR-based and proposed methods.

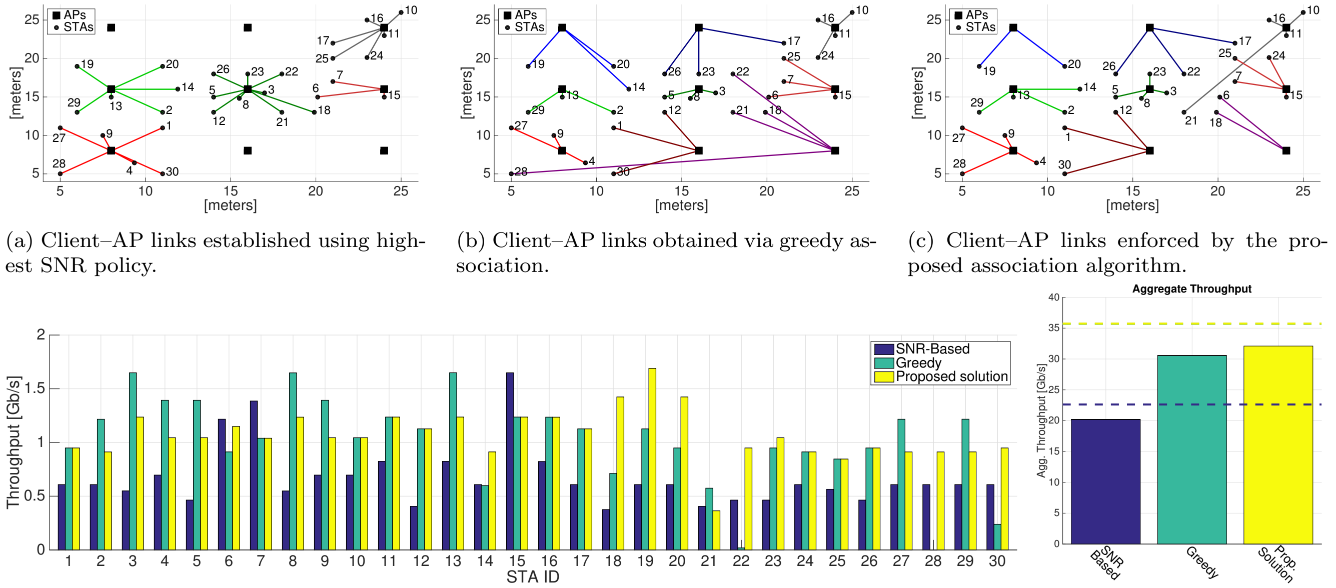
(b) Aggregate throughput comparison.

Figure 5: Enterprise mmWave network with 4 APs positioned on a grid and 10 client stations deployed using the pmf in Fig. 4. Client throughput performance attained with the SNR-based policy, the proposed association control mechanism and respectively exhaustive search. All clients are backlogged (saturation conditions) and equal airtime allocation is performed at each AP. Theoretical maximum shown with dashed lines. Simulation results.

in Fig. 4. Ten client stations are randomly distributed by extracting their positions using the probability mass function shown as a contour plot in the same figure. The probability decreases from red to blue, with the maximum probability of $\approx 2.9 \cdot 10^{-3}$ centered at coordinate $(15, 13)$, and the edge contour line corresponding to a probability of $\approx 0.2 \cdot 10^{-3}$. Note that the probability is never zero and there is a low chance to extract positions outside the outer contour line.

To obtain average results of measured throughputs with good statistical significance, we consider a total of 30 different deployments of this type, and compute the average individual throughput in each case over three simulation runs (i.e. 90 simulations in total). The black circles in Fig. 4 show an example of client locations used in simulation. The clients transmit at PHY rates between 693Mb/s and 6.756Gb/s, depending on their relative distance to the APs within range [3]. Given the small number of APs and stations in this first scenario, we will also compute the optimal solution of the problems (4) and (12) using exhaustive search.

In the second scenario we consider a more complex 30m×30m indoors deployment, where 9 ceiling mounted mmWave APs are placed on a grid layout, as depicted in Fig. 6a, and serve 30 randomly placed client stations. Here, the results we report are the averages computed over ten repetition of the simulation. Given the number of APs and client stations in this topology, finding the absolute optimum through exhaustive search is no longer feasible. In all scenarios APs transmit 1470-byte UDP packets in the downlink. Table 1 summarises the simulation parameters and scenarios we consider for evaluation.



(a) Client–AP links established using highest SNR policy.

(b) Client–AP links obtained via greedy association.

(c) Client–AP links enforced by the proposed association algorithm.

(d) Individual throughputs achieved with the SNR-based, greedy, and proposed methods.

(e) Aggregate throughput comparison.

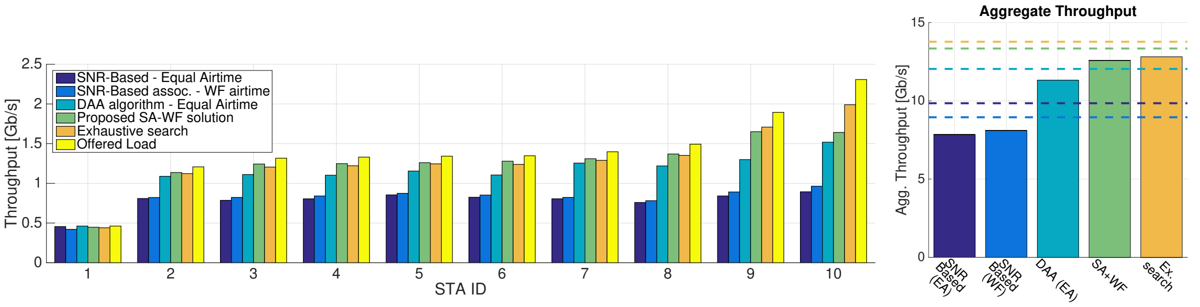
Figure 6: Enterprise mmWave network with 9 APs positioned on a grid and 30 client stations deployed randomly. Client associations enforced by and throughput performance attained with the SNR-based policy, the greedy association approach, and respectively the proposed association control mechanism. All clients are backlogged (saturation conditions) and equal airtime allocation is performed at each AP. Theoretical maximum shown with dashed lines. Simulation results.

6.2 Backlogged Traffic

We first investigate the performance of the association control scheme we propose for saturation scenarios, which involves solving the relaxed utility maximisation problem and executing an iterative rounding algorithm (Sec. 4). For the scenario with 4 APs, we compare the behaviour of our approach against the standard's default SNR-based association control policy and against the optimal solution of problem (4) obtained through exhaustive search. For the scenario with 9 APs, we compare the throughput performance of our approach only against the standard's default policy and the greedy association previously described. For a fair comparison, in all cases we allocate equal airtime to all clients of each AP (proportional fairness), with any association schemes.

We first illustrate in Fig. 5a the individual station throughputs achieved in the first scenario, where observe that with our scheme all the clients attain superior (up to 74% higher) throughput, as compared to the SNR-based policy. Overall, the proposed solution achieves a 1% utility gain, which corresponds to a 53% increase in aggregate network throughput, which we show in Fig. 5b. In addition, we run exhaustive searches over the solution space, to quantify the difference between the solution found by solving the relaxed problem and running the proposed iterative rounding algorithm, and the absolute optimum (yellow bars in Fig. 5). Observe that this difference is negligible – our method attains small throughput gains (up to 1.7%) for a subset of stations, and slightly lower (up to 2.9%) values for others. These small differences are mainly due to practical channel conditions that lead to frame collisions or reception failure, which are overlooked by the theoretical problem formulation. Overall, the utility loss of our method is only 0.0002%, which corresponds to an aggregate network throughput loss of 0.0035%. To add further perspective, in Fig. 5b we also plot with dashed lines the theoretical maximum throughput obtained numerically as a function of the optimal solution returned by each approach considered, noting only a 10% difference in all cases.

Turning attention to the 9 APs scenario, to gain a deeper understanding of the individual and aggregate throughput performance, we also illustrate the Client–AP links established using the proposed, SNR-based, and greedy methods. Note that using the SNR-based approach, client stations are clustered around the nearest AP, irrespective of their local density, as depicted in Fig. 6a. In effect, they can employ superior PHY bit rates, but often share a single AP with many others (e.g. 9 clients connected to the AP in the centre of the grid), while a subset of APs remain unutilised (4 APs in this case). The greedy approach fairs better as it distributes the load among access points, yet this is performed naïvely,



(a) Offered loads and throughputs with SNR-based approaches, DAA, and the proposed solution.

(b) Aggregate throughput comparison.

Figure 7: Enterprise mmWave network with 4 APs positioned on a grid and 10 client stations deployed using the pmf reported in Fig. 4. Client throughput performance attained with DAA [11], SNR-based policies, the proposed simulated annealing and water filling (SA-WF) based association control solution, and respectively exhaustive search. Clients have heterogeneous offered loads between 0.4–2.3Gb/s (finite load). Theoretical maximum shown with dashed lines. Simulation results.

which leads to half (or less) the throughput performance of the SNR-based strategy for some clients (e.g. clients 22 and 30). In contrast, our approach distributes clients among all APs with the goal of maximising network utility (sum of log throughputs), as shown in Fig. 6c. As such, clients may transmit at lower PHY rates, but are allocated more airtime, which translates into higher throughput.

Indeed, Fig. 6d demonstrates that with our scheme the majority of clients attain superior throughput performance (even >100% higher), while only a small fraction experience a minor performance hit, as compared to the SNR-based policy. Overall, our proposal attains a 2.5% utility gain, which corresponds to a 60% gain in the aggregate network throughput, as illustrated in Fig. 6e.

We conclude that **our scheme achieves substantial performance improvements under backlogged traffic conditions**. In what follows, we investigate the performance of the mechanism we introduced in Sec. 5 for finite load conditions.

6.3 Finite Load Conditions

Next we extend the performance analysis to finite load conditions, i.e. when stations have limited traffic demands (offered load). We consider the same indoor topologies, but with heterogeneous offered loads, whereby demand varies between 460Mb/s and 2.3Gb/s in the 4 APs scenario, and between 500Mb/s and 1.25Gb/s in the 9 APs scenario. Recall that maximising network utility in such circumstances, requires not only to find the appropriate association matrix, but also the airtimes allocated to each station at each AP (see Sec. 5).

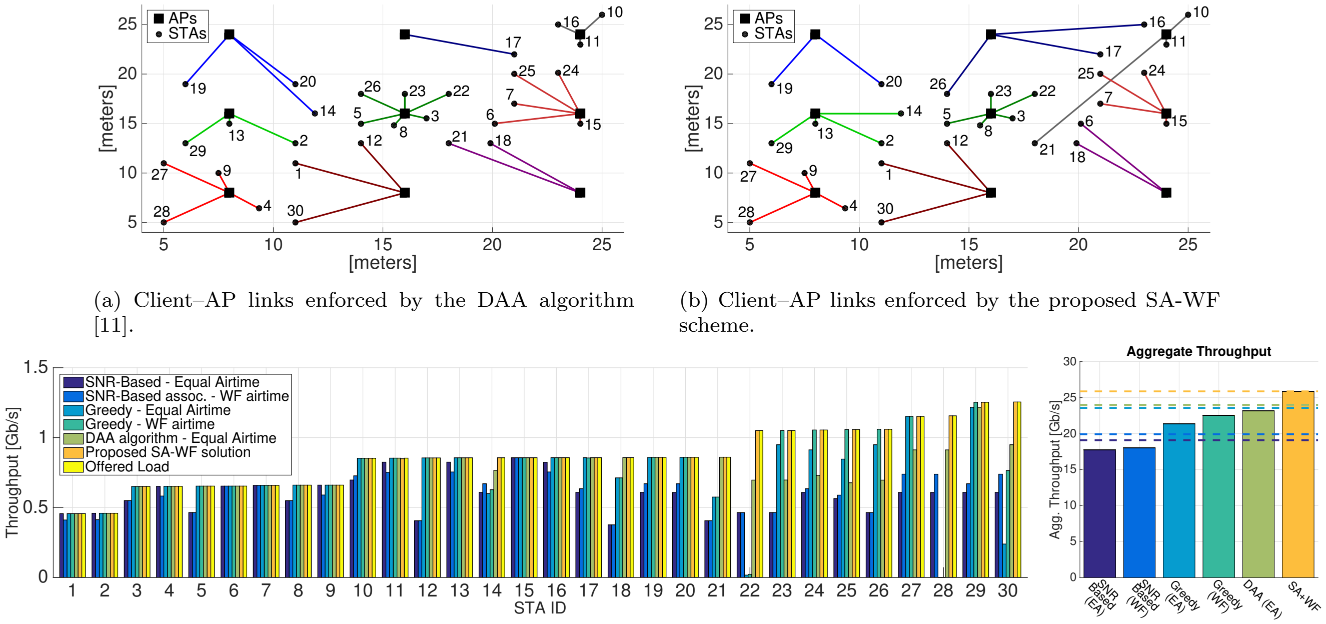
For comparison, we analyse the performance of the proposed simulated annealing and water filling based solution ("Proposed SA-WF solution") and that of:

- SNR-based association and equal airtime (EA) allocation;
- SNR-based association with airtime water filling (WF);
- The distributed DAA algorithm proposed in [11];[9]
- The optimal solution obtained through exhaustive search (only for the 4 APs topology).

Our simulated annealing algorithm works with the following parameters: $T0 = 20, \alpha = 0.7, q = NM/2, Tmin = 0.001$, $p = 0.1$, which we empirically found to yield good performance, as we will show in Sec. 6.6.

We illustrate the results of this experiments if Figs. 7 and 8, where we also show with yellow bars the offered load of each station. In the second case, we also depict the client associations enforced by our proposal (Fig. 8b) and the DAA scheme (Fig. 8a). First, note that also under finite load conditions, our solution largely performs very close to the optimum obtained through exhaustive search (Fig. 7a),

[9] Athanasiou *et al.* only address the association problem and do not consider airtime allocation [11]. As such, we use their approach with the same equal airtime (proportional fair) allocation strategy.



(a) Client–AP links enforced by the DAA algorithm [11].

(b) Client–AP links enforced by the proposed SA-WF scheme.

(c) Offered loads and throughputs with SNR-based, greedy, and DAA approaches, and respectively the proposed solution.

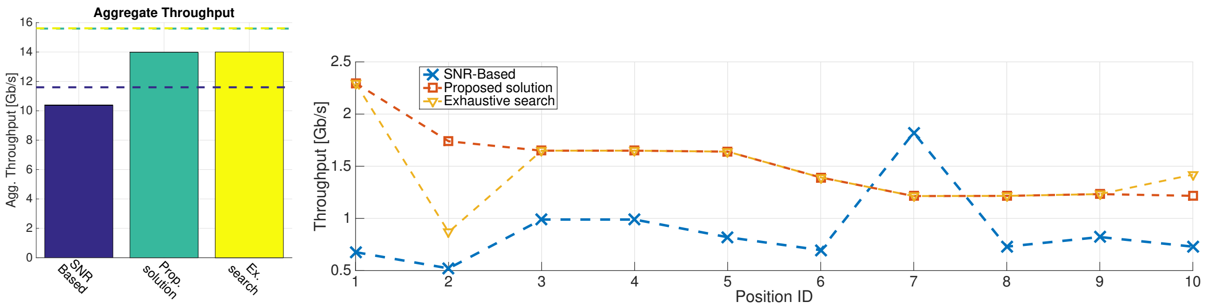
(d) Aggregate throughput comparison.

Figure 8: Enterprise mmWave network with 9 APs positioned on a grid and 30 client stations deployed randomly. Client associations enforced by and throughput performance attained with DAA [11], SNR-based policies, and respectively the proposed simulated annealing and water filling (SA-WF) based association control solution. Also shown is the performance with the greedy association approach working with equal time allocation (EA) and with water-filling (WF). Clients have heterogeneous offered loads between 0.5–1.25Gb/s (finite load). Theoretical maximum shown with dashed lines. Simulation results.

the difference in individual between the two being notable only at stations #10 (21%). Overall, our solution yield a 0.0004% smaller network utility, which corresponds to an aggregate throughput loss of 1.8% (hardly appreciable in Fig. 7b). On the other hand, the aggregate offered load exceeds the resources available in the network, while DAA and the SNR-based policy perform worse that the proposed simulated annealing and water filling based approach. In particular, observe in Fig. 7a that with our scheme almost all the clients attain a superior throughput performance, namely up to 96% and 27% higher, as compared to the SNR-based policy and DAA. Overall, we attain network utility gains up to 2.1%, corresponding to aggregate throughput gains of 11–60%, as illustrated in Fig. 7b.

Examining now the 9 AP topology, we note that DAA works distributively and thus manages to balance well the number of clients across different APs, as seen in Fig. 8a. However, the underlying assumption in this approach is that APs will always be able to accommodate any traffic demand, which is impractical, while airtime allocation at each AP is not considered explicitly. Consequently, although the network has sufficient resources to accommodate all demands in this scenario, some stations only receive a fraction of offered load with this approach (Fig. 8c). In particular, as the offered load increases, DAA largely accommodates only $\sim {}^{2}\!/_{3}$ of the individual demands (stations 22–30). The drawback of not explicitly accounting airtime at each AP is more obvious at the AP located in the bottom left corner. Even if the AP serves the same subset of clients with both the DAA algorithm (Fig. 8a) and the proposed SA–WF scheme (Fig. 8b), two of these clients (27 and 28) experience superior performance with our proposal. This is because our water filling procedure (Algorithm 3) takes into consideration the actual offered loads when allocating airtime to each associated client; in contrast, simply allocating equal airtime to each client with DAA associations proves sub-optimal. For completeness, we also compare the performance of our solution against the greedy approach, which maintains the same associations as shown in Fig. 6b. We consider this performs equal airtime allocation or is combined with the proposed water-filling scheme. Note that neither of the two match the performance of our solution or that of DAA. We remark that a greedy association of clients is only marginally better than an SNR-based approach under finite load conditions.

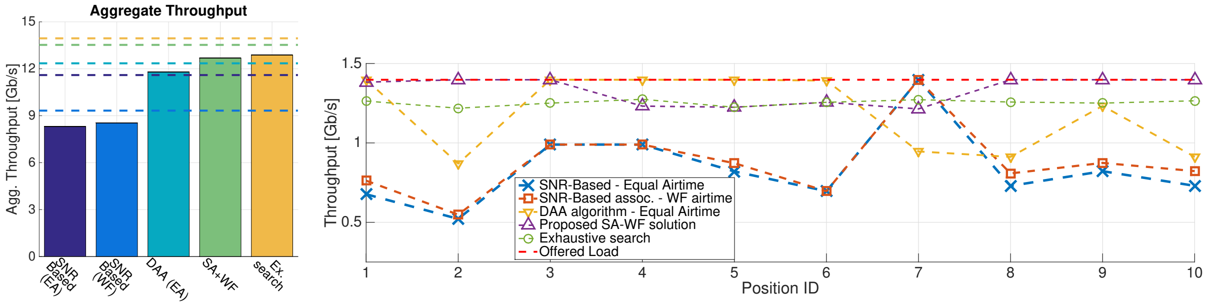
The simulated annealing and water filling based solution we propose obtains association and airtime allocation matrices that **successfully accommodate the demand of all stations (33% more than**



(a) Aggregate throughput comparison.

(b) Throughput of a single mobile station at 10 different locations.

Figure 9: Enterprise mmWave network with 4 APs positioned on a grid and 10 client stations moving following a random waypoint mobility model. Client throughput performance attained with the SNR-based association policy, the proposed solution, and the optimum obtained via exhaustive search. All clients are backlogged (saturation conditions) and equal airtime allocation is performed at each AP. Theoretical maximum shown with dashed lines. Simulation results.



(a) Aggregate throughput comparison.

(b) Offered loads and throughputs for a single station at 10 different locations.

Figure 10: Enterprise mmWave network with 4 APs positioned on a grid and 10 client stations moving using a random waypoint mobility model. Client throughput performance attained with DAA [11], SNR-based policy, the proposed simulated annealing and water filling (SA-WF) based association control solution, and the optimum obtained via exhaustive search. Clients have heterogeneous offered loads between 0.4–2.3Gb/s (finite load). Theoretical maximum shown with dashed lines. Simulation results.

DAA). Overall, our proposal increases network utility attainable with SNR-based policy and DAA by up to 2%, which effectively translates into an aggregate throughput gain of 11–32%, as shown in Fig. 8d.

6.4 User Mobility

In the following we evaluate the performance of the proposed methods when users move in the coverage area and either are backlogged or have finite demands. To this end, we work with a network deployment comprising 4 APs and 10 client stations. The APs are positioned again as in Fig. 4, while clients' positions evolve over time according to a random waypoint mobility model. More specifically, we start by randomly deploying the stations in the $[7, 23]$m$\times[7, 16]$m area, then simulate 100 seconds of user movements with velocities randomly distributed between 0.2m/s and 2.2m/s, pause intervals between movements uniformly distributed between 1s and 20s, and walk times uniformly distributed between 1s and 5s. We take ten snapshots of the client positions obtained with this mobility model (one every 10 seconds) and measure the user throughputs attained at each of these positions. We illustrate the results of this experiments in Figs. 9 and 10, where we plot the throughput of one station as a function of the position, as well as the average of the total (aggregate) throughput over all positions, in backlogged and respectively finite load scenarios.

We observe that, similar to the static scenarios, when stations are backlogged our proposal attains a 1.5% network utility gain, which translates into a 35% aggregate throughput gain. Also in this case, our proposal is very close to the optimal solution, as the difference is network utility is only 0.001% and the aggregate throughput only 0.0012% lower with the proposed scheme. Further, Fig. 9b demonstrates the throughput attained by an individual station with our approach is superior to that with the default

SNR-based policy in nine out of ten positions. It is important to note that, although the SNR-based policy offers higher throughput for the sampled station at position #7, this does not correspond to higher network-wide performance, as confirmed by Fig. 9a. In fact, our scheme offers 34.7% higher aggregate throughput, as compared to the SNR-based association policy.

We now examine a finite load condition scenario and compare the performance of the proposed simulated annealing and water filling based approach, as well as all other aforementioned schemes. Once again, our proposal attains superior results as compared to SNR-based and DAA mechanisms, while its performance is very close to the absolute optimal solution, as shown if Fig. 10. In particular, the simulated annealing and water filling mechanism attains a utility gain up to 2.8% higher than that of SNR-based and DAA approaches, which translates into an aggregate throughput gain of 38–52% (Fig. 10a). As compared to the optimal solution obtained through exhaustive search the network utility loss is limited to 0.003%, corresponding to only 0.012% lower aggregate throughput. Taking a closer look at the throughput of a single station at different locations, Fig. 10b confirms the SNR-based policy only offers superior performance to that of the proposed SA-WF and existing DAA scheme in one location (position #7). Our solution meets the offered load at 6/10 locations and the throughput is very close to that at the other 4/10, unlike DAA which works well in 5/10 locations, but offers significantly lower throughputs at the other 5/10. It is also interesting to note that the optimal solution (dashed green line) obtained through exhaustive search always under-performs in this example. This is easily explained by the fact that, in this particular example, the optimal solution reduces the performance of the sampled station, while improving the performance of other stations, in order to maximise the network utility. This is confirmed by the results we report in Fig. 10a, where we observe the solution found through exhaustive search obtains the highest aggregate throughput.

6.5 Impact of Obstacles

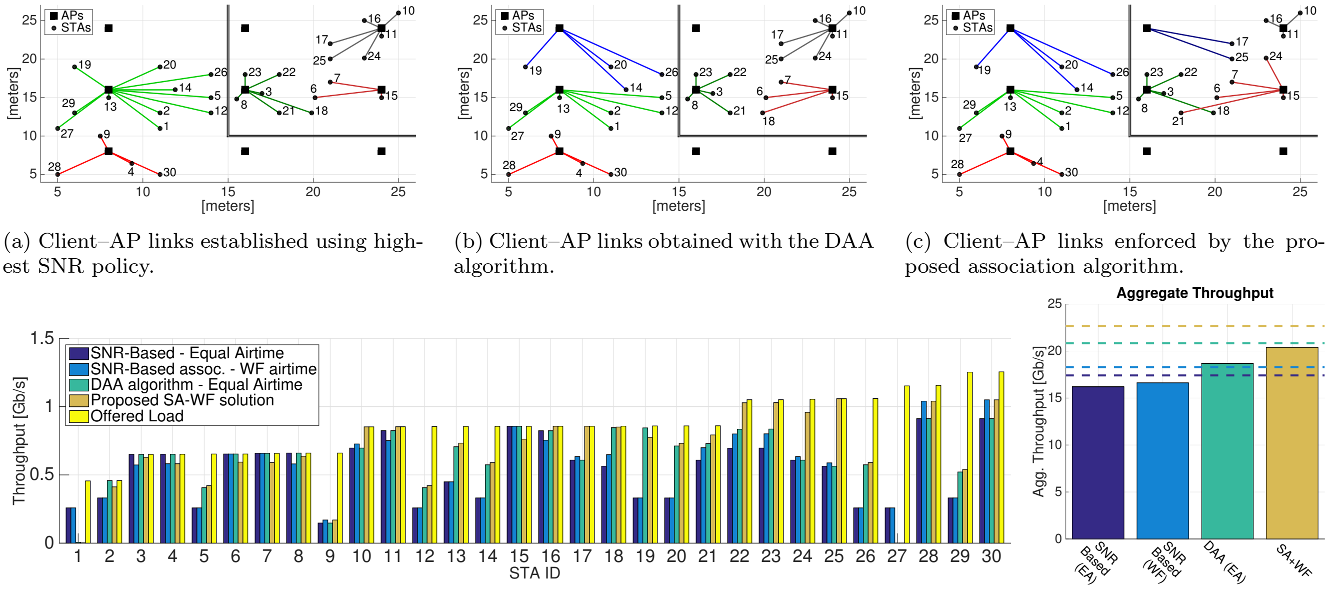
In this subsection we study the impact of obstacles on the association derived and throughput obtained with the proposed simulated annealing and water-filling based scheme, as well as with the SNR-based and DAA benchmarks. We consider a similar office environment which is now partitioned with walls. 9 APs are placed again on a grid and clients are randomly deployed, lying in different parts of the layout as shown in Fig. 11a. In this setting, links between clients and different APs are subject to obstacles. We consider finite heterogeneous load conditions (0.5–1.25Gb/s) and report the behaviour of all approaches in Fig. 11. Observe that the attenuation due to obstacles impacts on the association decision of all schemes, making association to the APs placed in the bottom-right part of the layout particularly problematic. This is indeed observable by comparing Figs. 11a–11c with Figs. 6a, 8a, and 8b. As a result the offered load of fewer clients can be satisfied, which results in overall lower aggregate throughput for all approaches. Nonetheless, the total throughput attained by the proposed solution exceeds that offered by the benchmarks considered, as seen in Fig. 11e.

6.6 Runtime Performance

Finally, we demonstrate that the proposed simulated annealing algorithm finds a solution rapidly, making it suitable for real-time operation in an enterprise mmWave network with a central controller. To this end we first take a closer look at the algorithm's runtime in the more complex network deployment scenario considered previously (9APs and 30 clients) and examine the utility at each step of the exploration.

As shown in Fig. 12a, the algorithm starts with the solution of the saturation problem, computes the utilities of each explored candidate solution, and in this case accepts all of them, even if the energy is negative (line 16 of Algorithm 2). By this procedure **the algorithm finds an optimum that satisfies all clients' offered loads within only 7 iterations**. This confirms that the simulated annealing parameters $T0, \alpha, q, Tmin$, and p we use are appropriate for the problem at hand.

To add further perspective, we simulate 400 topologies with 30–45 client stations and 9APs placed on a grid as before, and measure the execution time on a PC equipped with an Intel i7 CPU running at 3.1GHz. We plot the empirical CDF of these execution times in Fig. 12b. As expected, the runtime grows linearly with the number of clients, yet the median ranges between 180–355ms, which confirms the practical feasibility of our approach even in very dense settings where dynamics associated with pedestrian mobility can be tracked.



(a) Client–AP links established using highest SNR policy.

(b) Client–AP links obtained with the DAA algorithm.

(c) Client–AP links enforced by the proposed association algorithm.

(d) Individual throughputs achieved with the SNR-based, DAA, and proposed methods.

(e) Aggregate throughput comparison.

Figure 11: Enterprise mmWave network with 2 rooms separated by walls. 9 APs positioned on a grid and 30 client stations deployed randomly. Client associations enforced by and throughput performance attained with SNR-based, DAA, and respectively the proposed simulated annealing and water filling (SA-WF) based association control schemes. Clients have heterogeneous offered loads between 0.5–1.25Gb/s (finite load). Theoretical maximum shown with dashed lines. Simulation results.

7 Related Work

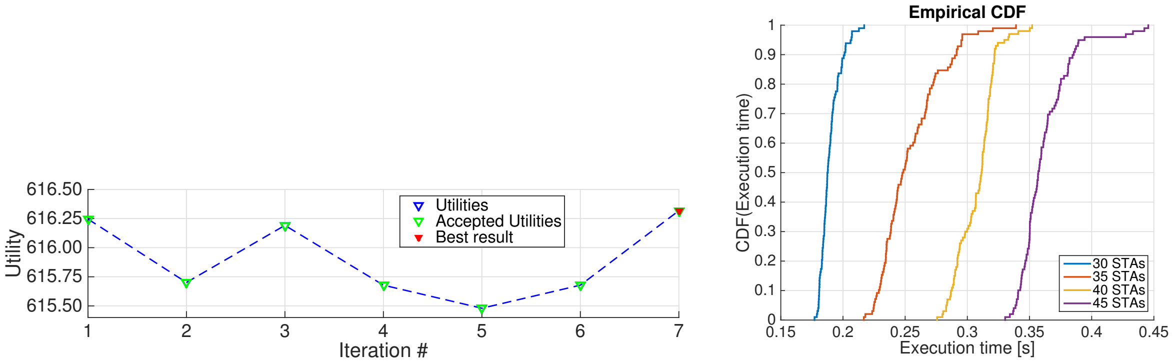
60GHz Characterisation: The feasibility of using mmWave technology for Gbps wireless connectivity has been the focus of several studies in recent years [16, 38, 39, 34]. Singh *et al.* contend that highly-directional links are feasible, but introduce terminal "deafness", shifting the focus from interference management to scheduling [16]. Interference regimes and the impact of mmWave base stations density are studied in [17]. Rappaport *et al.* employ prototype hardware to show steerable directional antennas work well over mmWave frequencies and have potential to support growing consumer data rates [38], while extensive measurements further confirm the feasibility of 60GHz outdoor pico-cells [39]. The feasibility of mmWave technology for top of the rack wireless communication in data centres was also demonstrated experimentally [36, 18]. Characterisation of indoor 802.11ad network performance, interference, and energy consumption has been undertaken in [19, 37]. In addition, software-radio based studies in office environments reveal 802.11ad networks can achieve high throughput coverage beyond a single room [34]. These particularities are a key driver for the association control problem in enterprise mmWave, which we address herein.

Association Control: User association was studied widely in the context of both Wi-Fi [8, 9, 40, 41] and cellular systems [42, 43, 10]. In multi-cell 802.11 networks, game theoretic approaches were employed to balance load [40, 41], while heuristics were proposed for scenarios where legacy clients share the network with high-throughput (802.11n) stations [8]. Optimal association in wireless mesh networks is tackled using an airtime-metric based mechanism in [9].

Son *et al.* address association control in cellular networks by jointly optimising partial frequency reuse and load-balancing schemes [42]. For multi-tier cellular deployments, a theoretical cell association framework is introduced in [43] and upper bounds on the achievable sum rate are derived. Ye *et al.* combine utility maximisation and simple biasing approaches, to achieve user association with load balancing goals [10].

Research into client association in 60GHz networks is sparse. Athansiou *et al.* address this problem from a load balancing perspective [11], though assume APs can always accommodate the demand of all clients, which is impractical, and pursue minimisation of maximum AP utilisation. They do not address distribution of APs' resource among clients.

Unlike previous work, we attack client association in enterprise mmWave networks as a utility max-



(a) Number of iterations required to compute the solution for the topology shown in Fig. 8b

(b) Empirical CDF of the execution time over 400 topologies with different sizes.

Figure 12: Runtime of the proposed simulated annealing algorithm. Simulation results.

imisation problem under both backlogged and finite load scenarios, and heterogeneous link rates. We give low complexity algorithms that achieve close to optimal performance, while ensuring fair airtime allocation at each AP. Our schemes are amendable to deployment on emerging SDN enabled infrastructure supporting 802.11v/k management amendments.

8 Conclusions

In this paper we tackled network utility maximisation in high-end mmWave networks, capturing distinctive terminal deafness and user demand constraints, as well as dissimilar link qualities. Despite inherent NP-completeness, for backlogged conditions we solved a relaxed version of the problem and gave a low-complexity rounding algorithm that attains near-optimal performance. For finite load scenarios, we proposed a mechanism that combines simulated annealing and water filling techniques to find both the optimal association matrix and airtime allocation vector. Using an NS-3 simulation tool we developed, we undertook a comprehensive evaluation campaign and showed that our solutions attain 60% higher throughput as compared to the standard's default SNR-based policy, whilst accommodating the demand of 33% more clients, as compared to recently proposed distributed association algorithms for mmWave networks.

Acknowledgements

The work of F. Gringoli was partly supported by the European Commission (EC) in the framework of the H2020-ICT-2014-1 project WiSHFUL (Grant agreement no. 645274).

References

[1] Y. Garty, Intel and Qualcomm Collaborate to Build Robust 802.11ad Ecosystem (Feb. 2016).
URL `https://tinyurl.com/11ad-sys`

[2] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, P. Popovski, Five disruptive technology directions for 5G, IEEE Communications Magazine 52 (2) (2014) 74–80.

[3] IEEE 802.11ad-Std., Wireless LAN Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band.

[4] Arstechnica, TP-Link unveils worlds first 802.11ad WiGig router (Jan. 2016).
URL `https://tinyurl.com/tp-link-11ad`

[5] Engadget, Acer introduces 'world's first' laptop with 802.11ad WiFi (Apr. 2016).
URL `https://tinyurl.com/acer-wigig`

[6] X. Tie, K. Ramachandran, R. Mahindra, On 60 GHz Wireless Link Performance in Indoor Environments, in: PAM, 2012.

[7] A. Giannoulis, P. Patras, E. Knightly, Mobile Access of Wide-Spectrum Networks: Design, Deployment and Experimental Evaluation, in: Proc. IEEE INFOCOM, Turin, Italy, 2013.

[8] D. Gong, Y. Yang, AP association in 802.11n WLANs with heterogeneous clients, in: Proc. IEEE INFOCOM, Orlando, USA, 2012.

[9] G. Athanasiou, T. Korakis, O. Ercetin, L. Tassiulas, A Cross-Layer Framework for Association Control in Wireless Mesh Networks, IEEE Transactions on Mobile Computing 8 (1) (2009) 65–80.

[10] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, J. G. Andrews, User Association for Load Balancing in Heterogeneous Cellular Networks, IEEE Transactions on Wireless Communications 12 (6) (2013) 2706–2716.

[11] G. Athanasiou, P. C. Weeraddana, C. Fischione, L. Tassiulas, Optimizing Client Association for Load Balancing and Fairness in Millimeter-Wave Wireless Networks, IEEE/ACM Transactions on Networking 23 (3) (2015) 836–850.

[12] F. Kelly, Charging and rate control for elastic traffic, European Transactions on Telecommunications 8 (1) (1997) 33–37.

[13] ONF, Software-Defined Networking (SDN) Definition, accessed Dec. 2016.

[14] IEEE 802.11v-Std., Wireless LAN Specifications Amendment 8: IEEE 802.11 Wireless Network Management.

[15] IEEE 802.11k-Std, Wireless LAN Specifications Amendment 1: Radio Resource Measurement of Wireless LANs.

[16] S. Singh, R. Mudumbai, U. Madhow, Interference Analysis for Highly Directional 60-GHz Mesh Networks: The Case for Rethinking Medium Access Control, IEEE/ACM Trans. Netw. 19 (5) (2011) 1513–1527.

[17] M. Rebato, M. Mezzavilla, S. Rangan, F. Boccardi, M. Zorzi, Understanding noise and interference regimes in 5G millimeter-wave cellular networks, in: Proc. European Wireless, 2016, pp. 1–5.

[18] Y. Zhu, X. Zhou, Z. Zhang, L. Zhou, A. Vahdat, B. Y. Zhao, H. Zheng, Cutting the cord: A robust wireless facilities network for data centers, in: Proc. ACM MobiCom, Maui, Hawaii, USA, 2014, pp. 581–592.

[19] T. Nitsche, G. Bielsa, I. Tejado, A. Loch, J. Widmer, Boon and Bane of 60 GHz Networks: Practical Insights into Beamforming, Interference, and Frame Level Operation, in: Proc. ACM CoNEXT, Heidelberg, Germany, 2015.

[20] R. Enns, M. Bjorklund, J. Schoenwaelder, A. Bierman, Network Configuration Protocol (NETCONF), RFC 6241 (Proposed Standard) (June 2011).

[21] G. R. MacCartney, T. S. Rappaport, A flexible millimeter-wave channel sounder with absolute timing, IEEE Journal on Selected Areas in Communications 35 (6) (2017) 1402–1418.

[22] IEEE 802.11ay Task Group, Enhanced Throughput for Operation in License-Exempt Bands above 45 GHz (2017).
URL `http://www.ieee802.org/11/Reports/tgay_update.htm`

[23] IEEE 802.11az Study Group, Next Generation Positioning (NGP) (2017).
URL `http://www.ieee802.org/11/Reports/ngp_update.htm`

[24] P. Patras, A. Garcia-Saavedra, D. Malone, D. Leith, Rigorous and Practical Proportional-fair Allocation for Multi-rate Wi-Fi, Ad Hoc Networks 36 (2016) 21–34.

[25] C. H. Papadimitriou, K. Steiglitz, Combinatorial Optimization: Algorithms and Complexity, Dover Publications Inc., 2000.

[26] S. Arora, B. Barak, Computational Complexity: A Modern Approach, Cambridge University Press, 2009.

[27] L. Lovász, On the ratio of optimal integral and fractional covers, Discrete Mathematics 13 (4) (1975) 383 –390.

[28] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2009.

[29] T. Cover, J. Thomas, Elements of Information Theory (2nd Ed), John Wiley & Sons, Inc., 2006.

[30] F. S. Hillier, G. J. Lieberman, Introduction to Operations Research, McGraw-Hill, 9th Ed, 2009.

[31] A. R. Conn, N. I. M. Gould, P. L. Toint, Trust-region Methods, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[32] P. J. M. Laarhoven, E. H. L. Aarts (Eds.), Simulated Annealing: Theory and Applications, Kluwer Academic Publishers, Norwell, MA, USA, 1987.

[33] D. J. Leith, Q. Cao, V. G. Subramanian, Max-min Fairness in 802.11 Mesh Networks, IEEE/ACM Transactions on Networking 20 (3) (2012) 756–769.

[34] S. Sur, V. Venkateswaran, X. Zhang, P. Ramanathan, 60 GHz Indoor Networking Through Flexible Beams: A Link-Level Profiling, in: Proce. ACM SIGMETRICS, Portland, Oregon, USA, 2015, pp. 71–84.

[35] A. Sheth, S. Seshan, D. Wetherall, Geo-fencing: Confining wi-fi coverage to physical boundaries, in: Proceedings of the 7th International Conference on Pervasive Computing, Pervasive '09, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 274–290.

[36] D. Halperin, S. Kandula, J. Padhye, P. Bahl, D. Wetherall, Augmenting data center networks with multi-gigabit wireless links, SIGCOMM Comput. Commun. Rev. 41 (4) (2011) 38–49.

[37] S. K. Saha, D. G. Malleshappa, A. Palamanda, V. V. Vira, A. Garg, D. Koutsonikolas, 60 GHz indoor WLANs: insights into performance and power consumption, Wireless Networks.

[38] T. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. Wong, J. Schulz, M. Samimi, F. Gutierrez, Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!, IEEE Access 1 (2013) 335–349.

[39] Y. Zhu, Z. Zhang, Z. Marzi, C. Nelson, U. Madhow, B. Y. Zhao, H. Zheng, Demystifying 60GHz Outdoor Picocells, in: ACM MobiCom, 2014.

[40] W. Xu, C. Hua, A. Huang, A Game Theoretical Approach for Load Balancing User Association in 802.11 Wireless Networks, in: IEEE GLOBECOM, 2010.

[41] O. Erçetin, Association games in IEEE 802.11 wireless local area networks, IEEE Trans. Wireless Comms. 7 (2008) 5136–5143.

[42] K. Son, S. Chong, G. D. Veciana, Dynamic association for load balancing and interference avoidance in multi-cell networks, IEEE Trans. Wireless Comms. 8 (7) (2009) 3566–3576.

[43] S. Corroy, L. Falconetti, R. Mathar, Dynamic cell association for downlink sum rate maximization in multi-cell heterogeneous networks, in: Proc. IEEE ICC, 2012, pp. 2457–2461.