Millimeter-Wave User Association and Low-Interference Beam Scheduling (Invited Paper)

Veljko Boljanovic
UCLA
Los Angeles, CA, USA
vboljanovic@ucla.edu

Shamik Sarkar
UCLA
Los Angeles, CA, USA
shamiksarkar@g.ucla.edu

Danijela Cabric
UCLA
Los Angeles, CA, USA
danijela@ee.ucla.edu

ABSTRACT

Due to highly directional communication links, millimeter-wave networks are often considered to be noise-limited. However, in dense networks with a small inter-site distance, a large number of users, and many serving beams, the inter- and intra-cell interference cannot be neglected. In this work, we aim to address this issue by proposing an optimization framework for joint user association and beam scheduling in dense millimeter-wave networks. The framework considers a sub-network of base stations that can serve multiple users at the same time using directional beams. The users with specific rate requirements can also steer multiple beams to enable spatial multiplexing or connect with more than one base station at a time. The framework includes an NP-hard combinatorial optimization problem that aims to maximize the number of associated users with satisfied rate requirements using a minimal number of intelligently scheduled non-interfering beams. We propose a low-complexity sub-optimal algorithm to solve the NP-hard problem, and then we numerically evaluate the proposed solution and demonstrate its advantages over the conventional association approaches that do not manage the interference.

CCS CONCEPTS

• **Networks** → **Network resources allocation**.

KEYWORDS

User association, beam scheduling, interference management, millimeter-wave network

ACM Reference Format:
Veljko Boljanovic, Shamik Sarkar, and Danijela Cabric. 2022. Millimeter-Wave User Association and Low-Interference Beam Scheduling (Invited Paper). In *Proceedings of The 6th ACM Workshop on Millimeter-Wave and Terahertz Networks and Sensing Systems (mmNets '22).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3555077.3556471

1 INTRODUCTION

Due to large available bandwidth, communication at millimeter-wave (mmW) frequencies is seen as the key enabler of high data rates in the fifth generation of cellular systems [1]. However, the use of mmW frequency bands comes at the cost of a higher propagation loss [2], which must be compensated for using a dense network deployment with a small inter-site distance and large antenna arrays at both base stations (BS)s and user equipments (UE)s. Antenna arrays enable high-gain directional communication links between BSs and UEs. When a BS is equipped with multiple radio frequency (RF) chains, it can steer multiple directional beams to serve multiple UEs simultaneously. On the other hand, a UE can leverage its own RF chains to boost the achievable rate through spatial multiplexing or to increase the robustness to the link blockage by connecting to multiple BSs at the same time, which is often referred to as multi-connectivity [3]. In dense mmW networks with a large number of UEs and serving beams, the power of inter- and intra-cell interference can increase significantly, which makes it challenging to satisfy UEs' data rate requirements.

Data rate requirements of UEs will be a crucial characteristic of next-generation cellular networks. For applications like augmented reality, virtual reality, high resolution video streaming, a strict requirement on data rates will be necessary. Thus, unlike the traditional approach of the network sum rate maximization, we must focus on maximizing the number of users whose rate requirements are satisfied.

In this work, we consider a system where the BSs are connected to a centralized processing unit (PU) with high computational power. In this setup, we propose an optimization framework which performs user association and beam scheduling simultaneously in dense mmW networks. In our proposed framework, we aim to: 1) maximize the number of UEs with satisfied rate requirements; 2) use minimal number of beams to do so; 3) intelligently schedule serving beams to avoid inter- and intra-cell interference from the main-lobes. We formulate a combinatorial optimization problem for these objectives and explain that it is NP-hard. Hence, we develop a sub-optimal algorithm to solve it efficiently in polynomial time. Our proposed solution allows moderate mobility of the UEs as we can repeatedly run our polynomial-time algorithm at regular time intervals. Using numerical simulations, we demonstrate that the proposed framework significantly reduces directional interference in the network and leads to a higher number of UEs with satisfied rate requirements compared to existing association schemes.

2 PREVIOUS WORK

The majority of previous work on user association [4–6] and resource allocation [7–12] in mmW networks used single-criterion objectives that maximize the network sum rate.

The work in [4] considered a user association framework which maximizes the rate of enhanced broadband UEs, while guaranteeing

mmNets '22, October 17, 2022, Sydney, NSW, Australia

ACM ISBN 978-1-4503-9509-0/18/06…$15.00
https://doi.org/10.1145/3555077.3556471

Table 1: Notation Summary

N_{BS}, N^{a}_{BS}, $N^{\text{RF}}_{\text{BS}}$	Number of BSs, BS antennas, and BS RF chains
N_{UE}, N^{a}_{UE}, $N^{\text{RF}}_{\text{UE}}$	Number of UEs, UE antennas, and UE RF chains
E_{BS}	Number of BSs that are known to each UE
P_{T}, P_{N}	Transmit power and noise power
f, BW	Carrier frequency and bandwidth
N_0	Power spectral density of noise
d	Inter-site distance
T	Number of time slots
λ	Scaling parameter in multi-criterion objective function
ϵ	Threshold for initial rounding
Q	Total number of optimization variables
$\mathcal{U}$, $\mathcal{B}$, $\mathcal{L}$, $\mathcal{T}$	Sets of all UEs, all BSs, all links, and all time slots
u, b, l, t	Indices for UEs, BSs, links, and time slots
$\mathcal{A}$, $\mathcal{V}_n$, $\mathcal{W}_n$	Sets of associated UEs, UEs that require n links, and UEs that require more than n links
v_u	Required number of links for UE u
R_u, R_{min}, R_{max}	Rate requirement of UE u, minimum possible rate requirement, and maximum possible rate requirement
$\mathbf{H}_{u,b}$	Channel matrix between UE u and BS b
$\mathbf{v}^l_{u,b}$, $\mathbf{w}^l_{u,b}$	Precoder and combiner on link l between UE u and BS b
$I^l_{u,b}$	Interference on link l between UE u and BS b
$c^l_{u,b}$	Capacity on link l between UE u and BS b
$\mathcal{I}_u$	Set of UEs that have at least one of their BS beams in common with UE u
s_u	Association variable that indicates if rate requirement of UE u is satisfied
$x^{l,t}_{u,b}$	Association variable that indicates if link l between UE u and BS b is used in time slot t
$a^t_{u,u'}$	Association variable that indicates if UE u' is allocated at least one BS beam that belongs to UE u in time slot t
$\bar{s}_u$, $\bar{x}^{l,t}_{u,b}$, $\bar{a}^t_{u,u'}$	Values of association variables after initial rounding
$\tilde{s}_u$, $\tilde{x}^{l,t}_{u,b}$, $\tilde{a}^t_{u,u'}$	Temporary values of association variables
$\hat{s}_u$, $\hat{x}^{l,t}_{u,b}$, $\hat{a}^t_{u,u'}$	Feasible values of association variables after proposed alg.

the reliability constraints for low-latency UEs. In [5] and [6], the authors proposed user association schemes with BS load balancing in mmW networks and heuristic algorithms to solve the optimization problems in polynomial time. However, users' rate requirements have not been considered in [4–6].

In mmW networks, resource allocation includes the distribution of beams, frequencies, and power resources. A joint allocation of all of these resources represents a very complex optimization problem. Thus, many previous algorithms aimed to either allocate only a subset of resources or provide a good sub-optimal solution to the joint allocation. In [7], the authors studied mmW beam training in a single-cell scenario and they proposed a scheme for beam allocation and conflict avoidance. While this scheme can eliminate the interference within the considered cell, inter-cell interference remains a problem. In contrast, [11] introduces a recursive polynomial-time algorithm that minimizes the number of inter-cell beam collisions between two BSs. The work in [8] and [9] proposed a sub-optimal sequential allocation of the system resources. System resources can also be allocated through a sub-optimal bisection-based greedy approach, as previously proposed in [10]. A recent work in [12] proposed a near interference-free beam scheduling scheme to minimize the amount of unfulfilled UEs' requirements. However, the authors assumed that the UEs had already been associated with the BSs, which limits the network performance optimization. Additionally, the network scenario considered in [12] is optimistic in terms of the beam interference since it assumes that all UEs have only one

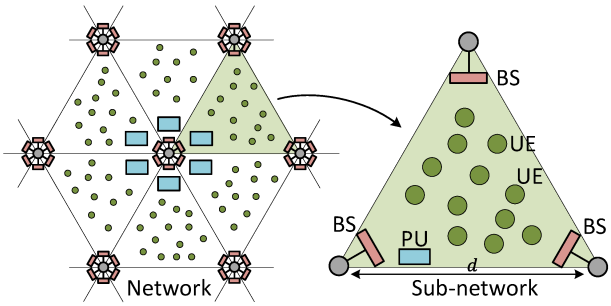


Figure 1: A simple illustration of the sub-network with $N_{\text{BS}} = 3$ and $N_{\text{UE}} = 12$. In general, the sub-network can have irregular shape with different inter-site distances.

RF chain and that they do not support spatial multiplexing and/or multi-connectivity.

In contrast to the previous work, we introduce a new joint user association and low-interference beam scheduling optimization framework that considers multi-RF chain BSs and UEs. The framework has the objective to maximize the number of the UEs with satisfied rate requirements, while suppressing the main-lobe interference in downlink.

3 OPTIMIZATION FRAMEWORK

In this section, we introduce the system model and the proposed user association and beam scheduling framework. Important notation is summarized in Table 1.

3.1 System Model

We consider downlink communication in one part of a mmW cellular network consisting of N_{BS} BSs from the set $\mathcal{B}$ and N_{UE} UEs from the set $\mathcal{U}$. The BSs are connected to a centralized PU with high computational power. A simple illustration of the sub-network with $N_{\text{BS}} = 3$ and $N_{\text{UE}} = 12$ is provided in Fig. 1. All communication links operate at the same carrier frequency within the bandwidth BW. Each BS has N^{a}_{BS} antennas and $N^{\text{RF}}_{\text{BS}}$ RF chains, i.e., it can steer $N^{\text{RF}}_{\text{BS}}$ beams simultaneously. Similarly, each UE has N^{a}_{UE} antennas and $N^{\text{RF}}_{\text{UE}}$ RF chains, where $N^{\text{RF}}_{\text{UE}} < N^{\text{RF}}_{\text{BS}}$. Each UE u has a specific data rate requirement R_u. We assume that each UE has already estimated the channel matrix between itself and E_{BS}, $E_{\text{BS}} \leq N_{\text{BS}}$, closest BSs. Using the control links, the channels estimates are reported to the corresponding BSs, which pass the estimates further to the PU. The channels and their estimates are assumed to be constant over a short period of time, e.g., the next scheduling period.

Let the matrix $\mathbf{H}_{u,b}$ represent the channel estimate between the UE u and BS b. The PU can determine the set $\mathcal{L}$ of $N^{\text{RF}}_{\text{UE}}$ links between u and b by using discrete Fourier transform (DFT) beamfoming matrices to estimate $N^{\text{RF}}_{\text{UE}}$ most dominant propagation directions in the channel. Thus, all transmit DFT beams at the BSs and all receive DFT beams at the UEs are determined through digital processing. The complexity of doing this scales as $O(N_{\text{UE}} N^{\text{RF}}_{\text{UE}} E_{\text{BS}} N^{\text{a}}_{\text{UE}} N^{\text{a}}_{\text{BS}})$. The PU can also identify the interfering (completely overlapping) DFT beams, i.e., conflicted links, at all BSs and UEs.

Let $\mathbf{v}^l_{u,b}$ and $\mathbf{w}^l_{u,b}$ be DFT beams that correspond to the l-th link between u and b. Assuming that the transmit power P_{T} is split between $N^{\mathrm{RF}}_{\mathrm{BS}}$ RF chains, the corresponding capacity $c^l_{u,b}$ is given as follows

$$c^l_{u,b} = \mathrm{BW}\log_2\left(1 + \frac{|\mathbf{w}^{l\mathrm{H}}_{u,b}\mathbf{H}_{u,b}\mathbf{v}^l_{u,b}|^2}{I^l_{u,b} + P_{\mathrm{N}}}\frac{P_{\mathrm{T}}}{N^{\mathrm{RF}}_{\mathrm{BS}}}\right) \tag{1}$$

where P_{N} is the post-combining noise power. With the noise power spectral density N_0, the noise power is defined as $P_{\mathrm{N}} = N^{\mathrm{a}}_{\mathrm{UE}}\mathrm{BW}N_0$. The interference power $I^l_{u,b}$ is calculated as $I^l_{u,b} = \sum_{b',u',l'} \frac{P_{\mathrm{T}}}{N^{\mathrm{RF}}_{\mathrm{BS}}} \times |\mathbf{w}^{l\mathrm{H}}_{u,b}\mathbf{H}_{u,b'}\mathbf{v}^{l'}_{u',b'}|^2$, where $u' \neq u$, b' goes over E_{BS} BSs that are known to the UE u, and l' goes over all links that belong to the UE u' and that are not conflicted with the link l. Note that with this definition of interference power, (1) represents a pessimistic capacity estimate because BSs are unlikely to use all of the links that do not have a conflict with l at once. If the channel between the UE u and BS b is not estimated, the capacity is $c^l_{u,b} = 0,\ \forall l$.

3.2 Proposed Framework

The existing frameworks for user association and resource allocation are often based on maximization of the network sum rate. In such frameworks, the available system resources are mainly allocated to the UEs with high-capacity links. Thus, many UEs may be left out with unsatisfied data rate requirements. Additionally, a large number of serving beams in a dense network can cause significant interference and further reduce the effective data rates. In this paper, we propose a novel framework which aims to maximize the number of users with satisfied data rates using a minimal number of beams (links) that are scheduled in space and time such that the interference from the main-lobes is suppressed.

Let $\mathcal{T}$ be a set of T time slots over which the scheduling is done. Let $\mathbf{s}$ be a vector of N_{UE} binary association variables, where s_u is 1 if the rate requirement of the UE $u \in \mathcal{U}$ is satisfied over the period of T slots, and 0 otherwise. Let $\mathbf{x}$ be a vector of $TN_{\mathrm{UE}}N^{\mathrm{RF}}_{\mathrm{UE}}N_{\mathrm{BS}}$ binary association variables for all links in the network, where $x^{l,t}_{u,b}$ is 1 is the UE u is served by the BS b using the link l in the time slot t, and 0 otherwise.

Based on the system model, the UEs and BSs can steer up to $N^{\mathrm{RF}}_{\mathrm{UE}}$ and $N^{\mathrm{RF}}_{\mathrm{BS}}$ beam simultaneously in any time slot t, which can be expressed as the following constraints

$$(\mathrm{C1}):\ \sum_{b,l} x^{l,t}_{u,b} \leq N^{\mathrm{RF}}_{\mathrm{UE}} s_u,\ \forall u \in \mathcal{U}, t \in \mathcal{T}, \tag{2}$$

$$(\mathrm{C2}):\ \sum_{u,l} x^{l,t}_{u,b} \leq N^{\mathrm{RF}}_{\mathrm{BS}},\ \forall b \in \mathcal{B}, t \in \mathcal{T}. \tag{3}$$

The variable s_u in (C1) ensures that beams are not allocated to the UEs whose rate requirements cannot be satisfied. Based on (C1), it is straightforward to see that a UE can get up to $TN^{\mathrm{RF}}_{\mathrm{UE}}$ links over T time slots.

Each UE has $E_{\mathrm{BS}}N^{\mathrm{RF}}_{\mathrm{UE}}$ potential links and thus $E_{\mathrm{BS}}N^{\mathrm{RF}}_{\mathrm{UE}}$ potential BS beams in every time slot. As described in Section 3.1, the PU can identify the conflicted beams at each BS. Namely, for each $u \in \mathcal{U}$, the PU can create an *interfering group*, i.e., a set $\mathcal{I}_u$ of the UEs that have at least one of their $E_{\mathrm{BS}}N^{\mathrm{RF}}_{\mathrm{UE}}$ BS beams in common with u. Let the BS index b' and the link index l' correspond to a beam that $u',\ u' \in \mathcal{I}_u$, has in common with u. Let $\mathbf{a}$ be a vector of binary association variables, where the element $a^t_{u,u'}$ is 1 if a UE u' is allocated at least one of the BS beams of the UE u in the time slot t, and 0 otherwise. Given the variables $x^{l,t}_{u,b}$ and indices b' and l', the variable $a^t_{u,u'}$ can be mathematically modeled as a logical OR using the following set of constraints:

$$(\mathrm{C3}):\ a^t_{u,u'} \leq \sum_{b',l'} x^{l',t}_{u',b'},\ \forall u \in \mathcal{U}, u' \in \mathcal{I}_u, t \in \mathcal{T}, \tag{4}$$

$$(\mathrm{C4}):\ a^t_{u,u'} \geq x^{l',t}_{u',b'},\ \forall u \in \mathcal{U}, u' \in \mathcal{I}_u, t \in \mathcal{T}, b', l', \tag{5}$$

$$(\mathrm{C5}):\ 0 \leq a^t_{u,u'} \leq 1,\ \forall u \in \mathcal{U}, u' \in \mathcal{I}_u, t \in \mathcal{T}. \tag{6}$$

With this definition of $a^t_{u,u'}$, we can introduce a constraint that suppresses the main-lobe interference at the BS side

$$(\mathrm{C6}):\ a^t_{u,u} + a^t_{u,u'} \leq 1,\ \forall u \in \mathcal{U}, u' \in \mathcal{I}_u, u \neq u', t \in \mathcal{T}. \tag{7}$$

The constraint in (C6) ensures that the BS beams that correspond to the links of the UE u cannot be allocated to both u and $u', u' \in \mathcal{I}_u \backslash \{u\}$, in the same time slot.

Similarly, the PU can identify the conflicted beams at each UE. Let a conflicted beam at the UE u be common for the links defined by the indices b'' and l''. Then the interference suppressing constraint for this beam can is expressed as

$$(\mathrm{C7}):\ \sum_{b'',l''} x^{l'',t}_{u,b''} \leq 1,\ \forall u \in \mathcal{U}, t \in \mathcal{T}. \tag{8}$$

Note that the number of (C3) - (C7) constraints depends on the number of conflicted beams at the BSs and UEs, and it may change from one scheduling period to another depending on the network density and topology.

Each associated UE u needs to satisfy its rate requirement R_u. This constraint can be expressed using the capacities in (1) as follows

$$(\mathrm{C8}):\ \sum_{b,l,t} c^l_{u,b} x^{l,t}_{u,b} \geq R_u s_u,\ \forall u \in \mathcal{U}. \tag{9}$$

Under the described interference suppressing constraints, our multi-criterion objective is to maximize the number of UEs with satisfied rate requirements using a minimal number of serving beams to do so. By reducing the number of serving beams, we also reduce the power of side-lobe interference, which is not suppressed by the constraints. The number of UEs with satisfied rate requirements and the number of serving beams (links) can be expressed as $\sum_u s_u$ and $\sum_{u,b,l,t} x^{l,t}_{u,b}$, respectively. Thus, the objective can be formulated as follows

$$\max_{\mathbf{s},\mathbf{x},\mathbf{a}} \sum_u s_u - \lambda \sum_{u,b,l,t} x^{l,t}_{u,b}. \tag{10}$$

The parameter λ is obtained through scalarization [13]. It can be shown that $\lambda = \frac{1}{K+1}$, where $K = \min\{N_{\mathrm{UE}}(TN^{\mathrm{RF}}_{\mathrm{UE}}-1)+1, N_{\mathrm{BS}}N^{\mathrm{RF}}_{\mathrm{BS}} \times (TN^{\mathrm{RF}}_{\mathrm{UE}}-1)/N^{\mathrm{RF}}_{\mathrm{UE}}+1\}$, guarantees that the number of satisfied UEs is maximized using a minimal number of serving beams. For the sake of brevity, we omit the details of scalarization.

Algorithm 1 Proposed greedy rounding algorithm

1: **Inputs:** $\mathbf{s}$, $\mathbf{x}$, ϵ
2: $\bar{s}_u \leftarrow 1$ if $s_u \geq \epsilon$, $\bar{s}_u \leftarrow 0$ otherwise
3: $\bar{x}_{u,b}^{l,t} \leftarrow 1$ if $x_{u,b}^{l,t} \geq \epsilon$, $\bar{x}_{u,b}^{l,t} \leftarrow 0$ otherwise
4: Determine $\bar{\mathbf{a}}$ based on $\bar{\mathbf{x}}$
5: **if** $\bar{\mathbf{s}}$, $\bar{\mathbf{x}}$, and $\bar{\mathbf{a}}$ are feasible **then**
6: $\hat{\mathbf{s}} \leftarrow \bar{\mathbf{s}}$, $\hat{\mathbf{x}} \leftarrow \bar{\mathbf{x}}$
7: **else**
8: Initialize: $\hat{\mathbf{s}} \leftarrow \mathbf{0}$, $\hat{\mathbf{x}} \leftarrow \mathbf{0}$, $\hat{\mathbf{a}} \leftarrow \mathbf{0}$, $\mathcal{A} = \{\emptyset\}$
9: Determine $v_u, \ \forall u$
10: **for** $n = 1 : TN_{\text{UE}}^{\text{RF}}$ **do**
11: Determine set $\mathcal{V}_n = \{u \mid v_u = n, \ u \notin \mathcal{A}\}$
12: Determine set $\mathcal{W}_n = \{u \mid v_u > n, \ u \notin \mathcal{A}\}$
13: **for** Each u in $\mathcal{V}_n$ **do**
14: Temporary variables: $\tilde{\mathbf{s}} \leftarrow \hat{\mathbf{s}}, \quad \tilde{\mathbf{x}} \leftarrow \hat{\mathbf{x}}, \quad \tilde{\mathbf{a}} \leftarrow \hat{\mathbf{a}}$
15: $\tilde{s}_u \leftarrow 1, \quad \tilde{x}_{u,b}^{l,t} \leftarrow \bar{x}_{u,b}^{l,t}, \ \forall b, l, t$ {n links}
16: Update $\tilde{\mathbf{a}}$ based on current $\tilde{\mathbf{x}}$
17: **if** $\tilde{\mathbf{s}}$, $\tilde{\mathbf{x}}$, and $\tilde{\mathbf{a}}$ are feasible **then**
18: $\hat{s}_u \leftarrow \tilde{s}_u, \quad \hat{x}_{u,b}^{l,t} \leftarrow \tilde{x}_{u,b}^{l,t}, \ \forall b, l, t$
19: $\mathcal{A} \leftarrow \mathcal{A} \cup u$
20: **end if**
21: **end for**
22: Repeat 13-21 for UEs in $\mathcal{W}_n$ using their n best links
23: **end for**
24: **end if**
25: **Outputs**: $\hat{\mathbf{s}}$, $\hat{\mathbf{x}}$

Finally, given the constraints and the objective, the linear optimization problem can be formulated as follows

$$\begin{aligned} \max_{\mathbf{s},\mathbf{x},\mathbf{a}} \quad & \sum_u s_u - \lambda \sum_{u,b,l,t} x_{u,b}^{l,t} \\ \text{s.t.} \quad & \text{(C1)} - \text{(C8)}, \\ & s_u, \ x_{u,b}^{l,t}, \ a_{u,u'}^{t} \in \{0,1\}, \quad \forall u, u', b, l, t. \end{aligned} \tag{11}$$

The optimization problem in (11) is a binary integer program (BIP), which is known to be NP-hard. This means that even a small-size problem with a few BSs and a moderate number of UEs and time slots has prohibitive computational complexity. To solve this linear optimization problem in polynomial time, we propose a low-complexity solution based on relaxation and greedy rounding in the next section.

4 PROPOSED ALGORITHM

We relax the BIP in (11) and formulate the following program with continuous optimization variables

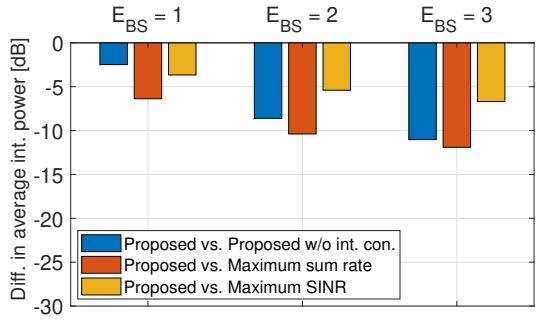
$$\begin{aligned} \max_{\mathbf{s},\mathbf{x},\mathbf{a}} \quad & \sum_u s_u - \lambda \sum_{u,b,l,t} x_{u,b}^{l,t} \\ \text{s.t.} \quad & \text{(C1)} - \text{(C8)} \\ & 0 \leq s_u, \ x_{u,b}^{l,t}, \ a_{u,u'}^{t} \leq 1, \quad \forall u, u', b, l, t. \end{aligned} \tag{12}$$

Let Q be the total number of variables in (12). The relaxed linear optimization problem can be solved in polynomial time. Specifically, the computational complexity scales as $O(Q^k)$, where k is the algorithm-dependent exponent. The solution to (12) is fractional, meaning that the values of s_u, $x_{u,b}^{l,t}$, and $a_{u,u'}^{t}$ are not necessarily equal to 0 or 1. For example, a UE u that experiences good channels can satisfy its rate requirement using only a fraction of its link capacities, i.e., the corresponding variables $x_{u,b}^{l,t}$ of the selected links may be significantly lower than 1. However, the proposed user association and beam scheduling framework requires all variables to be rounded to either 0 or 1.

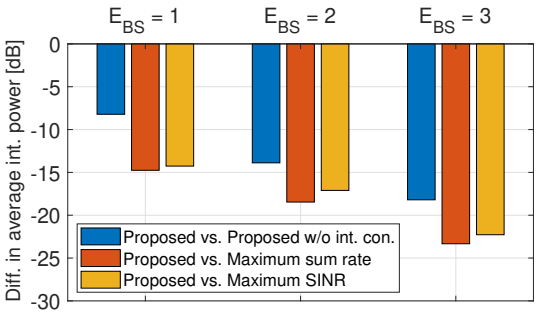
Here we propose a low-complexity greedy rounding algorithm. We first obtain the vector of associated UEs $\bar{\mathbf{s}}$, where each element $\bar{s}_u$ is set to 1 if $s_u \geq \epsilon$, and to 0 otherwise. Similarly, we obtain the vector of used links $\bar{\mathbf{x}}$ by setting $\bar{x}_{u,b}^{l,t}$ to 1 if $x_{u,b}^{l,t} \geq \epsilon$, and to 0 otherwise. The elements of the vector $\bar{\mathbf{a}}$ are calculated based on the elements of $\bar{\mathbf{x}}$. To avoid missing the links where only a fraction of the capacity is used, we assume that ϵ is a small constant value, e.g., $\epsilon = 0.1$. However, with a small ϵ, it is likely that too many links will be considered as 'used' and that the vectors $\bar{\mathbf{s}}$, $\bar{\mathbf{x}}$, and $\bar{\mathbf{a}}$ will violate multiple constraints and thus represent an infeasible solution. To obtain feasible vectors $\hat{\mathbf{s}}$, $\hat{\mathbf{x}}$, and $\hat{\mathbf{a}}$, we design a greedy algorithm, which aims to first associate the UEs that require only one link, then the UEs that require two links, and so on. We start by initializing the vectors $\hat{\mathbf{s}} = \mathbf{0}$, $\hat{\mathbf{x}} = \mathbf{0}$, $\hat{\mathbf{a}} = \mathbf{0}$, and define an empty set of associated UEs $\mathcal{A} = \{\emptyset\}$. Based on $\bar{\mathbf{x}}$, we determine a pessimistic estimate of the required number of links v_u for each u. Then we loop over $n = 1, ..., TN_{\text{UE}}^{\text{RF}}$, which represents all possible values of $v_u, \ \forall u$. For each n, we determine the set $\mathcal{V}_n$ of all UEs for which $v_u = n$, i.e., $\mathcal{V}_n = \{u \mid v_u = n, \ u \notin \mathcal{A}\}$. Similarly, we determine the set $\mathcal{W}_n = \{u \mid v_u > n, \ u \notin \mathcal{A}\}$. Next, we loop over all UEs in $\mathcal{V}_n$ to associate the UEs that satisfy the feasibility constraints. With every new associated UE, the vectors $\hat{\mathbf{s}}$, $\hat{\mathbf{x}}$, $\hat{\mathbf{a}}$, and the set $\mathcal{A}$ are updated. Since the estimates $v_u, \ \forall u$, are pessimistic, we loop over all UEs in $\mathcal{W}_n$ as well to check if any UE can be satisfied using their n best links, i.e., n links with the highest capacity. The proposed rounding algorithm is summarized in Algorithm 1. Note that the proposed algorithm always converges because the nested for-loops in Algorithm 1 have a finite number of iterations. The outer loop in line 10 has $TN_{\text{UE}}^{\text{RF}}$ iterations. The inner loops for the sets $\mathcal{V}_n$ and $\mathcal{W}_n$ depend on $v_u, \ \forall u$. In the most demanding case when all UEs require the maximum number of links, i.e., when $v_u = TN_{\text{UE}}^{\text{RF}}, \ \forall u$, only one of the inner loops is executed. Specifically, the loop for $\mathcal{W}_n$ is executed for $n = 1, ..., TN_{\text{UE}}^{\text{RF}} - 1$, while the loop for $\mathcal{V}_n$ is executed for $n = TN_{\text{UE}}^{\text{RF}}$. Regardless of n, the number of iterations in the executed inner loop is N_{UE}. Thus, in the worst case, the nested for loops have $TN_{\text{UE}}N_{\text{UE}}^{\text{RF}}$ iterations.

5 NUMERICAL EVALUATION

In this section, we numerically evaluate the proposed framework and we compare it with existing association frameworks, including the maximum sum rate and maximum signal-to-interference-plus-noise ratio (SINR). The objective of the maximum sum rate is to associate the UEs such that the overall data rate in the network is maximized. On the other hand, in the maximum SINR association,

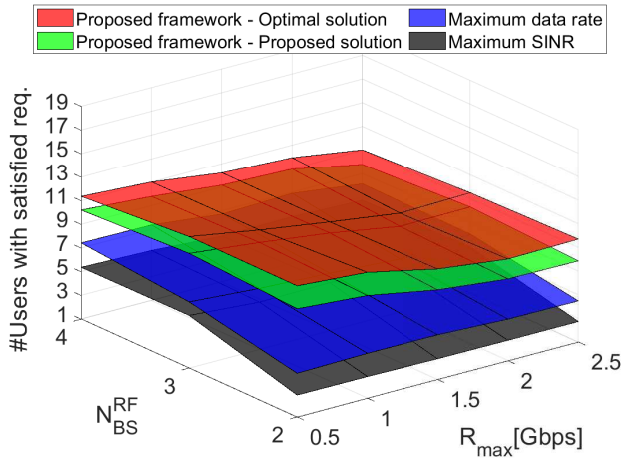


(a) $N_{\text{BS}}^{\text{RF}} = 4$

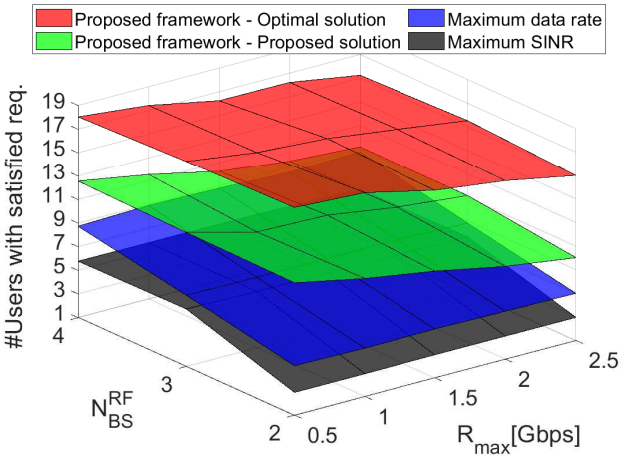


(b) $N_{\text{BS}}^{\text{RF}} = 12$

Figure 2: Difference in average interference power per link between the proposed framework, proposed framework without interference suppressing constraints, and benchmark frameworks.



(a) $E_{\text{BS}} = 1$



(b) $E_{\text{BS}} = 3$

Figure 3: Average number of UEs with satisfied rate requirements for the proposed and benchmark frameworks.

each UE selects a BS based on the maximum SINR, and then the BSs allocate their resources to the best UE candidates. We compare the frameworks in terms of the average interference power per scheduled link and the average number of UEs with satisfied rate requirements. It is worth noting that the interference power model from Section 3.1 was used only to calculate pessimistic capacity links, while here we consider the actual interference power from the scheduled links at *all* BS. The link capacities used to determine if a UE can satisfy its requirements are calculated based on the actual post-association interference power.

We consider a sub-network of $N_{\text{BS}} = 3$ BSs, similar as in Fig. 1. The inter-site distance is assumed to be $d = 200$ m. There are N_{UE} UEs and their positions are generated randomly within the triangle formed by the BSs. We consider an urban micro scenario with realistic mmMAGIC channels between the BSs and UEs. We generate the channels using the Quadriga simulator [14]. Based on the mmMAGIC channel model, the UEs can experience either line-of-sight or non-line-of-sight scenarios with the BSs. We assume the operating frequency $f = 28$ GHz, transmit power $P_{\text{T}} = 30$ dBm, bandwidth BW = 200 MHz, noise power spectral density $N_0 = -174\ \frac{\text{dBm}}{\text{Hz}}$, $N_{\text{UE}}^{\text{a}} = 8$ UE antennas, $N_{\text{UE}}^{\text{RF}} = 2$ UE RF chains, $N_{\text{BS}}^{\text{a}} = 32$ BS antennas, and a varying number $N_{\text{BS}}^{\text{RF}}$ of BS RF chains. The rate requirements are drawn uniformly from $R_u \sim U(R_{\min}, R_{\max})$, $\forall u$, where $R_{\min} = 0.2$ Gbps and $R_{\max}$ is the maximum possible requirement.

For our study of the average interference power in the network, we consider $T = 1$ time slot for simplicity and we assume $N_{\text{UE}} = 40$ and $R_{\max} = 1.5$ Gpbs. The results are presented in Fig. 2, where we highlight the difference in the average interference power per link between the proposed and benchmark frameworks. Numerical simulations are performed for different values of $N_{\text{BS}}^{\text{RF}}$, which is the the number of RF chains at each BS, and E_{BS}, which is the number of BSs that are known to each UE. Our results demonstrate that the proposed framework with low-interference beam scheduling leads to a significantly lower interference power per link than the benchmark frameworks that do not manage interference.

This is especially apparent when the number of RF chains (serving beams) at the BSs increases, which leads to more interference in the network. In interference-limited networks, a 20 dB decrease in interference power can increase the link capacity around 6x based on Shannon's formula. Additionally, Fig. 2 reveals the benefits of coordinated interference management. Namely, when the UEs estimate the channels with $E_{\mathrm{BS}} = 1$ BS, only intra-cell interference can be suppressed. On the other hand, with $E_{\mathrm{BS}} = N_{\mathrm{BS}}$, both inter- and intra-cell interference can be suppressed through coordinated scheduling among the BSs.

We assume $T = 8$ and $N_{\mathrm{UE}} = 20$ and we compare the frameworks in terms of the average number of UEs with satisfied rate requirements in Fig. 3. Numerical simulations are performed for different values of $N_{\mathrm{BS}}^{\mathrm{RF}}$, $R_{\max}$, and E_{BS}. The results indicate that the proposed framework leads to a higher number of satisfied UEs than the benchmarks. With $E_{\mathrm{BS}} = 1$, each UE is limited to only one BS and thus the problem of user association and beam scheduling can be considered on a cell-level. For this reason, the proposed low-complexity solution based on sequential greedy rounding is close to the optimal solution obtained by directly solving (11) using CVX [15]. On the other hand, with $E_{\mathrm{BS}} = N_{\mathrm{BS}} = 3$ and knowledge of the channel between any BS-UE pair, the performance gap between the proposed and optimal solutions increases. Nevertheless, the proposed low-complexity solution still performs better than the benchmarks. When the BSs have a low number of RF chains, the performance of the proposed framework is marginally affected. The reason for this is intelligent beam scheduling which can be leveraged to satisfy the UEs' rate requirements over time. On the other hand, an increase in the maximum rate requirement $R_{\max}$ leads to a slightly smaller number of satisfied UEs, as expected.

6 CONCLUSIONS AND FUTURE WORK

In this work, we proposed and mathematically formulated a new framework for user association and low-interference beam scheduling that maximizes the number of UEs with satisfied rate requirements. We also developed a sub-optimal low-complexity algorithm to solve the formulated NP-hard optimization problem. Numerical results showed that the proposed framework significantly reduces the directional interference and leads to a higher number of satisfied UEs compared to the benchmarks.

The results in this work represent a good starting point for the development of a more comprehensive framework. There are several important questions that need to be addressed in future work. First, if there are still available system resources, can they be allocated in a fair way to the UEs that could not satisfy their rate requirements? If so, can the benefits of low-interference beam scheduling be preserved after association of new UEs? Finally, given the allocated links and UEs' rate requirements, can power allocation at BSs be optimized on a sub-network level?

ACKNOWLEDGMENTS

This work was supported in part by NSF under grant 1718742. This work was also supported in part by the ComSenTer and CONIX Research Centers, two of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

REFERENCES

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang. What will 5G be? *IEEE Journal on Selected Areas in Communications*, 32(6):1065–1082, June 2014.

[2] Theodore S. Rappaport, Yunchou Xing, George R. MacCartney, Andreas F. Molisch, Evangelos Mellios, and Jianhua Zhang. Overview of millimeter wave communications for fifth-generation (5G) wireless networks—with a focus on propagation models. *IEEE Transactions on Antennas and Propagation*, 65(12):6213–6230, 2017.

[3] 3GPP. NR; Multi-connectivity; Overall description; Stage-2; 3GPP TS 37.340 v17.0.0, April 2022.

[4] Rui Liu and Guanding Yu. User association for millimeter-wave ultra-reliable low-latency communications. *IEEE Wireless Communications Letters*, 10(2):315–319, 2021.

[5] Alireza Alizadeh and Mai Vu. Load balancing user association in millimeter wave MIMO networks. *IEEE Transactions on Wireless Communications*, 18(6):2932–2945, 2019.

[6] Ehsan Moeen Taghavi, Alireza Alizadeh, Nandana Rajatheva, Mai Vu, and Matti Latva-aho. User association in millimeter wave cellular networks with intelligent reflecting surfaces. In *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, pages 1–6, 2021.

[7] Xuyao Sun, Chenhao Qi, and Geoffrey Ye Li. Beam training and allocation for multiuser millimeter wave massive MIMO systems. *IEEE Transactions on Wireless Communications*, 18(2):1041–1053, 2019.

[8] Zhongling Zhao, Jia Shi, Zan Li, Long Yang, Yue Zhao, and Wei Liang. Matching theory assisted resource allocation in millimeter wave ultra dense small cell networks. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pages 1–6, 2019.

[9] Jia Shi, Haris Pervaiz, Pei Xiao, Wei Liang, Zan Li, and Zhiguo Ding. Resource management in future millimeter wave small-cell networks: Joint PHY-MAC layer design. *IEEE Access*, 7:76910–76919, 2019.

[10] Wei Feng, Yanmin Wang, Dengsheng Lin, Ning Ge, Jianhua Lu, and Shaoqian Li. When mmWave communications meet network densification: A scalable interference coordination perspective. *IEEE Journal on Selected Areas in Communications*, 35(7):1459–1471, 2017.

[11] Ziyuan Sha and Zhaocheng Wang. Least pair-wise collision beam schedule for mmWave inter-cell interference suppression. *IEEE Transactions on Wireless Communications*, 18(9):4436–4449, 2019.

[12] Ziyuan Sha, Siyu Chen, and Zhaocheng Wang. Near interference-free space-time user scheduling for mmWave cellular network. *IEEE Transactions on Wireless Communications*, pages 1–1, 2022.

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

[14] S. Jaeckel, L. Raschkowski, K. Börner, L. Thiele, F. Burkhardt, and E. Eberlein. QuaDRiGa - Quasi Deterministic Radio Channel Generator, user manual and documentation. *Fraunhofer Heinrich Hertz Institute, Tech. Rep. v2.2.0*, 2019.

[15] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.