Context-Aware Resource Management and Performance Analysis of Millimeter Wave and Sub-6 GHz Wireless Networks

Omid Semiari

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Electrical Engineering

Walid Saad, Chair
Jeffrey H. Reed
Harpreet S. Dhillon
Paul E. Plassmann
Xiaoyu R. Zheng

June 5, 2017
Blacksburg, Virginia

Keywords: Heterogeneous Networks, Millimeter Wave Communications, Cellular Networks, Matching Theory, Context Information

Context-Aware Resource Management and Performance Analysis of Millimeter Wave and Sub-6 GHz Wireless Networks

Omid Semiari

ABSTRACT

Emerging wireless networks are foreseen as an integration of heterogeneous spectrum bands, wireless access technologies, and backhaul solutions, as well as a large-scale interconnection of devices, people, and vehicles. Such a heterogeneity will range from the proliferation of multi-tasking user devices such as smartphones and tablets to the deployment of multi-mode access points that can operate over heterogeneous frequency bands spanning both sub-6 GHz microwave (μW) and high-frequency millimeter wave (mmW) frequencies. This heterogeneous ecosystem will yield new challenges and opportunities for wireless resource management. On the one hand, resource management can exploit user and network-specific *context information*, such as application type, social metrics, or operator pricing, to develop application-driven, context-aware networks. Similarly, multiple frequency bands can be leveraged to meet the stringent quality-of-service (QoS) requirements of the new wireless services such as video streaming and interactive gaming. On the other hand, resource management in such heterogeneous, multi-band, and large-scale wireless systems requires distributed and low-complexity frameworks that can effectively utilize all available resources. The key goal of this dissertation is therefore to develop novel, self-organizing, and low-complexity resource management protocols – using techniques from *matching theory*, optimization, and machine learning – to address critical resource allocation problems for emerging heterogeneous wireless systems while explicitly modeling and factoring diverse network context information.

Towards achieving this goal, this dissertation makes a number of key contributions. First, a novel context-aware scheduling framework is developed for enabling dual-mode base stations to efficiently and jointly utilize mmW and μW frequency resources while maximizing the number of user applications whose stringent delay requirements are satisfied. The results show that the proposed approach will be able to significantly improve the QoS per application and decrease the outage probability. Second, novel solutions are proposed to address both network formation and resource allocation problems in multi-hop wireless backhaul networks that operate at mmW frequencies. The proposed framework motivates collaboration among multiple network operators by resource sharing to reduce the cost of backhauling, while jointly accounting for both wireless channel characteristics and economic factors. Third, a novel framework is proposed to exploit high-capacity mmW communications and device-level caching to minimize handover failures as well as energy consumption by inter-frequency measurements, and to provide seamless mobility in dense heterogeneous mmW-μW small cell networks (SCNs). Fourth, a new cell association algorithm is proposed, based on matching theory with minimum quota constraints, to optimize load balancing in integrated mmW-μW networks. Fifth, a novel medium access control (MAC) protocol is proposed to dynamically manage the wireless local area network (WLAN) traffic jointly over the unlicensed 60 GHz mmW and sub-6 GHz bands to maximize the saturation throughput and minimize the delay experienced by users. Finally, a novel resource management approach is proposed to optimize device-to-device (D2D) communications and improve traffic offload in heterogeneous wireless SCNs by leveraging social context information that is dynamically learned by the network. In a nutshell, by providing novel, context-aware, and self-organizing frameworks, this dissertation addresses fundamentally challenging resource management problems that mainly stem from large scale, stringent service requirements, and heterogeneity of next-generation wireless networks.

Context-Aware Resource Management and Performance Analysis of Millimeter Wave and Sub-6 GHz Wireless Networks

Omid Semiari

General Audience Abstract

The emergence of bandwidth-intensive applications along with vast proliferation of smart, multi-tasking handhelds have strained the capacity of wireless networks. Furthermore, the landscape of wireless communications is shifting towards providing connectivity, not only to humans, but also to automated cars, drones, and robots, among other critical applications. These new technologies will enable devices, machines, and things to be more intuitive, while being more capable, in order to improve the quality of life for human. For example, in future networked life, smartphones will predict our needs and help us with providing timely and relevant information from our surrounding. As an another example, autonomous vehicles and smart transportation systems with large number of connected safety features will minimize road incidents and yield a safe and joyful driving experience.

Turning such emerging services into reality will require new technology innovations that provide high efficiency and substantial levels of scalability. To this end, wireless communication is the key candidate to provide large-scale and ubiquitous connectivity. However, existing wireless networks operate at congested microwave (μW) frequency bands and cannot manage the exponential growth in wireless data traffic or support low latency and ultra-high reliability communications, required by many emerging critical applications. Therefore, the goal of this dissertation is to develop novel network resource utilization frameworks to efficiently manage the heterogeneous traffic in next-generation wireless networks, while meeting their stringent quality-of-service (QoS) requirements.

This transformative, fundamental research will expedite the deployment of communications at very high frequencies, at the millimeter wave (mmW) frequency bands, in next-generation wireless networks. The developed frameworks will advance new concepts from matching theory and machine learning for resource management in cellular networks, wireless local area networks (WLANs), and the intersection of these systems at both mmW and μW unlicensed frequency bands. This multi-band networking will leverage the synergies between mmW and μW wireless networks to provide robust and cost-effective solutions that enable the support of heterogeneous traffic from future wireless services. The anticipated results will transform the way in which spectral and time resources are used in both cellular networks and WLANs.

To my wife, my parents, and my brother.

Acknowledgments

First and foremost, I owe my deepest gratitude to my advisor, Dr. Walid Saad, for his continuous support of my Ph.D study, his patience, motivation, and immense knowledge and experience. I would like to thank you for the amount of time and effort, ideas, and funding you have generously dedicated to make my Ph.D. experience productive and stimulating. I am also grateful for believing in me and giving me the freedom to pursue diverse, yet coherent research directions which has made my Ph.D. study a joyful and unforgettable journey. I am thankful for your priceless advice and great supervision that have helped me to grow as a research scientist and find the right path for my future career.

I would like to thank the members of my Ph.D. advisory committee, Dr. Jeffery H. Reed, Dr. Harpreet Dhillon, Dr. Paul Plassmann, and Dr. Xiaoyu R. Zheng, for their valuable comments which have helped me to substantially improve the quality of this dissertation.

I am grateful for having the opportunity to work with great researchers and scientists during my Ph.D. study to publish scholarly papers. Special thanks to my co-authors, Dr. Mehdi Bennis, Dr. Behrouz Maham, Dr. Stefan Valentin, Dr. Zaher Dawy, Dr. Vincent Poor, and Dr. Merouane Debbah for their time and effort to provide me with their insightful comments to improve my work throughout the last four years. I also would like to thank the folks at Wireless@VT, specially my friends at NetSciWis lab for their help and support.

Last but not the least, words cannot express how grateful I am to my beloved wife, Mansooreh. Without her unparalleled love and support, I could have never been able to finish this dissertation. To my mother and my father for their continuous and unconditioned love. Thank you for always believing in me and for all of the sacrifices that you have made on my behalf. To my amazing brother, who has always inspired me, has given me the courage, and who has guided me in the most difficult stages of my life. I would also like to thank my parents-in-law, my brother's wife, my beautiful little nephew, and all of my friends who incentivized me to strive towards my goal.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

The exponential proliferation of new, highly-capable mobile wireless devices such as smartphones and tablets has significantly increased the growth in the demand for pervasive wireless access [1]. This rapid increase in the number of devices, combined with the emergence of advanced wireless networking technologies and systems will lead to a widespread adoption of bandwidth-intensive wireless services, such as social networking, interactive gaming, and multimedia streaming. This new breed of wireless services will effectively strain the capacity of existing wireless cellular systems and impose strict quality-of-service (QoS) requirements, in terms of the required data rates and tolerable application-specific delays. Naturally, existing wireless standards and network architectures, which have been originally designed primarily for voice services, will not be able to handle such stringent traffic and QoS requirements. Therefore, a substantial amount of research has recently emerged, in both industry and academia, with a focus on deriving a new generation of wireless networks that can properly cope with such emerging trends. In this chapter, we will overview the outcomes of these efforts and emerging technologies while outlining their accompanying research challenges and opportunities.

This chapter is organized as follows. In Section 1.1, key concepts of emerging wireless cellular networks and their challenges are discussed. The main contributions of this dissertation are presented in Section 1.2. Section 1.3 provides the list of publications and Section 1.4 presents the outline of this dissertation.

1.1 Emerging Wireless Heterogeneous Networks: Opportunities and Challenges

First, we briefly overview a number of key concepts [1] and technologies that are expected to shape the next generation of wireless networks and prove the platform within which the aforementioned,

[1] Here, we note that emerging future wireless techniques, such as massive MIMO and spectrum sharing will also play a key role in future cellular networks, but are not within the scope of this dissertation.

bandwidth-intensive wireless services will operate.

1.1.1 Small cell networks

Over the past decade, wireless cellular networks were typically reliant on high-power macrocellular base stations (MBSs) that provide coverage for large geographical areas (few kilometers). However, such a conventional deployment of base stations (BSs) may not be spectrally efficient and has been shown to be ineffective in catering for bandwidth-intensive wireless applications [2]. In addition, cell splitting gains are significantly limited by severe inter-cell interference as more MBSs are deployed. To overcome these challenges, one promising concept is to reduce the cell sizes by deploying low-power small base stations (SBSs) [3–6]. Such *small cell networks (SCNs)*, composed of densely deployed SBSs, such as picocells and femtocells, allow reducing the coverage holes and substantially increasing the spectral efficiency. Compared with conventional cellular networks, some of main challenges of SCNs include: [5]

- *Density:* A dense deployment of SCNs will introduce new challenges in terms of interference management and resource allocation. These challenges stem from the key features of SCNs such as the unplanned SBSs distribution, limited coverage, and limited backhaul capacities [3–6].
- *Cell association:* Conventional cell association approaches assign each user to the MBS with maximum RSSI or maximum SINR. In SCNs, such association mechanisms may result in unbalanced load distributions, since most of the users will be assigned to the SBSs with higher transmit powers.
- *Backhaul:* Unlike conventional MBS-based cellular networks, in SCNs, providing a fiber backhaul for connecting a dense number of SBSs to the core network will not be cost-effective from an economic perspective, due to high cost of deployment and leasing of fiber backhaul. In addition, fiber backhaul solution is not feasible for SBSs that are deployed in adverse locations. Therefore, wireless backhaul solutions are being considered as a viable solution for the SCNs as they can enable each SBS to connect to the core network via a single-hop or a multi-hop wireless backhaul link. Supporting high capacity and reliable wireless backhaul is one of the key challenges of SCNs deployment [7–16].
- *Mobility management:* Due to the reduced cell sizes in SCNs, handovers (HOs) happen more frequently for mobile users, compared to the macro cellular networks [17–25]. Handover introduces packet loss and latency which can be detrimental to the QoS of wireless data communication. Conventional handover techniques which are based on the received signal power will be inefficient in SCNs, since they can adversely increase the number of handovers and degrade the performance.

In addition, we note that the capacity scales linearly with the number of cells. Hence, SCNs alone will not be able to meet the required capacity to accommodate orders of magnitude increase in mobile data traffic [26]. This, in turn, has led to the emergence of additional technologies, as discussed in the following sections, which are expected to co-exist with SCNs.

1.1.2 Millimeter wave communications

Most mobile communication systems today operate at the sub-6 GHz frequency bands, in particular within the frequency spectrum range of 300 MHz- 3 GHz. One of the key problems of operating at such frequency bands is that the transceivers and their RF components, including power amplifiers, low noise amplifiers, mixers, and antennas, are bulky and often power inefficient [26]. In addition, these frequency bands are very congested and expensive to lease. Moreover, low-frequency signals can effectively penetrate the obstacles and reflect from the object. Although this is desired in general, it makes the interference mitigation more challenging for dense SCNs. These challenges have provided incentives for mobile network operators (MNOs) to consider the possibility of operating cellular networks at higher frequency bands at which high bandwidth is readily available [26].

The *millimeter wave* (mmW) spectrum band, ranging from 30-300 GHz is an attractive solution as it offers a significantly large available bandwidth that can reach, up to 1 GHz. Indeed, in the past few years, mmW has been utilized in several industrial standards, such as IEEE 802.15.3c and IEEE 802.11ad [27]. Operating at mmW frequencies allows the use of small-size antenna arrays with large number of elements that can achieve high beamforming gains. In fact, the path loss and rain attenuation at 28 GHz band can be compensated by using directional antennas. Recent field measurements have shown that mmW transmission over distances up to 300 meters is feasible which makes mmW a very natural solution for covering cell sizes in SCNs [28].

However, mmW signals are highly susceptible to the blockage. For example, metal-coated glass walls can attenuate the mmW signal for up to 50 dB, while penetration loss for brick walls is even much higher [29]. This characteristic of mmW frequencies will substantially reduce the inter-cell interference, however, it can also be detrimental for the desired signals. Another practical challenge for mmW communications is *deafness*. Deafness occurs whenever transceivers fail to align their antenna beams in the desired direction. To perform beam alignment, mmW transceivers must follow a process, called *beam training*, which essentially assists transceivers to find coefficients that maximize the beamforming gain.

Unfortunately, blockage and deafness can frequently occur, due to the movements of user or objects in the environment. Therefore, guaranteeing reliable transmissions for QoS-sensitive applications will be challenging for mmW networks [7, 26, 30–34]. To this end, mmW communication has to co-exist with conventional cellular networks operating at microwave (μW) frequencies [26, 31, 35, 36]. For such integrated mmW-μW networks, novel protocol and resource management approaches are required to efficiently allocate available resources at both frequency bands to the users, while considering the QoS constraints.

1.1.3 Device-to-device communications and caching

One of the key enabling technologies to increase the spectral efficiency and decrease the transmission delay is to enable mobile devices to directly communicate with one another [37–43]. In fact, *device-to-device (D2D)* over cellular communication is defined as a direct transmission between two mobile users without traversing the BSs or core network. Therefore, D2D communication can be an effective technology for reducing the traffic from BSs and the backhaul. Moreover, D2D communications have a high relevance to expedite the deployment of proximity-based applications, such as content distribution and location-aware advertisement. Nonetheless, D2D links may interfere with the uplink and downlink cellular transmissions. In addition, at sub-6 GHz, D2D communications require each UE to discover nearby devices, initialize a D2D link, and allocate the resources for D2D and cellular link transmissions. Therefore, novel networking protocols and resource management techniques must be developed for D2D-enabled cellular networks [37–40].

Along with D2D communications, a promising solution for reducing peak hour traffic is to store viral data contents at users' devices for future usage. This concept, referred to as *caching*, will allow D2D-enabled SCNs to avoid redundant transmissions at the access network, if the required content has been effectively cached in nearby devices [44–47]. Modern user devices benefit from sufficiently large storage capacities which enables the network to store large amount of data at each device. In addition, caching at SBSs will allow reducing the backhaul traffic which is one the major bottlenecks of SCNs. The achievable gains of caching for offloading traffic from access and backhaul networks depend on how much nearby users are interested in similar data contents. Therefore, novel context-aware resource allocation approaches must be developed that leverage similarities in users' interests to increase traffic offloads via D2D communications.

1.1.4 Context-aware heterogeneous networks

Prior to the introduction of modern smart handhelds, mobile devices had limited capabilities and typically were unable to carry out complicated tasks or run highly complex wireless applications. Therefore, the majority of the network optimization has traditionally been done at the MBSs. In fact, MBSs typically act as a centralized control centers that provide service to mobile devices. Considering the capabilities of modern smart handhelds, they are able to run different applications simultaneously and provide useful information such as user's location, user's trajectory, and among other useful information [48–54]. Therefore, conventional resource allocation solutions may not be optimal anymore, in the sense that they do not leverage the capabilities of smart devices. To this end, new resource allocation approaches have recently been introduced that aim to exploit useful information extracted from smart devices, which is known as *context information (CI)*, in the network optimization.

Context, as a research notion, has been introduced and exploited in many fields of informatics since 1960s and refers to the idea that computers can sense, react and possibly adapt their functionalities based on the information they acquire from their environment [55]. The term *context awareness* was first explicitly introduced in the research area of pervasive computing in [56] and refers, in general, to the ability of computing systems to acquire and reason about the CI and adapt the corresponding applications accordingly.

In wireless networks, the term context-aware is used to describe the knowledge extracted from the environment that can be jointly used with physical layer metrics, such as channel state informa-

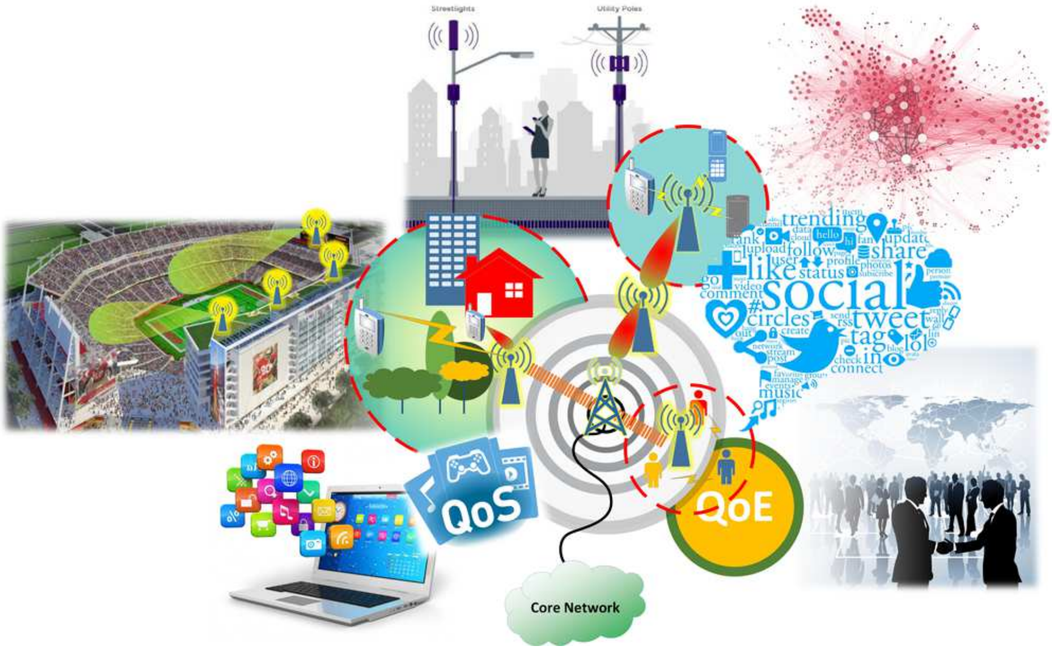


Figure 1.1: Context-aware heterogeneous network.

tion (CSI), to improve resource management and scheduling. In this sense, *context-aware resource management* is defined as any scheduling scheme that exploits CI. Fig. 1.1 shows a high-level description of future context-aware heterogeneous networks (HetNets), enabled by the key concepts introduced in previous subsections.

Although context-aware networks have been studied in computer science and other disciplines for years, this literature is quite immature in wireless communications. Here, we fist develop a classification for different types of CI for wireless-oriented problems. For each context category, we overview the existing literature on context-aware resource allocation approaches.

In general, CI can be classified into two broad groups as follows:

Human-centric CI:

This CI category includes all the information that is extractable from user behavior in real life that may directly or indirectly affect the wireless network, e.g. data usage patterns, mobility, or mmW link blockage by a human body. In practice, human behavior is too complicated and cannot be studied completely using formal analytical models. However, such complicated behavior can be abstracted into some level of CI by using standard techniques including, monitoring, learning, and predicting.

With this in mind, one can see that human-centric CI depends on the temporal and spatial correlations of the people activities in real life and their impact on data traffic in wireless network. Based on the existing works in the literature [51–54, 57–63], we introduce four different dimensions for the user-centric CI, as defined in Table 1.1.

Table 1.1: Dimensions of human-centric CI

Dimension	**Definition**
Physical	Including the location of the user, user path and trajectory. This also includes environmental location information such as the obstacles, moving objects, and whether. [57–60]
Social	The relationship with, and the density, flow, type, and behavior of, surrounding people. [51–54, 61]
Task	The functional relationship of the user with other people and objects, and the benefits (e.g. resources available, monitory incentives, etc.) or constraints. [62, 63]
Temporal	The temporal context is embedded within everything, and is what gives a *current* situation meaning, based upon *past* situations/occurrences, expected *future* events, and the higher-level temporal context relating to the time of day, week, month, or season. [57–61]

Our classification in Table 1.1 is in line with the classification presented in [64] for user-centric and multidisciplinary context-aware frameworks. However, we mainly focus on the CI that can be exploited in wireless resource allocation problems. Next, we review the body of work in the literature that takes user driven CI into account.

In [57], a supervised learning algorithm is presented to predict the user's mobility. Using long-term handover information and short-term CSI, authors have formulated the prediction as a classification problem. In [58], authors presented an anticipatory resource allocation scheme for wireless video streaming. The key assumption of this work is that the channel state can be predicted for the upcoming time slots by tracking the pathloss of the user's trajectory. In this regard, the authors proposed an optimization problem to adjust the video requested rate per time slot to have enough buffered video on the one hand while consuming the minimum spectrum on the other hand. Authors in [59] extended the work in [58], but considered imperfect rate prediction. In addition, in [60], authors presented a location-based adaptive video quality planning, content pre-fetching, and long-term radio resource management.

The work in [51] adopts an analytical model for the epidemic information spreading among mobile users of an ad hoc network. In [52], resource allocation in a wireless local area networks (WLAN) is defined as an optimization problem while taking the notion of social distance into account. The authors in [53] and [54] extend the work in [52] by introducing new utility functions which again account for the social distance of users, extracted from the social graph.

In addition, the work in [62] considered the pricing as a key concept to provide seamless mobility for mobile users in future wireless networks. In fact, authors outlined the major issues in designing resource allocation and pricing in heterogeneous wireless access networks. Moreover, the authors in [63] proposed a game theoretic analysis for service competition and pricing in heterogeneous wireless access networks. As we discuss next, there is another CI category that mainly stems from the network characteristics, rather than the user behavior.

Table 1.2: Dimensions of Network-centric CI

Context	Description
Power	Network may give higher priority and hence, more resources to the device with a low battery state in order to finish a task in a shorter time. [65, 66]
QoS	QoS requirements by different services impacts the resource allocation. [65–69]
Traffic type	Based on the type of traffic, e.g. voice, video, etc., network determines which frequency band (WiFi, LTE, mmW, etc.) is suitable to serve the user. [67, 68, 70]
LCD size	Small screen sizes may not require high quality video streaming. [68]
Storage capacity	User devices with more available storage capacity are able to cache more content, thus, providing more opportunities for device to device communications. [70]
Multi-band operation and Radio interfaces	Based on the capability of the user devices to operate in different frequency bands, network can manage vertical handoffs and traffic offloads. [67, 71]

Network-centric CI:
This class of CI includes any information that is specified by network capabilities, the requirements of the requested traffic, and any other information that is not directly impacted by the human user. As we discuss later, it is often more convenient to adopt context-aware resource allocation approaches based on this type of CI, compared with the human-centric CI. That is because network-centric CI typically presents a wireless network metric, while it may not be straightforward to translate human-centric CI, e.g. social interrelationships among users, directly into a quantitative wireless metric.

In Table 1.2, we list some of the network-centric CI that are used in the literature. The work in [67] proposes a radio access technology (RAT) selection scheme from which small cell base stations autonomously offload delay-tolerant traffic into the unlicensed frequency band. In [71], a load-aware user-centric RAT selection scheme is proposed that allows to offload traffic to WiFi small cells, while minimizing feedback overhead and better accounting for user preferences.

In addition, the work in [68] adopts a context-aware user-cell association approach that takes the QoS requirements of different traffics into account. The QoS requirements are determined based on the context features, including application as well as the hardware in use. In [70], the authors proposed a user-cell association and backhaul resource management by envisioning the popularity of the cached content at the SBSs and the estimated incoming file requests.

In [69], the authors proposed a context-aware resource management approach that jointly optimizes resource allocation at uplink and downlink. The approach of [69] processes the applications' profile and traffic patterns in each of the cells to ensure that user requirements are satisfied while guaranteeing the best network performance in terms of throughput. Moreover, the work presented in [65] considered power constraints of user devices for delay-optimal resource allocation in the uplink. Furthermore, in [66], the authors surveyed different cross-layer resource allocation approaches in wireless networks that address delay-energy tradeoffs as well as lookahead scheduling algorithms. Although the various approaches presented in [66] may not directly exploit UE driven CI, they provide a useful guidelines to develop context-aware approaches for future wireless networks.

The body of work we overviewed in this section presents interesting context-aware concepts in wireless networks. However, it is mostly focused on resource management for conventional cellular networks and disregards specific challenges in heterogeneous networks (HetNets), discussed in Section 1.1.1-1.1.3, such as resource allocation for integrated mmW-μW communications and D2D-enabled SCNs. In this dissertation, our goal is to use the notion of context-awareness to address some of the challenging problems in future HetNets, as outlined in the next section.

1.2 Contributions

The main contribution of this dissertation is to provide novel analytical frameworks that bring forward new ideas from matching theory, machine learning, and optimization, to address some of the fundamental challenges of future wireless networks by developing novel, context-aware resource management algorithms and protocols, as well as providing performance analysis for various scenarios in both cellular and local area networks. In particular, in this dissertation, we provide a comprehensive study for the following problems: 1) Traffic management for heterogeneous mmW-μW cellular networks to increase users' quality-of-experience (QoE), while considering the unique constraints of mmW signal propagation, 2) Tractable analysis of joint network formation and resource allocation in multi-hop mmW backhaul networks, 3) Mobility management for integrated mmW-μW networks, 4) Load balancing and cell association for heterogeneous mmW-μW cellular networks, 5) Performance analysis of joint mmW-μW WLANs, and 6) Leveraging the synergies between wireless and social networks to enhance the overall QoS delivered by small cell-based cellular systems.

In fact, in this dissertation, we answer the following fundamental questions:

1) *Subject to the intermittent nature of mmW signals, how can a cellular system maximize the QoS for applications with stringent delay requirements, while leveraging the large bandwidth at mmW frequencies?*

   To answer this question, in Chapter 3, we introduce the concept of integrating mmW communications into μW systems at the medium access control (MAC) layer and we propose a joint scheduler that dynamically manages the mmW-μW resources, while taking into account the constraints of each frequency band. The proposed framework is shown to effectively leverage the large bandwidth at the mmW frequencies to maximize the number of severed user applications, while achieving robustness against blockage and increasing the

QoS by exploiting μW frequencies. To achieve this goal, the proposed scheduler employs a set of context information, including the LoS probability per user, as well as the delay requirement per user application. To learn this context information, we propose a novel Q-learning model with fast convergence that finds the line-of-sight (LoS) probability per user by exploring three states, including mmW LoS, mmW non-LoS (NLoS), and μW, over the past transmissions. We solve this context-aware scheduling problem by proposing two novel algorithms that run jointly over the mmW and μW frequency bands. The main outcomes of this research direction can be summarized as follows:

- The proposed context-aware scheduling framework for integrated mmW-μW networks is shown to significantly improve the QoS per user application.
- The proposed framework for QoS provisioning per user application is shown to maximize the QoE for the users that run multiple applications simultaneously.
- The results show that, compared with conventional scheduling schemes, such as a proportional fair scheduler, the proposed approach significantly decreases the outage probability. In addition, the proposed integrated mmW-μW network completely outperforms the single-mode networks with only mmW or μW resources.
- We show that the beam-training overhead at the mmW frequency band substantially affects the statistics of the outage.
- The complexity of the proposed approach is shown to be polynomial with respect to the network size.

2) *How can mmW frequencies be exploited to achieve a low-cost, yet reliable backhaul solution for dense small cell networks, while considering the limitations of mmW communications?*

MmW frequencies offer a large bandwidth which can be exploited to achieve high data rates which makes them a promising candidate for supporting backhaul connectivity for dense small cell networks. However, mmW links are limited in range and susceptible to blockage. To address these challenges, in Chapter 4, we propose a framework that allows small cells to connect to the core network over multi-hop mmW backhaul links. In addition, to achieve robustness against blockage, we motivate cooperation among network operators, such that a BS of one operator can support backhaul connections for another operator's BS the experience blockage. Within the proposed framework, we develop two novel self-organizing algorithms to solve backhaul network formation and resource allocation problems. The key outcomes of this research are:

- We show that the conventional deferred acceptance algorithm fails to guarantee two-sided stability for the backhaul resource allocation problem. On the other hand, we prove that the proposed resource management algorithm yields a two-sided stable allocation of mmW frequency resources to demanding small cells.
- We prove that the proposed algorithms converge in polynomial time with respect to the network size, are distributed, guarantee two-sided stability, and thus, are suitable to realize a self-organizing backhaul network with dense small cell deployments.
- The results show that the proposed cooperative scheme achieves a performance that is close to the optimal solution found by the exhaustive search. In addition, the proposed

scheme significantly outperforms the non-cooperative approach in which operators do not share their resources.

- It is shown that cooperation among operators facilitates flexible mmW backhaul solutions that are more immune against the blockage.
- We demonstrate that the proposed framework can effectively capture the economic aspects of cooperation among operators. In particular, our results provide a design guideline that subject to the resource price and budget constraints, determines the operation region for the proposed cooperative backhaul network.

3) *How to minimize the number of handovers, power consumption, and handover failure for mobile users in emerging dense heterogeneous networks?*

   One of the critical challenges of future wireless networks is to provide seamless HO for mobile users, without degrading the QoS. Considering the various coverage areas of each small cell as well as their dense deployment, a mobile user may potentially be required to perform frequent HOs which results in an excessive power consumption to perform inter-frequency search, HO failures, and low QoS. To address this challenging problem, in Chapter 5, we propose a novel mobility management framework that allows users to mute their HO and cell search process, while traversing small cells. This scheme is realized in integrated mmW-μW networks where high-speed mmW links can be leveraged, whenever available, to cache the content at the mobile user device. Meanwhile, the control and paging information are handled over the μW frequency band. We provide a comprehensive performance analysis for the proposed mobility management scheme, using a geometric framework. Furthermore, we propose a distributed HO mechanism, based on dynamic matching games, to associate mobile users to the BSs. The major outcomes of this research are:

   - Fundamental results on the caching capabilities, including caching probability, duration, and the average achievable rate of caching are derived for mobile users. Moreover, the impact of caching on the number of HOs, energy consumption, and the average handover failure (HOF) is analyzed.
   - We propose a novel mobility management algorithm that finds the best HO policy for a mobile user, i.e., to choose between: a) executing an HO to a target cell, b) being connected to the macrocell base station, or c) perform a transparent HO by using the cached content.
   - The proposed dynamic matching algorithm is proved to converge to a *dynamically stable* association between mobile users and BSs. This key result shows the merit of proposed approach to realize future self-organizing networks.
   - The results show that the proposed mobility management framework yields significant performance gains, in terms of reducing the number of HO failures, energy consumption, and the HO probability.
   - The results also show that with low-complexity and overhead, the proposed distributed algorithm is capable of offloading traffic from the macrocell base stations, even for the users with relatively high speeds.

4) *How to achieve a balanced load distribution in emerging wireless heterogeneous networks*

with both mmW and μW radio access technologies?

Future cellular networks will offer connectivity over multiple RATs using mmW and μW frequencies. In Chapter 6, we show that the conventional cell association schemes, including max-RSSI and max-SINR associations, will result in drastically unbalanced load distribution across mmW and μW BSs. To alleviate this problem, we propose a novel distributed approach, based on *matching games with minimum quota constraint*, which yields the following key results:

- We show that, for cell association problems with minimum quota constraints, the standard deferred acceptance algorithm may not admit a feasible solution.
- The proposed cell association algorithm is proved to converge to a feasible Pareto optimal and stable matching between users and BSs.
- Simulation results show that, compared with conventional max-SINR and max-RSSI with cell range expansion, the proposed approach achieves significant performance gains in terms of maximizing the sum rate, while maintaining a balanced load across mmW and μW RATs.

5) *What are the performance gains that can be achieved by leveraging unlicensed 60 GHz mmW band in WLANs?*

Exploiting the large available bandwidth at the unlicensed 60 GHz mmW band, jointly with the unlicensed μW frequencies, will be a key enabler to support bandwidth-intensive and delay sensitive emerging applications such as virtual reality in WLANs. To enable such an integrated mmW-μW WLAN, in Chapter 7, we propose a novel MAC protocol that enables users to dynamically leverage the bandwidth available at the 60 GHz mmW band and alleviate the excessive delay caused by the contention-based medium access over the μW frequencies. To analyze the performance of the proposed MAC protocol, we adopt a Markov chain model for backoff time and derive the probability of transmission over each RAT, as well as the system's saturation throughput. Next, the key findings of this research are outlined.

- We introduce a novel MAC protocol that relies on dynamic *fast session transfer* between mmW and μW RATs, which is shown to be backward compatible with legacy IEEE 802.11 WLANs.
- The proposed MAC protocol inherently captures the constraints of each mmW and μW frequency bands, including the intermittent channel at the 60 GHz band, directional mmW transmissions, and the level of congestion observed over the sub-6 GHz bands.
- Simulation results corroborate the analytical derivations and show that the proposed integrated mmW-sub 6 GHz MAC protocol yields significant performance gains, in terms of maximizing the saturation throughput.
- Both analytical and simulation results will show that the proposed MAC scheme effectively minimizes the delay experienced by the users and is suitable to support low-latency communications in WLANs.
- The results also provide insights on the tradeoffs between the achievable gains and the overhead introduced by the fast session transfer procedure.

6) *How to study the users' common interests and exploit that context information to reduce the backhaul traffic in heterogeneous cellular networks with D2D capabilities?*

   Next-generation wireless networks will support proximity services enabled with D2D communications. In Chapter 8, we show that leveraging the D2D capability along with the underlying correlation in users' interests will results in substantial offloading gains for wireless backhaul networks. The key idea to achieve such offloading gains is to let users with common interests to form D2D clusters, cache popular content (depending on the users' interests within the cluster), and allow users to directly serve one another over D2D links. This framework will decrease the number of users' requests, submitted to SBSs for receiving popular content, and consequently reduce the backhaul traffic of small cells. The main findings of this research are:

   - We adopt a graphical learning approach that studies the common attributes from the users' profiles in social networks, and translates them into social tie strength among users.
   - Using this context information, we propose a novel resource allocation framework that enables users with social ties to form social clusters, use the cached content within the cluster, and avoid redundant requests for the popular content from BSs.
   - The context-aware resource management problem is formulated as a one-to-many *matching game with externalities*, capturing the impact of interdependent social ties on the social cluster formation.
   - We propose a novel distributed resource allocation algorithm that leverages the social context in addition to the channel state information. We show that the complexity of the proposed algorithm, in terms of signaling overhead, is polynomial with respect to the network size.
   - We prove that the proposed algorithm effectively handles the externalities observed in the matching game and yields a two-sided stable allocation of resource blocks to the users.
   - The results show that with manageable complexity, the proposed context-aware approach can offload a large amount of traffic from the backhaul-constrained small cell network.

1.3 List of Publications

As a byproduct of the above contributions, this dissertation has led to the following key publications:

1.3.1 Journal papers

[J1] O. Semiari, W. Saad, and M. Bennis "Context-Aware Scheduling of Joint Millimeter Wave and Microwave Resources for Dual-Mode Base Stations," *IEEE Transactions on Wireless*

Communications, accepted and to appear, May. 2017. Available at `http://arxiv.org/abs/1606.08971`.

[J2] O. Semiari, W. Saad, M. Bennis, and Z. Dawy "Inter-Operator Resource Management for Millimeter Wave, Multi-Hop Backhaul Networks," *IEEE Transactions on Wireless Communications*, accepted and to appear, May. 2017. Available at `https://arxiv.org/pdf/1704.03620`.

[J3] O. Semiari, W. Saad, S. Valentin, M. Bennis, and H. V. Poor "Context-Aware Resource Allocation in Small Cell Networks: How Social Metrics Improve Wireless Resource Allocation," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 5927-5940, Nov. 2015. Available at `https://arxiv.org/abs/1505.04220`.

[J4] O. Semiari, W. Saad, M. Bennis, and B. Maham "Caching Meets Millimeter Wave Communications for Enhanced Mobility Management in 5G Networks," submitted to an *IEEE Transaction*, March. 2017. Available at `https://arxiv.org/abs/1701.05125`.

1.3.2 Conference papers

[C1] O. Semiari, W. Saad, and M. Bennis, "Downlink Cell Association and Load Balancing for Joint Millimeter Wave-Microwave Cellular Networks," in *Proc. of the IEEE Global Communications Conference (GLOBECOM'16), Mobile and Wireless Networks Symposium*, Washington DC, USA, Dec. 2016.

[C2] O. Semiari, W. Saad, and M. Bennis, "Context-Aware Scheduling of Joint Millimeter Wave and Microwave Resources for Dual-Mode Base Stations," in *Proc. of the IEEE International Conference on Communications (ICC'16), Mobile and Wireless Networks Symposium*, Kuala lumpur, Malaysia, May 2016.

[C3] O. Semiari, W. Saad, Z. Dawy, and M. Bennis, "Matching Theory for Backhaul Management in Small Cell Networks with mmWave Capabilities," in *Proc. of the IEEE International Conference on Communications (ICC'15), Mobile and Wireless Networks Symposium*, London, UK, June 2015.

[C4] O. Semiari, W. Saad, S. Valentin, M. Bennis, and B. Maham, "Matching Theory for Priority-based Cell Association in the Downlink of Wireless Small Cell Networks," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*, Florence, Italy, May 2014.

[C5] O. Semiari, W. Saad, and S. Valentin, "On Self-Organizing Resource Allocation for Social Context-Aware Small Cell Networks," in *Proc. of the 1st KuVS Workshop on Anticipatory*

Networks, Stuttgart, Germany, September 2014.

[C6] O. Semiari, W. Saad, M. Bennis, and B. Maham “Mobility Management for Heterogeneous Networks: Leveraging Millimeter Wave for Seamless Handover”, submitted to *an IEEE Conference*, 2017.

[C7] O. Semiari, W. Saad, M. Bennis, and M. Debbah “Performance Analysis of Integrated Sub-6 GHz-Millimeter Wave Wireless Local Area Networks”, submitted to *an IEEE Conference*, 2017.

[C8] T. Zeng, O. Semiari, and W. Saad, “Spatio-Temporal Motifs for Optimized Vehicle-to-Vehicle (V2V) Communications”, submitted to *an IEEE Conference*, 2017.

[C9] T. Zeng, O. Semiari, and W. Saad, “Exploring Spatial Motifs for Device-to-Device Network Analysis (DNA) in 5G Networks”, submitted to *an IEEE Conference*, 2017.

1.4 Outline of the Dissertation

The rest of this dissertation is organized as follows.

- Chapter 2 provides an overview on the framework of matching theory.
- Chapter 3 presents the proposed context-aware resource management scheduling scheme for integrated mmW-μW cellular networks.
- Chapter 4 presents the proposed cost-effective and multi-hop backhaul solution for SCNs that leverage mmW communications.
- In Chapter 5, the proposed mobility management framework for integrated mmW-μW SCNs is presented and its performance is analyzed.
- Chapter 6 addresses the load balancing and cell association problems in heterogeneous SCNs with both mmW and μW RATs.
- In Chapter 7, the proposed MAC protocol is presented for the integrated mmW-μW WLANs. Moreover, the performance of the proposed framework is comprehensively studied.
- Chapter 8 presents the proposed context-aware resource management approach for D2D-enabled SCNs that exploit user social context information.
- Chapter 9 concludes the dissertation.

Note: The notations used in the subsequent chapters will be specific to the chapter in which they are introduced.

Chapter 2

Matching Theory

Prior to providing an in-depth discussion for each of the aforementioned problems in Section 1.2, we overview matching theory [72] as a powerful mathematical tool that will prove useful in addressing the resource management problems studied in this dissertation. As we discussed in Chapter 1, various emerging technologies such as dense SBS deployments, D2D communications, caching, mmW communications, and others, will shape future wireless networks. Nonetheless, a successful integration of these technologies in one unified system is contingent upon adopting novel resource management techniques that can capture the following characteristics:

- *Distributed implementation:* Existing cellular networks use MBSs as coordination points that can exchange resource allocation information among one another, manage inter-cell interference, and allocate network resources (frequency channels, time slots, power) to the users. However, such a conventional centralized resource management will not effectively work in future wireless networks, due to the following reasons: 1) Dense deployment of SBSs requires a substantial backhaul infrastructure to provide full coordination among neighboring cells which may not be viable, 2) Emerging wireless networks are expected to leverage the unlicensed spectrum (e.g. LTE-U), and thus, maintaining coordination among cellular BSs and Wi-Fi access points (APs) may not be feasible, 3) Moving airborne BSs, such as drones, along with terrestrial SBSs will change the conventional macrocell network architectures and make the centralized network optimization more challenging.
- *User-centric:* D2D communication is one of the central components for emerging wireless networks that can increase the spectral efficiency and reduce delay. In addition to these benefits, D2D communication will be a key enabler for proximity services, such as smart-home applications, that allows the direct communication among devices. In such ad-hoc network scenarios, BSs/APs may provide minimum coordination and wireless devices must perform critical tasks, such as nearby user discovery, channel estimation, as well as resource allocation. [27].
- *Fast convergence:* Dense deployments of SBSs in HetNets, uplink/downlink decoupling, and association of a single user to multiple SBSs are some of the new features of emerging

wireless networks. These enabling techniques will clearly add to the complexity of cell association, since each user has the flexibility to be associated to more number of BSs, compared with the legacy cellular networks.

- *Network heterogeneity support:* Next-generation wireless networks will provide service by using multiple RATs at the mmW and sub-6 GHz frequencies. As such, resource management algorithms must take into account the heterogeneous characteristics of communication over different frequency bands. In addition, new capabilities for performing vertical handovers across RATs will further increase the complexity of resource management. [73].
- *Context-awareness:* As discussed in Chapter 1, advanced resource management algorithms must enable users to extract CI, and act based on their local information. In fact, collecting CI from all the users and providing network-wide information may results in a large signaling overhead and latency.

Therefore, resource management algorithms for next-generation wireless systems have to be distributed, user-centric, fast, robust, context-aware, and consequently *self-organizing* [74]. In fact, self-organization allows SBSs or even UEs to use some intelligence to make fast resource allocation decisions [27]. To this end, new optimization schemes and game-theoretic solutions are introduced in the literature to address the specific challenges of the future HetNets. However, optimization problems are mostly suitable for centralized implementations provided with the network-wide information, which may result in a significant overhead and complexity. This become even more important when dealing with combinatorial integer programming problems such as channel allocation [74].

During the past few years, *matching theory* is envisioned as a promising approach to solve the resource allocation problems in HetNets [74–76]. The rest of this chapter is organized as follows. In 2.1, we first review the fundamentals of matching theory. In Section 2.2, we elaborate the benefits and challenges of using matching theory to solve resource allocation problems in context-aware HetNets. Note that the various developments of matching algorithms are provided individually in Chapters 3, 4, 5, 6, and 8.

2.1 Fundamentals of Matching Theory

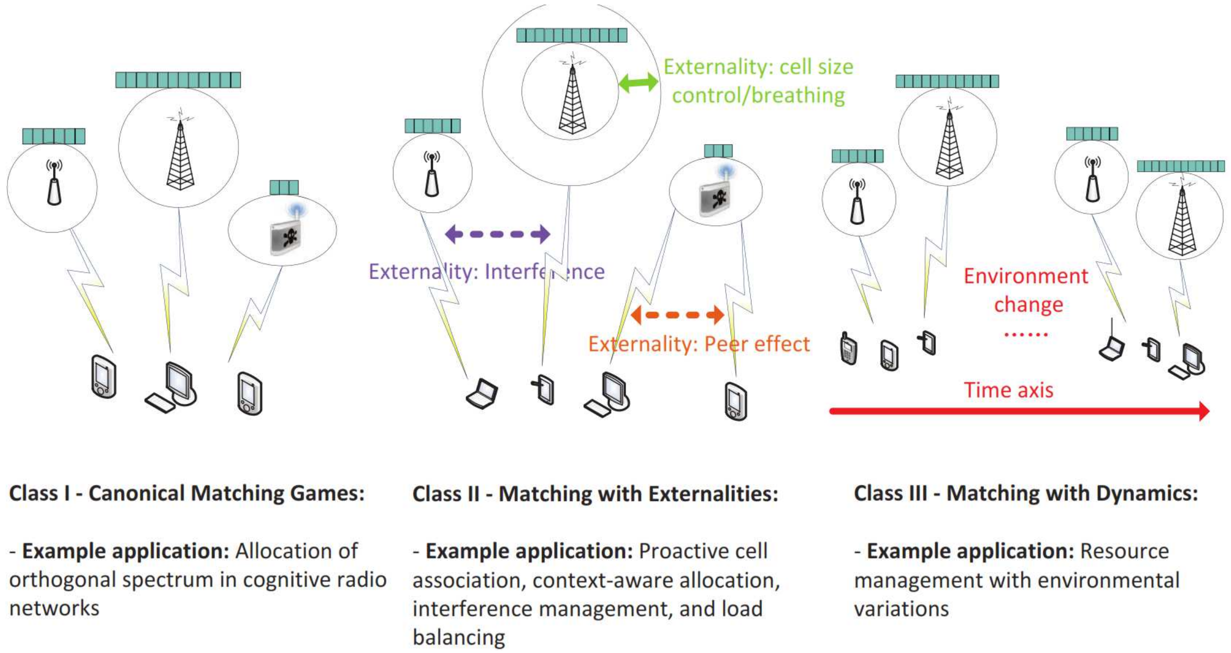
Matching theory is a mathematical framework in economics and applied mathematics to study the formation of mutual beneficial relationships and in particular to solve assignment problems. In wireless communications literature, matching theory has recently become attractive to solve resource allocation problems due to exhibiting useful properties, as we discuss in detail.

In wireless networks, we are interested in *matching problems* that assign network resources (e.g. time slots, frequency channels, power, etc.) to the demanding entities (e.g. devices, stations). The goal of the resource allocation matching problem is to optimally allocate the resources to the users, given the constraints of the wireless network. Following, we define the key components and terminologies of every matching problem in the context of wireless resource allocation:

- **Two disjoint sets of players:** Matching problem μ assigns the players in one set (resources $r \in \mathcal{R}$) to the players of the other set (users $m \in \mathcal{M}$). We denote $\mu(m) \subseteq \mathcal{R}$ as the set of all resources allocated to user m. Similarly, $\mu(r) \subseteq \mathcal{M}$ denotes the set of all users that are assigned to resource r.
- **Quota:** For each player, quota determines the maximum number of players with which it can be matched. Let q_m and q_r be the quota of user m and resource r, respectively. Then, $|\mu(m)| \leq q_m$ and $|\mu(r)| \leq q_r$.
- **Preference relation and strategy:** A player uses the preference relation $\succ$ to rank the players of the other set. This ranking, which is called the *strategy* of the player, is based on some metrics that are important for the player. Each player can quantify its preference relation by assigning a utility U to each player of the other set. Let $\succ_m$ and $U_m(r)$ denote, respectively, the preference relation of user m and the utility that user m gives to resource r. Then, $r_1 \succ_m r_2$ if and only if $U_m(r_1) \geq U_m(r_2)$.
- **Utility function:** Determines the utilities each player assigns to the players of the other set. In wireless networks, $U_m(.) : \mathcal{R} \longrightarrow \mathbb{R}^n$, where $n = |\mathcal{R}|$. In essence, $U_m(.)$ is the objective function of user m, thus, m tries to be matched to the resource r that maximizes $U_m(.)$.
- **Solution of the matching problem:** The solution of the matching problem is a function $\mu : \mathcal{M} \longrightarrow \mathcal{R}$, such that 1) $\forall m \in \mathcal{M}, \mu(m) \subseteq \mathcal{R}$ and $|\mu(m)| \leq q_m$, 2) $\forall r \in \mathcal{R}, \mu(r) \subseteq \mathcal{M}$ and $|\mu(r)| \leq q_r$, and 3) $r \in \mu(m)$ if and only if $m \in \mu(r)$. We denote $(m, r) \in \mu$ if m and r are matched, and $(m, r) \notin \mu$ otherwise.
- **Two-sided stable matching solution:** Consider $(m, r) \notin \mu$, such that $\exists m' \in \mu(r), m \succ_r m'$ and $\exists r' \in \mu(m), r \succ_m r'$, then m and r can *block* the matching μ by leaving their current assigned player, r' and m', and creating a new matched pair (m, r). Matching μ^* is two-sided stable, if there exists no pair of players that blocks μ^*.

One common method to classify matching problems is based on the quota of the players. Matching problem is *one-to-one*, if the quota of all players is one. In addition, matching problem is *many-to-one*, if the quota of at least one player is greater than one, while for the players of the other set quota is one. Finally, if there is at least one player per each set with a quota greater than one, then the problem is called *many-to-many* matching. For one-to-one and many-to-one matching problems, the *deferred acceptance (DA) algorithm* proposed in [72], the seminal work by Gale and Shapely, is always guaranteed to converge to a two-sided stable matching.

In [77], a new wireless-oriented method for classifying matching problems is presented, as shown in Fig. 2.1. This method classifies matching problems into three classes: 1) Canonical matching, 2) Matching with externalities, and 3) Matching with dynamics. Simply stated, canonical matching represents the basic form of matching problems in which the preference of the players do not vary within the timeframe of the resource allocation (i.e., fixed strategies). Moreover, class II of matching problems includes scenarios in which preferences of players are interdependent. Hence, players may change their preferences (i.e. varying strategies) within the timeframe of the resource allocation. The third class, i.e. matching with dynamics, represents matching problems in which the strategy of players in the current resource allocation may depend on the strategies in the past resource allocations.


Externality: cell size control/breathing
Externality: Interference
Externality: Peer effect
Environment change
......
Time axis

Class I - Canonical Matching Games:

- **Example application:** Allocation of orthogonal spectrum in cognitive radio networks

Class II - Matching with Externalities:

- **Example application:** Proactive cell association, context-aware allocation, interference management, and load balancing

Class III - Matching with Dynamics:

- **Example application:** Resource management with environmental variations

Figure 2.1: Wireless-oriented classification of matching problems [77].

2.2 Matching Theory for Wireless Resource Allocation: Opportunities and Challenges

As we discussed earlier in this chapter, self-organizing frameworks must be developed to optimize the resource management in dense, user-centric, and heterogeneous next-generation wireless networks. As such, wireless networks may need a fundamental paradigm shift from the conventional centralized network optimization into robust, fast, and distributed resource management. In this dissertation, we advance matching theory and prove it useful to solve various challenging resource allocation problems for emerging wireless networks. In particular,

- In Chapter 3, we show how matching theory can be applied to solve an NP-hard scheduling problem with minimum unsatisfied relations. We show that this problem belongs to a class of *matching games with externalities*, and propose a novel algorithm to solve it. In addition, we prove that the proposed matching theoretic framework converges in polynomial time and guarantees a two-sided stable allocation of frequency resources to user applications.
- In Chapter 4, we solve the complex multi-hop backhaul problem by proposing a framework based on matching theory. In fact, we pose the original problem as two interdependent matching games to solve network formation and resource allocation problems. Beyond fast convergence as well as guaranteeing stability, another key feature of the proposed distributed solution is to allow multiple network operators to share the backhaul resources and reduce the backhauling cost of their networks. In fact, due to the conflict-of-interests among network operators to either increase their revenue or decrease the backhauling cost, centralized

solutions may not be feasible. That is because competing operators may not share a control center that manages backhaul resources and has access to the operators' confidential information, such as pricing mechanisms.

- In Chapter 5, we show that the mobility management problem in HetNets can be formulated as a *dynamic matching game*, in which mobile users play a key role on choosing the handover strategy. In addition, we show that the concept of dynamic stability is important in distributed handover mechanisms to effectively offload mobile users' traffic from the MBS to small cells.
- In Chapter 6, we demonstrate how matching theory can be used to solve the load balancing problem in future wireless networks with both mmW and sub-6 GHz RATs. We show that the cell association with load balancing can be formulated effectively as a *matching game with minimum quota constraints*. By using this framework, we prove that the proposed self-organizing cell association scheme can achieve significant performance gains, compared with conventional association approaches with cell range expansion.
- In Chapter 8, we show that exploiting the social context information and the formation of social clusters will result in interdependent utilities for users in a D2D-enabled small cell network. In this regard, we propose a resource allocation framework, based on matching theory, that successfully captures the *peer effects* induced by the social context information. Moreover, we show that the proposed self-organizing approach yields a comparable performance to the optimal solution.

Aside from these promising aspects of matching theoretic resource management, one may need to address the following challenges when dealing with the matching theory problems:

- For some practical scenarios in wireless networks, it may be difficult or even infeasible to find a stable solution. For canonical matching games, classical algorithms such as deferred acceptance algorithm can be applied to find a stable solution. Nonetheless, as we show in the rest of this dissertation, this algorithm fails to converge to a stable matching in many important network scenarios, e.g., for matching games with externalities.
- Matching theory solutions are built upon an iterative signaling mechanism between two sides of the game, for example, users and BSs. While it is typically assumed that such signaling is performed over error-free control channels, any error in decoding matching signals may substantially delay the convergence of algorithm.
- Last but not the least, matching theory algorithms may yield a sub-optimal solution. Therefore, it is imperative to find the performance gap between a matching theoretic approach and the optimal solution. For the problems that we study in this dissertation, we show that our proposed frameworks will achieve a close-to-optimal performance.

With that in mind, next, we present the first research work that addresses the scheduling problem in integrated mmW and sub-6 GHz wireless cellular networks.

Chapter 3

Context-Aware Scheduling of Joint Millimeter Wave and Microwave Resources for Dual-Mode Base Stations

3.1 Background, Related Works, and Summary of Contributions

Communication at high frequency, mmW bands is seen as promising approach to overcome the scarcity of the radio spectrum while providing significant capacity gains for tomorrow's wireless cellular networks [7, 10, 27]. However, field measurements [7] have shown that the availability of mmW links can be highly intermittent, due to various factors such as blockage by different obstacles. Therefore, meeting the stringent QoS constraints of delay-sensitive applications, such as HDTV and video conferencing, becomes more challenging at mmW frequencies compared to sub-6 GHz frequencies [7, 26, 30–34].

Such strict requirements can be achieved by deploying dual-mode SBSs that can support high data rates and QoS by leveraging the available bandwidth at both mmW and μW frequency bands [26]. Indeed, in order to provide robust and reliable communications, mmW networks must coexist with small cell LTE networks that operate at the conventional μW band [26, 32–35]. However, differences in signal propagation characteristics and in the available bandwidth lead to a significant difference in the achievable rate and the QoS over mmW and μW frequency bands, thus, yielding new challenges for joint mmW-μW user scheduling [35, 78]. In addition, QoS provisioning in dual-mode mmW-μW networks requires overcoming two key challenges: 1) a joint scheduling over both frequency bands is required, since resource allocation over one band will affect the allocation of the resources over the other frequency band and 2) the QoS constraints per user application (UA) will naturally dictate whether the traffic should be served via mmW resources, μW resources, or both. Therefore, robust and efficient scheduling algorithms for dual-mode SBSs are required that exploit *context information per UA*, including the CSI, maximum tolerable delay, and the required load to maximize users' QoE.

3.1.1 Related works

The work in [26] provides an overview on possible mmW-μW dual-mode architectures that can be used to transmit control and data signals, respectively, at μW and mmW frequency bands. To cope with the intermittent mmW link quality, the authors in [79] formulate the handover decision problem as a Markov decision process (MDP) in mmW networks. In addition, the work in [80] studies the problem of RAT selection and traffic aggregation where each user can simultaneously be connected to multiple BSs. In [36], the authors develop an RAT selection scheme for mmW-μW networks via a multi-armed bandit problem that aims to minimize the cost of handoffs for the UEs. Furthermore, the authors in [81] propose a cross-layer resource allocation scheme for full-duplex communications at the 60 GHz mmW frequency band.

Although interesting, the body of work in [26], [36,79,80], and [81] does not address the scheduling problem in mmW-μW networks. In fact, [36,79,80] focus only on the cell association problem without taking into account, explicitly, the joint allocation of mmW and μW resources. Moreover, existing works such as in [79] and [81] are solely focused on the mmW network, while completely neglecting the impact of the communications over the μW frequencies.

In [31], the authors propose an energy-efficient resource allocation scheme for cellular networks, leveraging both μW and unlicensed 60 GHz mmW bands. In [35], the resource allocation problem for ultra-dense mmW-μW cellular networks is studied under a model in which the cell association is decoupled in the uplink for mmW users. However, this work does not consider any QoS constraint in mmW-μW networks. The problem of QoS provisioning for mmW networks is studied in [32–34], and [75]. In [32], the authors propose a scheduling scheme that integrates device-to-device mmW links with 4G system to bypass the blocked mmW links. The work in [33] presents a mmW system at 60 GHz for supporting uncompressed high-definition (HD) videos for WLANs. In [34], the authors evaluated key metrics to characterize multimedia QoS, and designed a QoS-aware multimedia scheduling scheme to achieve the trade-off between performance and complexity.

Nonetheless, [33] and [34] do not consider multi-user scheduling and multiple access in dual-mode networks. In addition, conventional scheduling mechanisms, such as in [31–33], and [34], identify each UE by a single traffic stream with a certain QoS requirement. In practice, however, recent trends have shown that users run multiple applications simultaneously, each with a different QoS requirement. Although the applications at a single device experience the same wireless channel, they may have different QoS requirements. For example, the QoS requirements for an interactive video call are more stringent than updating a background application or downloading a file. With this in mind, each user's QoE must be defined as a function of the number of QoS-satisfied UAs. Accounting for precise, application-specific QoS metrics is particularly important for scheduling mmW resources whose channel is highly variable, due to large Doppler spreads and short channel coherence time. In fact, conventional scheduling approaches fail to guarantee the QoS for multiple applications at a single UE.

3.1.2 Summary of contributions

The main contribution of this chapter is to propose a novel, context-aware scheduling framework for enabling a dual-band base station to jointly and efficiently allocate both mmW and μW resources to user applications. This proposed context-aware scheduler allows each user to seamlessly run multiple applications simultaneously, each having its own distinct QoS constraint. To this end, the proposed scheduling problem is formulated as an optimization problem with minimum unsatisfied relations (min-UR) and the goal is to maximize the number of satisfied UAs. To solve this NP-hard problem, a novel scheduling framework is proposed that considers a set of context information composed of the UAs' tolerable delay, required load, and the LoS probability, to jointly select and schedule UAs over the μW and the mmW frequency bands. The resource allocation problem at μW band is modeled as a matching game that aims to assign resource blocks (RBs) to the candidate UAs. To solve this game, a novel algorithm is proposed that iteratively solves the UA selection and the resource allocation problems. We show that the proposed algorithm is guaranteed to yield a two-sided stable matching between UAs and the μW RBs. Over the mmW band, the scheduler assigns priority, based on the context information, to the remaining UAs that were not scheduled over the μW band. Consequently, over the mmW band, we show that the scheduling problem can be cast as a 0-1 Knapsack problem. To solve this problem, we then propose a novel algorithm that allocates the mmW resources to the selected UAs. Moreover, we show that the proposed, two-stage scheduling framework can solve the context-aware dual-band scheduling problem in polynomial time with respect to the number of UAs. Simulation results show that the proposed approach significantly improves the QoS per application, compared to the proportional fair and round robin schedulers.

The rest of this chapter is organized as follows. Section 3.2 presents the problem formulation. Section 3.3 presents the proposed context-aware scheduling solution over the μW band. The proposed context-aware scheduling solution over the mmW band is proposed in Section 3.4. Simulation results are analyzed in Section 3.5. Section 3.6 concludes the chapter.

3.2 System Model

Consider the downlink of a dual-mode SBS that operates over both μW and mmW frequency bands. The coverage area of the SBS is a planar area with radius d centered at $(0,0) \in \mathbb{R}^2$. Moreover, a set $\mathcal{M}$ of M UEs is deployed randomly and uniformly within the SBS coverage. UEs are equipped with both mmW and μW RF interfaces which allow them to manage their traffic at both frequency bands [82]. The antenna arrays of mmW transceivers can achieve an overall beamforming gain of $\psi(y_1, y_2)$ for a LoS UE located at $(y_1, y_2) \in \mathbb{R}^2$ [10]. Meanwhile, the μW

Table 3.1: Variables and notations

Notation	**Description**	**Notation**	**Description**
M	Number of UEs	$\mathcal{M}$	Set of UEs
κ_m	Number of UAs per UE m	A	Total number of UAs
τ	Time slot duration	τ'	Beam training overhead
L	Path loss	π_t	Scheduling decision at time slot t
K_1	Number of μW RBs	$\mathcal{K}_1$	Set of μW RBs
K_2	Number of mmW RBs	$\mathcal{K}_2$	Set of mmW RBs
$\mathcal{G}_{t,1}$	Set of UAs to be scheduled over μW band	$\mathcal{G}_{t,2}$	Set of UAs to be scheduled over mmW band
w_1	Bandwidth of μW RBs	w_2	Bandwidth of mmW RBs
g_{kt}	μW channel over RB k at time slot t	$\zeta \in \{0,1\}$	$\zeta = 1$ if link is LoS, otherwise, $\zeta = 0$.
h_{kt}	mmW channel over RB k at time slot t	$p_{k,1}$	Transmit power over μW RB k
ρ_a	LoS probability of the link for UA a	$p_{k,2}$	Transmit power over mmW RB k
P_1	Total transmit power over μW band	P_2	Total transmit power over mmW band
$\boldsymbol{r}$	Q-learning reward vector	J	Number of QoS classes
$\mathcal{A}$	Set of all UAs across all UEs	$\mathcal{A}_j$	Set of UAs with j-th QoS class
b_a	Total required bits for UA a	$b_a^{\text{rec}}(t)$	Received bits by UA a during time slot t
$\lambda_{t,1}$	Number of satisfied UAs over μW band	$\lambda_{t,2}$	Number of satisfied UAs over mmW band

transceivers have conventional single element, omni-directional antennas to maintain low overhead and complexity at the μW frequency band [83]. In our model, each UE $m \in \mathcal{M}$ runs κ_m UAs. We let $\mathcal{A}$ be the set of all UAs with $A = \sum_{m \in \mathcal{M}} \kappa_m$ as the total number of UAs across all UEs.

3.2.1 Channel model and multiple access at mmW and μW frequency bands

The downlink transmission time is divided into time slots of duration τ. For the μW band, we consider an orthogonal frequency division multiple access (OFDMA) scheme in which multiple UAs can be scheduled over K_1 resource blocks (RBs) in the set $\mathcal{K}_1$ at each time slot with duration τ. Therefore, the achievable *μW rate* for an arbitrary UA a at RB k and time slot t is:

$$R_a(k,t) = w_1 \log_2 \left(1 + \frac{p_{k,1} |g_{kt}|^2 10^{-\frac{L_1(y_1,y_2)}{10}}}{w_1 N_0} \right). \tag{3.1}$$

Here, w_1 is the bandwidth of each RB at μW band, and g_{kt} is the Rayleigh fading channel coefficient over RB k at time slot t. The total transmit power at μW band, P_1, is assumed to be distributed uniformly among all RBs such that $p_{k,1} = P_1/K_1$. This uniform power allocation assumption is due to the fact that at a high SNR regime, as is expected in small cells with relatively short-range links, optimal power allocation policies such as the popular water-filling algorithm will ultimately converge to a uniform power allocation [84]. The path loss $L_1(y_1, y_2)$ follows the log-distance model with parameters α_1, β_1, and ξ_1^2 that represent, respectively, the path loss exponent, the path loss at 1 meter distance, and the variance of the shadowing for the μW band.

Over the mmW band, directional transmissions are inevitable to overcome the significantly high path loss at the mmW frequencies. Therefore, the multiple access scheme at the mmW band should support directional transmissions, while maintaining low complex designs for transceivers. Thus, the SBS uses a time division multiple access (TDMA) scheme to schedule UAs [7], which is in line with the existing standards such as WirelessHD and IEEE 802.15.3c [85]. We let $\mathcal{G}_{t,2}$ be the set of UAs that must be scheduled over the mmW band at slot t. During each time slot, for UAs that are assigned to the mmW band, the SBS transmits to UA $a \in \mathcal{G}_{t,2}$ an OFDM symbol of duration $\tau_{a,t}$ composed of K_2 RBs. In practice, the mmW transceivers must align their beams during a *beam training* phase, in order to achieve the maximum beamforming gain [86]. This training phase will introduce a non-negligible overhead on the TDMA system, which can become particularly significant as the number of mmW users increases. Hence, a beam training overhead time $\tau' < \tau$ is considered per transmission to a UA over the mmW band. In practice, duration of τ' can reach up to 1.54 milliseconds, depending on the beam resolution [86]. Therefore, the effective time for data transmission to UAs in $\mathcal{G}_{t,2}$ will be $\sum_{a\in\mathcal{G}_{t,2}} \tau_{a,t} = \tau - |\mathcal{G}_{t,2}|\tau'$, where $|\mathcal{G}_{t,2}|$ denotes the cardinality of the set $\mathcal{G}_{t,2}$.

The large-scale channel effects over the mmW links follow the popular model of [10]:

$$L_2(y_1, y_2) = \beta_2 + \alpha_2 10\log_{10}(\sqrt{y_1^2 + y_2^2}) + \chi, \tag{3.2}$$

where $L_2(y_1, y_2)$ is the path loss at mmW frequencies for all UAs associated with a UE located at $(y_1, y_2) \in \mathbb{R}^2$. In fact, (3.2) is known to be the best linear fit to the propagation measurement in mmW frequency band [10], where α_2 is the slope of the fit and β_2, the intercept parameter, is the pathloss (dB) for 1 meter of distance. In addition, χ models the deviation in fitting (dB) which is a Gaussian random variable with zero mean and variance ξ_2^2. Overall, the total achievable *mmW rate* for UA a at time slot t is given by

$$R_a(t) = \begin{cases} \sum_{k=1}^{K_2} w_2 \log_2\left(1+\frac{p_{k,2}\psi(y_1,y_2)|h_{kt}|^2 10^{-\frac{L_2(y_1,y_2)}{10}}}{w_2 N_0}\right), & \zeta_{at} = 1, \\ 0, & \zeta_{at} = 0, \end{cases} \tag{3.3}$$

where $\zeta_{at} = 1$ indicates that a LoS link is feasible for UA a, otherwise, $\zeta_{at} = 0$ and the link is blocked by an obstacle. In fact, ζ_{at} is a Bernoulli random variable with probability of success ρ_a, and is identical for all UAs that are run by the same UE. Moreover, w_2 is the bandwidth of each RB, h_{kt} is the Rician fading channel coefficient at RB k of slot t [87], and N_0 is the noise power spectral density. Furthermore, $p_{k,2}$ denotes the SBS transmit power at RB k of mmW frequency band. The total transmit power at mmW band, P_2, is assumed to be distributed uniformly among all RBs, such that $p_{k,2} = P_2/K_2$.

Let $\mathcal{G}_{t,1}$ be the set of UAs that must be scheduled over the μW band at time slot t. During each time slot, a UA can be scheduled only at one frequency band, i.e., $\mathcal{G}_{t,1} \cap \mathcal{G}_{t,2} = \emptyset$.

The proposed dual-band multiple access scheme is shown in Fig. 3.1, where each color identifies a single, distinct UA.

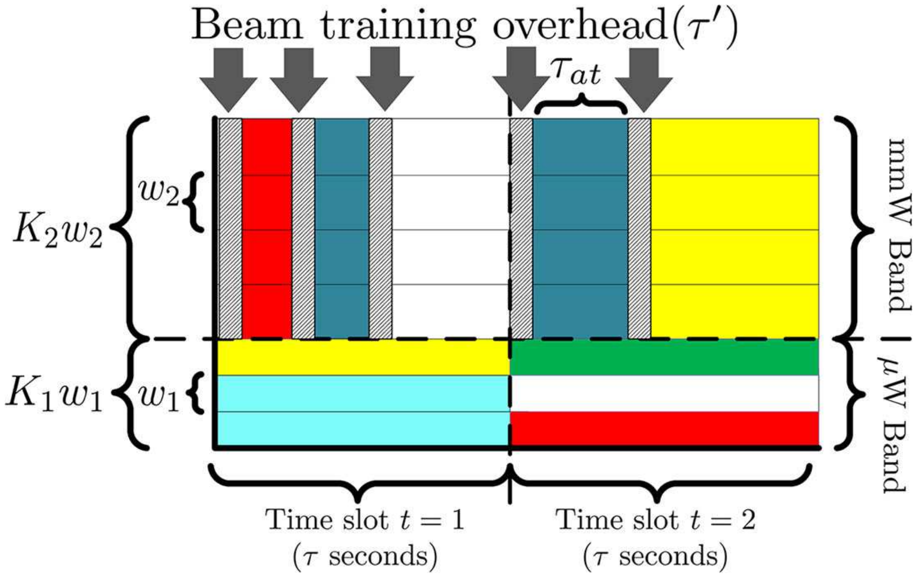


Figure 3.1: Example of resource allocation of the dual-band configuration. Colors correspond to different UAs that may run at different UEs.

3.2.2 Traffic model with QoS constraints

We assume a non-full buffer traffic model in which an arbitrary UA a has a total of b_a bits of data to receive. In addition, each UA has an application-specific tolerable delay which specifies its QoS class, as formally defined next.

Definition 1. The *QoS class*, $\mathcal{A}_j$, is defined as the set of all UAs stemming from all UEs that can tolerate a maximum packet transmission delay of j time slots.

Each UA in our system can belong to one out of a total of J QoS classes, $\mathcal{A}_j, j = 1, \cdots, J$ with $\bigcup_{j=1}^{J} \mathcal{A}_j = \mathcal{A}$, and $\mathcal{A}_j \cap \mathcal{A}_{j'} = \emptyset, j \neq j'$. Due to system resource constraints, not all UAs can be served instantaneously and, thus, a transmission delay will be experienced by the UAs. In essence, to transmit a data stream of size b_a bits to UA $a \in \mathcal{A}_j$, an average data rate of $b_a/j\tau_2$ during j consecutive time slots is needed, otherwise, the UA experiences an outage due to the excessive delay.

The scheduling decision π_t at a given slot t is a function that outputs two vectors $\boldsymbol{x}_t$ and $\boldsymbol{\tau}_t$ that determine, respectively, the resource allocation over μW and mmW bands. In fact, $\boldsymbol{x}_t$ includes the variables $x_{akt} \in \{0, 1\}$ with $a \in \mathcal{A}, k \in \mathcal{K}_1$ where $x_{akt} = 1$ indicates that μW RB k is allocated to UA a at slot t, otherwise, $x_{akt} = 0$. In addition, each element $\tau_{at} \in [0, \tau], a \in \mathcal{A}$, of $\boldsymbol{\tau}_t$ determines the allocated time to UA a over mmW band. The required bits for UA a at slot t, $b_a^{\text{req}}(t)$, depend on the number of bits received during previous slots, $\sum_{t'=0}^{t-1} b_a^{\text{rec}}(t')$. In other words, $b_a^{\text{req}}(t) = b_a - \sum_{t'=0}^{t-1} b_a^{\text{rec}}(t')$, with $b_a^{\text{req}}(1) = b_a$ and $b_a^{\text{rec}}(0) = 0$. For a given policy π_t, the required load at time slot $t+1$, $b_a^{\text{req}}(t+1)$, can be written recursively as

$$b_a^{\text{req}}(t+1) = b_a^{\text{req}}(t) - b_a^{\text{rec}}(t) = b_a^{\text{req}}(t) - \left[\tau \sum_{k=1}^{K_1} R_a(k,t) x_{akt} + R_a(t)\tau_{at}\zeta_{at}\right]. \tag{3.4}$$

From (3.4), we observe that policy π_t depends on the scheduling decisions during previous time slots $\{\pi_1, \pi_2, ..., \pi_{t-1}\}$. Thus, we define $\pi = \{\pi_1, \pi_2, ..., \pi_t, ..., \pi_J\} \in \boldsymbol{\Pi}$ as a long-term *scheduling policy*, where $\boldsymbol{\Pi}$ is the set of all possible scheduling policies.

Next, we use (3.4) to formally define the *QoS criterion* for any UA $a \in \mathcal{A}_j$ as

$$\mathbb{1}(a \in \mathcal{A}_j; \pi) = \begin{cases} 1 & \text{if } \sum_{t'=1}^{j} b_a^{\text{rec}}(t') \geq b_a, \\ 0 & \text{otherwise,} \end{cases} \tag{3.5}$$

where $\mathbb{1}(a \in \mathcal{A}_j; \pi) = 1$ indicates that under policy π, enough resources are allocated to UA $a \in \mathcal{A}_j$ to receive b_a bits within j slots, while $\mathbb{1}(a \in \mathcal{A}_j; \pi) = 0$ indicates that UA a is going to experience an outage. We define the outage set $\mathcal{O}^\pi = \{a | \mathbb{1}(a \in \mathcal{A}_j; \pi) = 0, j = 1, \cdots, J\}$ as the set of UAs in outage.

Prior to formulating the problem, we must note the following inherent characteristics of dual-mode scheduling: 1) If mmW link with a high LoS probability is not feasible for a UE, scheduling over the mmW band can cause outage to the associated UAs, specifically for delay-intolerant UAs, 2) larger range of supported rates is available for UAs compared to the conventional single-band systems. Hence, for some UAs, the required rate exceeds the achievable rate at μW band. Therefore, effective dual-mode scheduling should not only rely solely on CSI, but it must also leverage UA-specific metrics, herein referred to as *context information* as formally defined next.

Definition 2. At any slot t, the tuple $\mathcal{C} = (\mathcal{A}_{j\geq t}, \boldsymbol{b}^{\text{req}}(t), \boldsymbol{\rho})$ defined as *context information*, is composed of the delay constraints of UAs, $\mathcal{A}_{j\geq t} = \bigcup_{j=t}^{J} \mathcal{A}_j$, the required load per UA, $\boldsymbol{b}^{\text{req}}(t) = \{b_a^{\text{req}}(t) | a \in \mathcal{A}_{j\geq t}\}$, and the LoS probability of each UA, $\boldsymbol{\rho} = \{\rho_a | a \in \mathcal{A}_{j\geq t}\}$.

Note that exploiting the context information at any time slot t properly links the scheduling policy π_t to the history, since from (3.5), $\mathcal{C}$ at slot t depends on $\pi_t', t' = 1, \cdots, t-1$.

3.2.3 Problem formulation

Our goal is to find a scheduling policy $\pi^* \in \boldsymbol{\Pi}$ that satisfies (3.5) for as many UAs as possible over J time slots. The general long-term scheduling problem for slots $t = 1, \cdots, J$ can be solved separately at each slot t to find $\pi_t^*(\mathcal{C}, \text{CSI}) = (\boldsymbol{x}_t^*, \boldsymbol{\tau}_t^*)$, while the time-dependency of scheduling decisions is captured by exploiting the context information. Therefore, the scheduling problem at an arbitrary slot t can be formulated as follows:

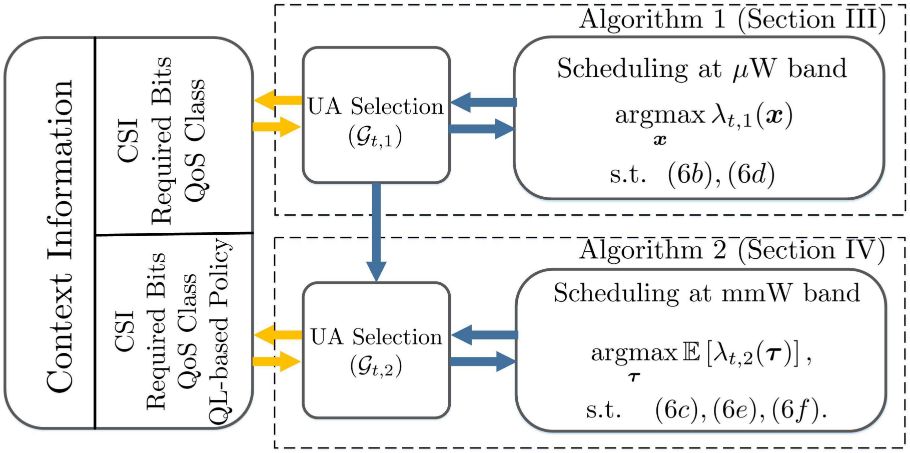


Figure 3.2: The structure of the context-aware scheduler.

$$\underset{\boldsymbol{x}_t,\boldsymbol{\tau}_t}{\text{argmax}}\, \lambda_{t,1} + \mathbb{E}\left[\lambda_{t,2}\right], \tag{3.6a}$$

$$\text{s.t.} \quad \tau \sum_{k=1}^{K_1} R_a(k,t) x_{akt} \geq b_a^{\text{req}}(t), \qquad \forall a \in \mathcal{A}_t \cap \mathcal{G}_{t,1}, \tag{3.6b}$$

$$R_a(t)\tau_{at}\zeta_{at} \geq b_a^{\text{req}}(t-1), \qquad \forall a \in \mathcal{A}_t \cap \mathcal{G}_{t,2}, \tag{3.6c}$$

$$\boldsymbol{x}_t \in \mathcal{X} = \left\{ x_{akt} \in \{0,1\} \middle| \sum_{a\in\mathcal{A}} x_{akt} \leq 1, \sum_{k=1}^{K_1} x_{akt} \leq K_1, \forall a \in \mathcal{A}_{j\geq t} \right\}, \tag{3.6d}$$

$$\boldsymbol{\tau}_t \in \mathcal{Y} = \left\{ \tau_{at} \in [0,\tau] \middle| \sum_{a\in\mathcal{G}_{t,2}} \tau_{at} + |\mathcal{G}_{t,2}|\tau' \leq \tau, \forall a \in \mathcal{G}_{t,2} \right\}, \tag{3.6e}$$

$$\boldsymbol{x}_t, \boldsymbol{\tau}_t \in \mathcal{Z} = \left\{ \boldsymbol{x}_t \in \mathcal{X}, \boldsymbol{\tau}_t \in \mathcal{Y} \middle| \sum_{k=1}^{K_1} x_{akt}\tau_{at} = 0 \right\}, \tag{3.6f}$$

where $\lambda_{t,1}$ and $\lambda_{t,2}$ denote, respectively, the number of satisfied UAs scheduled at μW and mmW bands at slot t. Given a decision policy π_t, $\lambda_{t,2}$ is a random variable that depends on ζ_{at} at the mmW band. In fact, the expectation in 3.6a is taken over ζ_{at}, for all $a \in \mathcal{G}_{t,2}$. However, $\lambda_{t,1}$ is deterministic, if the slot duration τ is smaller than the μW channel coherence time.

The problem (3.6a)-(3.6f) falls into a class of optimization problems, referred to as *Min-UR*, which are known to be NP-hard [88]. Although linear systems with equality or inequality constraints can be solved in polynomial time, using an adequate linear programming method, least mean squared methods are not appropriate for infeasible systems when the objective is to maximize satisfied relations [88].

With this in mind, we propose a two-stage solution that solves (3.6a)-(3.6e) in polynomial time, as illustrated in Fig. 3.2. The scheduling at μW band is considered first in order to reliably schedule as many UAs as possible with small required loads over the μW band. The motivation for serving

UAs first at μW band is due to the fact that transmissions at μW frequencies are robust against blockage. Unlike μW frequencies, mmW communication is highly susceptible to blockage and, thus, scheduling UAs only at the mmW band can potentially cause outage for delay-intolerant UAs. To this end, a hierarchy scheme is proposed based on the UAs' QoS class, CSI, and the required loads. Moreover, the UAs selection and scheduling are jointly done at SBS using an iterative algorithm. Then, for the remaining UAs that were not scheduled at the μW band, we propose a joint UA selection criterion and scheduling algorithm that introduces a hierarchy to the UAs, based on the context information, and maximizes the number of satisfied UAs.

3.3 Context-Aware UA Selection and Resource Allocation at μW Band

Before scheduling at mmW band, the goal of the scheduler is to first find an allocation $\boldsymbol{x}_t^*$ at each slot t over μW band that satisfies

$$\underset{\boldsymbol{x}_t^*}{\operatorname{argmax}}\, \lambda_{t,1}(\boldsymbol{x}_t^*), \tag{3.7a}$$

$$\text{s.t.} \quad (3.6\text{b}), (3.6\text{d}). \tag{3.7b}$$

The downlink scheduling problem in (3.7a)-(3.7b) is an inconsistent combinatorial problem of matching users to resources which does not admit a closed-form solution and has an exponential complexity [89]. Hence, the solution of (3.7a)-(3.7b) depends on which UAs are chosen to be scheduled at μW band, i.e., the set $\mathcal{G}_{t,1}$. To this end, we introduce a hierarchy for UA selection by grouping the different UAs into the following sets:

$$\mathcal{G}_{t,1}^{(1)} = \{a \in \mathcal{A}_j | j = t, b_a^{\text{req}}(t) > 0\}, \tag{3.8}$$

$$\mathcal{G}_{t,1}^{(2)} = \{a \in \mathcal{A}_j | j > t, b_a^{\text{req}}(t) > 0\}. \tag{3.9}$$

In fact, the UAs in $\mathcal{G}_{t,1}^{(1)}$ have higher priority than $\mathcal{G}_{t,1}^{(2)}$, since they must be served during the current time slot, otherwise, they will experience an outage. In addition, for UAs of the same set, the UA that satisfies the following has the highest priority:

$$a^* = \underset{a}{\operatorname{argmin}} \frac{b_a^{\text{req}}(t)}{\sum_{k \in \mathcal{K}_1} R_a(k,t)}, \tag{3.10}$$

where (3.10) selects UA a^* that minimizes the ratio of the required load to the achievable rate. To ensure that the constraints set for the selected UAs $a \in \mathcal{G}_{t,1}$ is feasible, i.e. $\lambda_{t,1}(\boldsymbol{x}) = |\mathcal{G}_{t,1}|$, the UA selection has to be done jointly while solving (3.7a)-(3.7b). Following, we propose a framework that solves (3.7a)-(3.7b) for a given $\mathcal{G}_{t,1}$.

3.3.1 Scheduling as a matching game

For a selected set of UAs at μW band, $\mathcal{G}_{t,1}$, we propose a novel resource allocation scheme at μW band based on matching theory concept, introduced in Chapter 2 [76, 90–92]. As explained in Chapter 2, a matching game is defined as a two-sided assignment problem between two disjoint sets of players in which the players of each set are interested to be matched to the players of the other set, according to *preference relations*. At each time slot t of our scheduling problem, $\mathcal{K}_1$ and $\mathcal{G}_{t,1}$ are the two sets of players. A preference relation $\succ$ is defined as a complete, reflexive, and transitive binary relation between the elements of a given set. Here, we let $\succ_a$ be the preference relation of UA a and denote $k \succ_a k'$, if player a prefers RB k over RB k'. Similarly, we use $\succ_k$ to denote the preference relation of RB $k \in \mathcal{K}_1$.

In the proposed scheduling problem, the preference relations of UAs depend on both the rate and the QoS constraint which will be captured via well-designed, individual utility functions for UAs and SBS resources, as defined later in this section.

3.3.2 Scheduling at μW band as a matching game

Each scheduling decision $\pi_{t,1}$ determines the allocation of RBs to UAs during time slot t over the μW band. Thus, the scheduling problem at μW frequency band can be defined as a *one-to-many matching game*:

Definition 3. Given two disjoint finite sets of players $\mathcal{G}_{t,1}$ and $\mathcal{K}_1$, the scheduling decision at time slot t, $\pi_{t,1}$, can be defined as a *matching relation*, $\pi_{t,1} : \mathcal{G}_{t,1} \rightarrow \mathcal{K}_1$ that satisfies 1) $\forall a \in \mathcal{G}_{t,1}, \pi_{t,1}(a) \subseteq \mathcal{K}_1$, 2) $\forall k \in \mathcal{K}_1, \pi_{t,1}(k) \in \mathcal{G}_{t,1}$, and 3) $\pi_{t,1}(k) = a$, if and only if $k \in \pi_{t,1}(a)$.

In fact, $\pi_{t,1}(k) = a$ implies that $x_{akt} = 1$, otherwise $x_{akt} = 0$. Therefore, $\pi_{t,1}$ is indeed the scheduling decision that determines the allocation at μW band. One can easily see from the above definition that the proposed matching game inherently satisfies the constraint (3.6d). Next, we need to define suitable utility functions to determine the preference profiles of UAs and RBs. Given matching $\pi_{t,1}$, we define the utility of UA a for $k \in \mathcal{K}_1$ at time slot t as:

$$\mathit{\Psi}_a(k,t;\pi_{t,1}) = \begin{cases} 0 & \text{if } \sum\limits_{k' \in \pi_{t,1}(a)} R_a(k',t)\tau \geq b_a^{\text{req}}(t), \\ R_a(k,t) & \text{otherwise.} \end{cases} \tag{3.11}$$

The utility of μW RBs $k \in \mathcal{K}_1$ for UA $a \in \mathcal{G}_{t,1}$ is simply the rate

$$\mathit{\Phi}_k(a,t) = R_a(k,t). \tag{3.12}$$

Using these utilities, the preference relations of UAs and RBs at a given time slot t will be

$$k \succ_a k' \Leftrightarrow \mathit{\Psi}_a(k,t;\pi_{t,1}) \geq \mathit{\Psi}_a(k',t;\pi_{t,1}) \tag{3.13}$$

$$a \succ_k a' \Leftrightarrow \mathit{\Phi}_k(a,t) \geq \mathit{\Phi}_k(a',t), \tag{3.14}$$

Algorithm 1 Context-Aware UA Selection and Resource Allocation Algorithm at μW Band

```
Inputs: $\mathcal{G}_{t,1}^{(1)}$, $\mathcal{G}_{t,1}^{(2)}$, $\boldsymbol{b}^{\text{req}}(t)$, $R_a(k,t)$.
Outputs: $\boldsymbol{x}$; $\mathcal{G}_{t,1}$.
Initialize: $\mathcal{G}_{t,1} = \emptyset$,
 1: $\mathcal{G}'_{t,1} = \mathcal{G}_{t,1}^{(1)}, \mathcal{K}_a = \mathcal{K}_1, \forall a \in \mathcal{G}'_{t,1}$.
 2: Add UA $a^* \in \mathcal{G}'_{t,1}$ with smallest $b_a^{\text{req}}(t)/\sum_{k\in\mathcal{K}_1} R_a(k,t)$ to $\mathcal{G}_{t,1}$ and remove it from $\mathcal{G}'_{t,1}$.
 3: Update the preference ordering of UAs $a \in \mathcal{G}_{t,1}$ and RBs $k \in \mathcal{K}_1$, using (3.11) and (3.12).
 4: Using $\succ_a$, a UA $a \in \mathcal{G}_{t,1}$ is tentatively assigned to its most preferred RB in $\mathcal{K}_a$.
 5: From the tentative list of UA applicants plus $\pi_{t,1}(k)$ for RB $k$, only the most preferred UA, based on $\succ_k$, is
    assigned to $k$. Next, $k$ is removed from the applicants' $\mathcal{K}_a$ sets.
 6: Each UA $a$ updates $b_a^{\text{req}}(t)$ and $\succ_a$ based on (3.13).
 7: repeat Steps 3 to 6
 8: until $\mathbb{1}_a(\boldsymbol{x}) = 1$, or $\mathcal{K}_a = \emptyset, \forall a \in \mathcal{G}_{t,1}$.
 9: if $\exists a \in \mathcal{G}_{t,1}, \mathbb{1}_a(\boldsymbol{x}) \neq 1$ then
10:     Remove $a^*$ from $\mathcal{G}_{t,1}$ and go to Step 3.
11: end if
12: if $\exists k, \sum_{a\in\mathcal{G}_{t,1}^{(1)}\cup\mathcal{G}_{t,1}^{(2)}} x_{akt} = 0$ then let $\mathcal{G}'_{t,1} = \mathcal{G}_{t,1}^{(2)}$ and go to Step 2.
13: end if
```

for $\forall a, a' \in \mathcal{G}_{t,1}$, and $\forall k, k' \in \mathcal{K}_1$. Given this framework, we propose a joint UA selection and matching-based scheduling algorithm that maximizes $\lambda_{t,1}$.

3.3.3 Proposed context-aware scheduling algorithm at μW band

To solve the proposed game and find a suitable outcome, we use the concept of two-sided *stable matching* between UAs and RBs, defined as follows [90]:

Definition 4. A pair $(a,k) \notin \pi_{t,1}$ is said to be a *blocking pair* of the matching $\pi_{t,1}$, if and only if $a \succ_k \pi_{t,1}(k)$ and $k \succ_a \pi_{t,1}(a)$. Matching $\pi_{t,1}$ is *stable*, if there is no blocking pair.

A stable scheduling decision, $\pi_{t,1}$, ensures fairness for the UAs. That is, if a UA a envies the allocation of another UA a', then a' must be preferred by the RB $\pi_{t,1}(a')$ to a, i.e., the envy of UA a is not justified. For conventional matching problems, the popular DA algorithm is normally used to find a stable matching [74,76,90]. However, DA cannot be applied directly to our problem because it assumes that the quota for each UA is fixed. The quota is defined as the maximum number of RBs that a UA can be matched to. In our problem, however, quotas cannot be predetermined, since the number of RBs needed to satisfy the QoS constraint of a UA in (3.7b) depends on the channel quality at each RB, as well as the context information. In fact, the adopted utility functions in (3.11) depend on the current state of the matching. Due to the dependency of UAs' preferences to the state of the matching, i.e. x_{akt} variables, the proposed game can be classified as a *matching game with externalities* [74]. For matching games with externalities, DA may not converge to a two-sided stable matching. Therefore, a new matching algorithm must be found to solve the problem.

To this end, we propose a novel context-aware scheduling algorithm shown in Algorithm 1.

Algorithm 1 first allocates the RBs to the UAs in $\mathcal{G}_{t,1}^{(1)}$. At every iteration, each UA a^* given by (3.10) is added to the set $\mathcal{G}_{t,1}$ of the matching game. In Steps 4 to 10, the algorithm assigns RBs $k \in \mathcal{K}_1$ to UAs $a \in \mathcal{G}_{t,1}$ as follows. Each UA $a \in \mathcal{G}_{t,1}$ is tentatively assigned to its most preferred RB $k \in \mathcal{K}_a$. Next, from the tentative list of candidate UAs as well as current assignment $\pi_{t,1}(k)$, the scheduler allocates RB k only to the most preferred UA, based on $\succ_k$. The RB k is removed from the set $\mathcal{K}_a$ corresponding to each candidate UA $a \in \mathcal{G}_{t,1}$. Based on the allocated RBs, the UAs update $b_a^{\text{req}}(t)$ and $\succ_a$. This process is repeated until the rate constraints for UAs are satisfied, $\mathbb{1}_a(\boldsymbol{x}) = 1$, or $\mathcal{K}_a = \emptyset$ for UAs $a \in \mathcal{G}_{t,1}$. Then, if some of the RBs are left unallocated, the algorithm follows Steps 2 to 14 to add UAs from $\mathcal{G}_{t,1}^{(2)}$ to $\mathcal{G}_{t,1}$.

Theorem 1. *The proposed Algorithm* 1 *is guaranteed to yield a two-sided stable matching between UAs and* μ*W RBs.*

Proof. See Appendix A.1. □

Given $\mathcal{G}_{t,1}$ by Algorithm 1 at μW band, the scheduling problem at slot t is now reduced to choosing a subset of unscheduled UAs and allocate mmW resources to them such that the number of satisfied UAs is maximized.

3.4 Context-Aware UA Selection and Resource Allocation at mmW Band

We let $\mathcal{G}'_{t,2}=\{a \in \mathcal{A}_{j\geq t} | a \notin \mathcal{G}_{t,1}, b_a^{\text{req}}(t) > 0\}$ be the set of UAs that have not been scheduled over the μW band. Here, the scheduling problem over the mmW band at slot t can be formulated as a stochastic min-UR problem as follows:

$$\underset{\boldsymbol{\tau}_t}{\operatorname{argmax}}\, \mathbb{E}\left[\lambda_{t,2}(\boldsymbol{\tau}_t)\right], \tag{3.15a}$$

$$\text{s.t.} \quad (3.6\text{c}), (3.6\text{e}), (3.6\text{f}). \tag{3.15b}$$

Here, we note that ζ_{at} in (3.6c) is a Bernoulli random variable with success probability ρ_a. Hence, for any allocation $\boldsymbol{\tau}_t$, the number of satisfied constraints $\lambda_{t,2}$ is a random variable. Although the exact distribution of $\lambda_{t,2}$ may not be found for a general infeasible problem as (3.15a)-(3.15b), we can approximate the distribution of outage ratio at slot t, $P_{\text{out,t}} = 1 - [(\lambda_{t,1} + \lambda_{t,2})/A_{j=t}]$, as follows:

Proposition 1. Let $\boldsymbol{\tau}_t$ be a feasible solution for the subset of constraints in (3.15b) associated with UAs $a \in \mathcal{G}_{t,2} \subseteq \mathcal{G}'_{t,2}$. Given $\lambda_{t,1}$ and $0 \leq P_{th} < 1 - \frac{\lambda_{t,1}}{A_{j=t}}$, where P_{th} is an outage threshold, the CDF of the outage ratio at slot t, $F_{P_{out,t}}(P_{th})$ can be approximated by,

$$F_{P_{out,t}}(P_{th}) \approx 1 - \frac{\mathit{\Gamma}\left(\lfloor (1-P_{th})A_t - \lambda_{t,1} + 1\rfloor, \lambda_{ave}\right)}{\lfloor (1-P_{th})A_t - \lambda_{t,1}\rfloor !}, \tag{3.16}$$

where $\lfloor . \rfloor$ is the floor function, $\mathit{\Gamma}(.,.)$ is the incomplete gamma function, and

$$\lambda_{ave} = \mathbb{E}\left[\lambda_{t,2}(\boldsymbol{\tau}_t)\right] = \sum_{a \in \mathcal{G}_{t,2}} \rho_a. \tag{3.17}$$

Proof. See Appendix A.2. □

From (3.17) and (3.29), we can see that the objective function increases as UAs with higher ρ_a are satisfied, however, the approximation of the distribution becomes less accurate.

We note that if LoS probabilities ρ_a are known by the SBS, the proposed scheduling problem over the mmW band becomes equivalent to a 0-1 stochastic Knapsack optimization problem [93]. However, in practice, the explicit values of ρ_a may not be available at the SBS. In Section 3.4.2, we will introduce a learning approach using which the SBS can determine if $\rho_a \geq \rho_{th}$, where ρ_{th} is a constant value. By learning which UAs satisfy $\rho_a \geq \rho_{th}$, the SBS assigns priority to the UAs that are more likely to be at a LoS link from the SBS. This information along with the QoS classes of UAs will allow the scheduler to group UAs into the following non-overlapping subsets:

$$\mathcal{G}_{t,2}^{(1)} = \{a \in \mathcal{A}_{j=t} \cap \mathcal{G'}_{t,2} | \rho_a \geq \rho_{th}\}, \tag{3.18}$$

$$\mathcal{G}_{t,2}^{(2)} = \{a \in \mathcal{A}_{j=t} \cap \mathcal{G'}_{t,2} | \rho_a < \rho_{th}\}, \tag{3.19}$$

$$\mathcal{G}_{t,2}^{(3)} = \{a \in \mathcal{A}_{j>t} \cap \mathcal{G'}_{t,2} | \rho_a \geq \rho_{th}\}, \tag{3.20}$$

$$\mathcal{G}_{t,2}^{(4)} = \{a \in \mathcal{A}_{j>t} \cap \mathcal{G'}_{t,2} | \rho_a < \rho_{th}\}. \tag{3.21}$$

In fact, the SBS will adopt a greedy approach that assigns priority to sets $\mathcal{G}_{t,2}^{(i)}$ with $i = 1$ as highest and $i = 4$ as lowest priority. That is due to the fact that UAs in $\mathcal{G}_{t,2}^{(1)}$ cannot tolerate further delays. Moreover, they belong to UEs with high possibility of LoS access to SBS. In addition, UAs in $\mathcal{G}_{t,2}^{(2)}$ are in second priority, since they cannot tolerate more delay, while having a low ρ_a. Moreover, UAs $\mathcal{G}_{t,2}^{(3)}$ are assigned to a third priority, since they can tolerate more delays and have high probability to be at LoS mmW link with SBS. The least priority is assigned to UAs in $\mathcal{G}_{t,2}^{(4)}$ as they can tolerate further delays, while having low ρ_a.

Furthermore, for the UAs of the same set, the highest priority is given to a UA a^* that satisfies:

$$a^* = \operatorname*{argmin}_a \frac{b_a^{\text{req}}(t)}{R_a(t)}. \tag{3.22}$$

In other words, the SBS selects the UA that requires the least time resource to be satisfied. Similar to μW band scheduling, the SBS must ensure that the constraints set for selected UAs $a \in \mathcal{G}_{t,2}$ is feasible. Therefore, the UA selection has to be done jointly while solving (3.15a)-(3.15b). Next, we propose a joint UA selection and scheduling algorithm at mmW band.

Algorithm 2 Context-Aware UA Selection and Resource Allocation Algorithm at mmW Band

```
Inputs: $\mathcal{G}_{t,2}^{(i)}, i = 1, ..., 4$, $\boldsymbol{b}^{\text{req}}(t)$, $R_a(t)$.
Output: $\boldsymbol{\tau}; \mathcal{G}_{t,2}$.
Initialize: $\mathcal{G}_{t,2} = \emptyset$.
for $i = 1; i \leq 4; i++$ do
    for $j = 1 : |\mathcal{G}_{t,2}^{(i)}|$ do
        Find UA $a^* \in \mathcal{G}_{t,2}^{(i)}$ from (3.22), set $\tau_{a^*,t} = b_{a^*}^{\text{req}}(t)/R_{a^*}(t)$ and add $a^*$ to $\mathcal{G}_{t,2}$.
        if (3.6e) is not satisfied then
            Remove $a^*$ from $\mathcal{G}_{t,2}$. Break.
        end if
    end for
end for
```

3.4.1 Proposed context-aware scheduling algorithm over the mmW band

Over the mmW band, the objective is to serve as many UAs as possible in order to offload more traffic from the μW band, subject to UAs delay constraints. With this in mind, we propose Algorithm 2 to solve (3.15a)-(3.15b). The algorithm follows the priority criterion introduced in (3.19)-(3.21). Starting with the set $\mathcal{G}_{t,2}^{(1)}$, the scheduling process is a 0-1 Knapsack problem composed of $|\mathcal{G}_{t,2}^{(1)}|$ items all with the same benefit ρ_{th} and weights equal to the required time $\tau_{a,t} = \frac{b_a^{\text{req}}(t)}{R_a(t)}$. This problem can be simply solved by sorting the required time in increasing order and adding UAs one by one to the set $\mathcal{G}_{t,2}$. The algorithm follows the process for the remaining sets and converges, once the entire mmW time slot duration is allocated and no additional time is available for more UAs. From Algorithms 1 and 2, we observe that resource allocation at any time slot affects the scheduling at both mmW and μW bands for the subsequent time slots. Therefore, the proposed UA selection and scheduling schemes at one frequency band are not independent of those at the other frequency band, thus requiring *joint scheduling* for the dual-mode system.

The above solution requires the SBS to determine for which UAs the condition $\rho_a \geq \rho_{th}$ is satisfied. Next, we introduce a learning scheme that enables the UEs to obtain this information by monitoring successful LoS transmissions from the SBS over time and send it to the SBS. Clearly, ρ_a is the same for the UAs that run at an arbitrary UE, since they experience the same wireless channel.

3.4.2 Q-learning model to evaluate the LoS probability

In a real-world cellular network, the UEs will be surrounded by many objects and, thus, the SBS may never know in advance whether an LoS mmW link will be available or not. Therefore, scheduling UAs of UEs that are experiencing a high blockage probability not only wastes network resources, it may drastically degrade QoS for delay intolerant UAs.

In practice, ρ_a depends on many parameters such as the distance between the UE and the SBS, or blockage by human or other surrounding objects. Although finding a closed-form relation of ρ_a with these parameters may not be feasible in general, the UEs can learn whether they have a high

LoS probability based on transmissions from the SBS over time. The UEs will then update and send this information to the SBS at each time slot. Clearly, a simple averaging over time would not work, since the environment is dynamic and ρ_a may change over time. To this end, we propose a learning framework, based on Q-learning (QL) [94], in order to determine UAs with $\rho_a \geq \rho_{th}$ without knowing the actual ρ_a values. QL is a reinforcement learning algorithm that determines optimal policy without detailed modeling of the system environment [94, 95]. The proposed QL model is formally defined by the following key elements:

- *Agents:* UEs $m \in \mathcal{M}$.
- *States:* Depending on whether a UA of a given UE is being scheduled over μW or mmW bands, there are three possible states for the UA: 1) UA is served by the SBS over a LoS mmW link (S_1), 2) UA is scheduled over mmW band, but no LoS link is possible (S_2), and 3) UA is scheduled over μW band (S_3).
- *Action:* At any state, a UE can make a decision d chosen from a set $\mathcal{D} = \{d_1, d_2\}$ where d_1 and d_2, respectively, stand for whether to schedule this user's UAs at the current frequency band or switch to the other frequency band.
- *State transition probability:* $T(S_i, d, S_j)$ denotes the probability of transition from state S_i to S_j if decision $d \in \mathcal{D}$ is chosen by the UE. Hence, $T(S_i, d_2, S_3) = 1$ for $i = 1, 2$, and $T(S_3, d_2, S_1) = 1 - T(S_3, d_2, S_2) = \rho_a$. In addition, $T(S_3, d_1, S_3) = 1$ and $T(S_i, d_1, S_1) = 1 - T(S_i, d_1, S_2) = \rho_a$ for $i = 1, 2$.
- *Reward:* The UE receives rewards $\boldsymbol{r} = [r_1, -r_2, r_3]$, respectively, for each of its UAs being at states S_1, S_2, and S_3, where $r_2 > r_1 > r_3 > 0$. The rewards are assumed the same for all UAs $a \in \mathcal{A}$. The reward values affect both the convergence and the policy. For instance, for large negative rewards, i.e., $r_2 \gg r_3$, the optimal policy for the UA is to choose μW, even for large ρ_a values. The long-term reward for choosing mmW band by UA $a \in \mathcal{A}$ is $r_1\rho_a - (1 - \rho_a)r_2$. Therefore, we can set $\boldsymbol{r}$ such that only for $\rho_a \geq \rho_{th}$, mmW band be preferred by UA a. That is, $r_1\rho_a - (1 - \rho_a)r_2 \geq r_3$ which implies

$$\rho_{th} = \frac{r_3 + r_2}{r_1 + r_2}, \quad r_2 > r_1 > r_3 > 0. \tag{3.23}$$

At any time slot, each UA that is selected for scheduling will explore one of the three states. Consequently, this UA's corresponding UE will achieve a reward associated with the current residing state. We note that the UEs do not have any prior knowledge about the transition probabilities in advance. However, QL provides a model-free approach that instead of estimating ρ_a, it allows UE to find the best decision while residing at each state. This is done by the notion of Q-values $Q(S, d)$ which represents the value of decision d while being at state S. Starting from an initial Q-values, UA can find true values via an iterative process as follows:

$$Q(S, d) \leftarrow (1 - \alpha)Q(S, d) + \alpha \left[\boldsymbol{r}(S') + \gamma \max_{d'} Q(S', d')\right], \tag{3.24}$$

where α and γ are predetermined constants. It can be shown that updating the Q-table based on (3.24) maximizes the long-term expected reward: $\bar{r} = \lim\limits_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{r}(S(t))$ [94]. Moreover, given

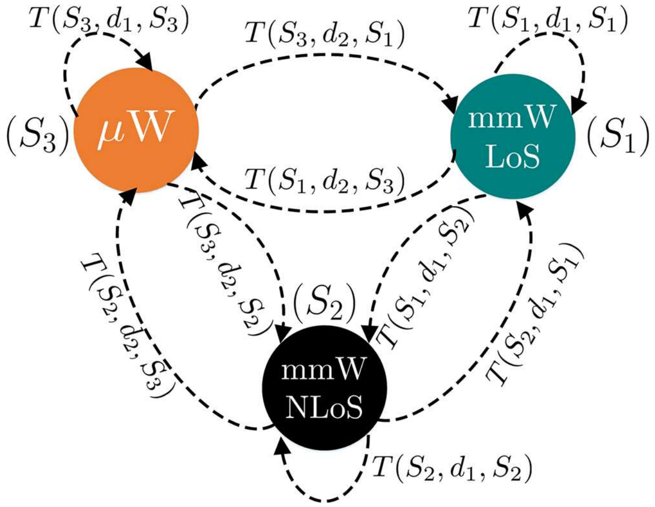


Figure 3.3: QL model with state transition probabilities.

the converged Q values, the following sufficient condition can be used to find a subset of UAs with $\rho_a \geq \rho_{th}$:

$$Q(S_i, d_2) \leq Q(S_i, d_1), i = 1, 2 \text{ and } Q(S_3, d_1) \leq Q(S_3, d_2) \Rightarrow \rho_a \geq \rho_{th}. \tag{3.25}$$

We note that if there is only one UE that is running only one UA, the criterion given by (3.24) leads to making optimal decisions in terms of maximizing the expected reward. However, the multi-user resource allocation cannot be done only based on $\rho_a \geq \rho_{th}$ criterion. On the one hand, assigning mmW resources only to UAs with high ρ_a and small required load will result low spectral efficiency. Moreover, UAs with small ρ_a and large $b_a^{\text{req}}(t)$ will not meet their delay requirement, if they are scheduled over the μW frequency band. However, even with small ρ_a, it is still probable for these UAs to be served over a LoS mmW link. Therefore, multi-user scheduling enforces SBS to exploit per UA context information, i.e., required load per UA, delay constraint, as well as UEs-SBS channel diversity. Here, it worth noting that exploiting side information such as the geographical location information of buildings could also facilitate learning the LoS probabilities [35, 96].

3.4.3 Complexity analysis of the proposed two-stage solution

With this in mind, we can make the following observation with regard to the proposed two-stage solution in Algorithms 1 and 2 for the original problem in (3.6a)-(3.6f).

Theorem 2. *The proposed long-term scheduling algorithm composed of Algorithm* 1 *and Algorithm* 2 *solves the problem* (3.6a)-(3.6f) *in polynomial time with respect to the number of UAs.*

Proof. See Appendix A.3. □

Table 3.2: Simulation parameters

Notation	Parameter	Value
P_1, P_2	Transmit power	30 dBm
(Ω_1, Ω_2)	Available Bandwidth	(10 MHz, 1 GHz)
ω	Bandwidth per RB	180 KHz
K-factor	Rician K-factor	2.4 [87]
(ξ_1, ξ_2)	Standard deviation of mmW path loss	$(10, 5.2)$ [10]
(α_1, α_2)	Path loss exponent	$(3, 2)$ [10]
(β_1, β_2)	Path loss at 1 m	$(38, 70)$ dB
ψ	Antenna gain	18 dBi
τ	Time slot duration	10 ms
τ'	Beam-training overhead	0.1 ms
N_0	Noise power spectral density	-174 dBm/Hz
J	Number of QoS classes	5 [97]
κ	Number of UAs per UE	3
$\boldsymbol{r}$	Reward vector	$[3, -16, 1]$

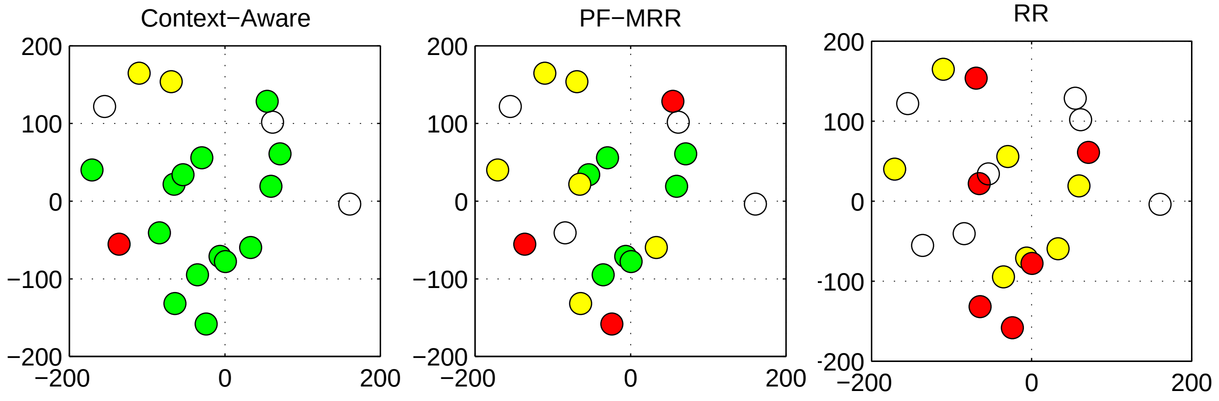
3.5 Simulation Results

For simulations, we consider an area with diameter $d = 200$ meters with the SBS located at the center [10]. UEs are distributed uniformly within this area with a minimum distance of 5 meters from the SBS. Each UE has κ UAs chosen randomly and uniformly from J QoS classes. The main parameters are summarized in Table 3.2. All statistical results are averaged over a large number of independent runs. We compare the performance of the proposed context-aware algorithm with two well-known resource allocation approaches:

- Proportional Fair Scheduler with minimum rate requirement (PF-MRR): The PF scheduling for multi-carrier systems with minimum rate requirement is different than the conventional approach. In [98], a simple approach is proposed to implement PF-MRR which we modify to apply to the dual-mode system. At μW band, RB k is assigned to the UA $a \in \mathcal{A}_t$ that satisfies

$$a = \underset{a\in\mathcal{A}_t}{\operatorname{argmax}} \frac{R_a(k,t)}{\bar{R}_a^{\text{rec}}(t) + R_a^{\text{req}}(t)}, \tag{3.26}$$

  where $\bar{R}_a^{\text{rec}}(t)$ is the achieved average rate up to time slot t, and $R_a^{\text{req}}(t) = b_a^{\text{req}}(t)/\tau$ is the required average rate at slot t to meet the QoS constraint of UA a. UAs $a \in \mathcal{A}_{t'\geq t}$ with unsatisfied rate requirement are scheduled at mmW band where $\tau_{a,t} = \frac{b_a^{\text{req}}(t)}{R_a(t)}$ is allocated to the UA $a = \operatorname{argmax}_a \frac{R_a(t)}{\bar{R}_a^{\text{rec}}(t)+R_a^{\text{req}}(t)}$, while $\sum_a \tau_{a,t} = \tau - |\mathcal{G}_{t,2}|\tau'$.
- Round Robin Scheduler (RR): At μW band, the scheduler allocates equal number of RBs to each $a \in \mathcal{A}_t$. Unsatisfied UAs $a \in \mathcal{A}_{t'\geq t}$ are scheduled at mmW band with $\tau_{a,t} = \frac{\tau-|\mathcal{G}_{t,2}|\tau'}{|\mathcal{G}_{t,2}|}$.



(a) Proposed approach (b) PF-MRR (c) RR

Figure 3.4: Performance comparison between scheduling approaches for $M = 20$ and $b_a = 5$ Mbits. For the cell edge UEs, ρ_a is sampled randomly from $[0, 1]$ and for others $\rho_a = 1$.

3.5.1 Quality-of-experience of the users

Fig. 3.4 shows a snapshot of a given network realization in which specific UEs are represented by circles. Each UE is associated with $\kappa = 3$ UAs, each having a required load of $b_a = 5$ Mbits. We note that for an arbitrary UA $a \in \mathcal{A}_j$, the required load b_a (bits) can be translated into data rate $b_a/(j\tau)$. For example, $b_a = 5$ Mbits for $a \in \mathcal{A}_5$ is equivalent to 100 Mbits/s data rate. The results from this figure show each user's satisfaction by indicating how many UAs per UE are satisfied. In Fig. 3.4, the colors red, yellow, and green are used, respectively, to indicate one, two, and three satisfied UAs per UE. Moreover, circles with no color represent UEs with no serviced UA. Clearly, in Fig. 3.4, we can see that the proposed approach significantly improves the overall system performance by providing service to more UEs, compared to both PF-MRR and RR schemes. In addition, we observe that the proposed context-aware approach outperforms PF-MRR and RR schemes by satisfying the QoS needs of more applications, which naturally leads to a higher quality-of-experience per user.

3.5.2 Outage probability vs number of UEs

The overall outage probability, P_{out}, is defined as the ratio of the number of QoS violations over the total number of UAs which will be given by:

$$
\begin{aligned}
P_{\text{out}}(\pi) &= 1 - \frac{1}{A}\left(\sum_{t=1}^{J}\lambda_{t,1}(\pi) - \sum_{t=1}^{J}\lambda_{t,2}(\pi)\right) \qquad (3.27)\\
&= 1 - \frac{1}{A}\sum_{t=1}^{J}\sum_{a\in\mathcal{A}_t}\mathbb{1}(a;\pi) = 1 - \frac{1}{A}|\mathcal{O}^{\pi}|,
\end{aligned}
$$

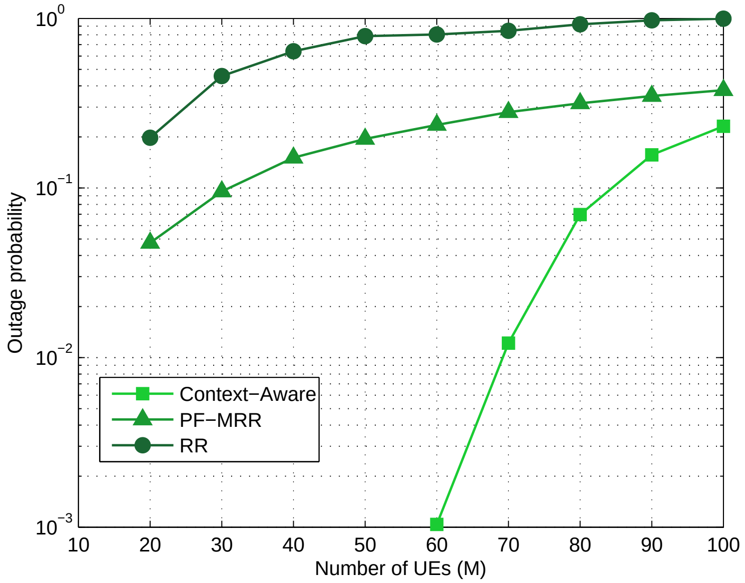


Figure 3.5: Performance comparison between scheduling approaches versus the number of UEs, M. All users are at LoS. $b_a = 1$ Mbits and $\rho_a = 1$ for all $a \in \mathcal{A}$.

Since P_{out} is a random variable, we will study whether the proposed scheduling policy $\pi^* \in \boldsymbol{\Pi}$ guarantees $\mathbb{P}(P_{\text{out}}(\pi^*) \geq P_{th}) \leq \epsilon$, where P_{th} is the maximum tolerable outage probability and ϵ is a pre-defined threshold. This can be written as $F_{P_{\text{out}}}(P_{th}) \geq 1 - \epsilon$, where $F_{P_{\text{out}}}(.)$ is the cumulative distribution function (CDF) of P_{out}.

Fig. 3.5 shows the outage probability as the number of UEs varies, for the three considered approaches. Fig. 3.5 shows that the outage probability increases as the number of UAs increases. In fact, the results show the number of UAs that can be satisfied for a given outage threshold. Clearly, the proposed algorithm outperforms the PF-MRR and RR scheduling approaches. For example, for a 0.01 outage probability, the proposed context-aware approach satisfies up to 210 UAs, considering $\kappa = 3$ UAs per UE. However, the baseline approaches fail to achieve this performance. In fact, the outage probability is always greater than 0.04 for both the RR and the PF-MMR approaches over all network sizes. Finally, from Fig. 3.5, we can clearly see that the proposed approach can always guarantee the QoS for up to 180 UAs on average, which is three times greater than the number of satisfied UAs resulting from the PF-MRR and RR approaches.

3.5.3 Impact of Q-learning

Fig. 3.6 shows the gain of the proposed QL approach. The QL gain is defined as the respective number of satisfied UAs with and without QL. The results presented in Fig. 3.6 show that more gain is achievable as the number of UAs increases. This stems from the fact that, as the number of UAs increases, it becomes more probable that more number of UEs be at a LoS connection with the BS. Fig. 3.6 shows that the QL-based information allows scheduling UAs with higher

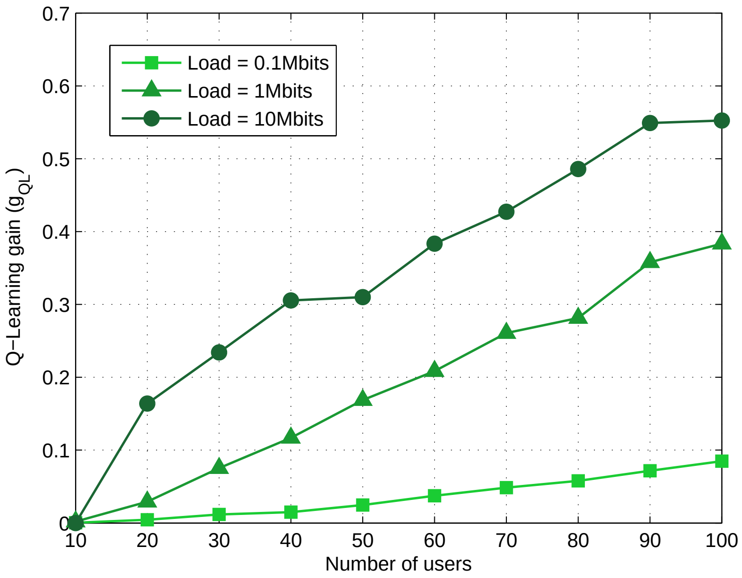


Figure 3.6: Gain of Q-Learning vs number of users for different load values.

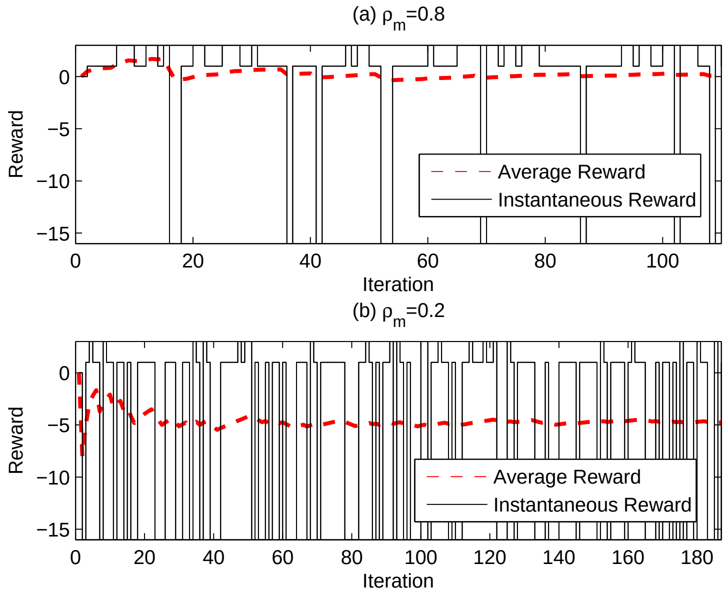


Figure 3.7: State exploration and convergence of Q-Learning

LoS probabilities. More interestingly, Fig. 3.6 shows that the gain increases as the required load per UA increases. This is due to the fact that with more strict QoS constraints, it is become more important to allocate mmW resources only to the UAs with higher probability of LoS.

Fig. 3.7.a and Fig. 3.7.b show both random state exploration and the resulting long-term rewards. The LoS probabilities $\rho_a = 0.8$ and $\rho_a = 0.2$ are considered, respectively, in Fig. 3.7.a and Fig. 3.7.b. The results in Fig. 3.7.a show that up to 110 iterations is needed for the QL algorithm to

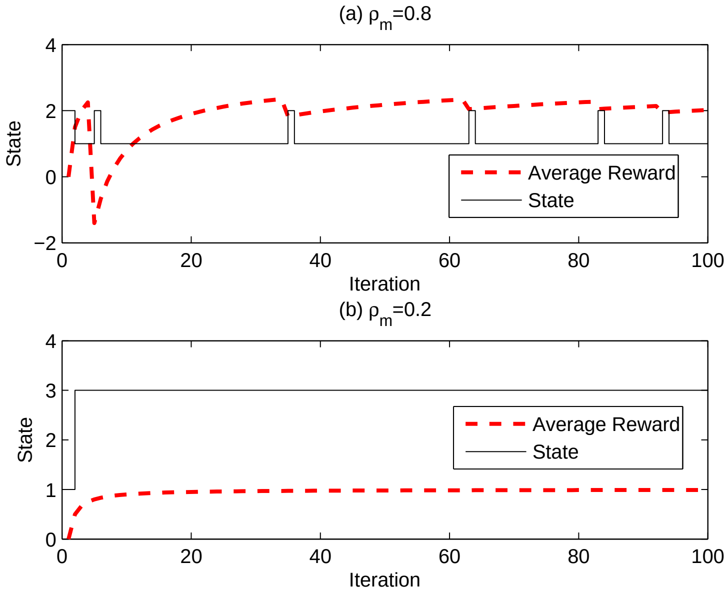


Figure 3.8: State transition and average reward resulted by the optimal policy.

converge, for $\rho_a = 0.8$. However, Fig. 3.7.b shows that the algorithm will converge within less than 190 iterations for $\rho_a = 0.2$. Moreover, the average reward is higher in Fig. 3.7.a, since the UAs with $\rho_a = 0.8$ are often served over mmW LoS links, while in Fig. 3.7.b, mmW links are frequently blocked. Real-life field measurements have shown that the blockage duration can be very long, exceeding several hundreds of milliseconds [99]. This long duration will allow the proposed QL algorithm to converge, before the blockage environment changes.

In Fig. 3.8.a and Fig. 3.8.b, the average reward and state transitions are shown when the optimal QL policy is followed, respectively, for $\rho_a = 0.8$ and $\rho_a = 0.2$. Clearly, when LoS probability is high, the optimal policy is to schedule the UA over mmW band, as shown in Fig. 3.8.a. In addition, compared to Fig. 3.7, we can see that the QL policy will substantially increase the average reward compared to the random frequency band selection. For example, for $\rho_a = 0.2$, the average reward is increased from -5 in Fig. 3.7.b to 1 in Fig. 3.8.b.

3.5.4 Outage probability vs the required load

In Fig. 3.9, we show the outage probability as the required load per UA varies, for the three scheduling approaches. In this figure, we can see that the outage probability decreases as the required load per UA decreases. In addition, from Fig. 3.9, we can see that the proposed context-aware approach yields significant gains, compared to the PF-MRR and RR schemes. In fact, the proposed approach guarantees the required loads up to 2 Mbits per UA, for 0.01 outage probability. However, the baseline PF-MRR and RR approaches can guarantee, respectively, less than 0.2 and 0.1 Mbits load per UA for the same outage probability.

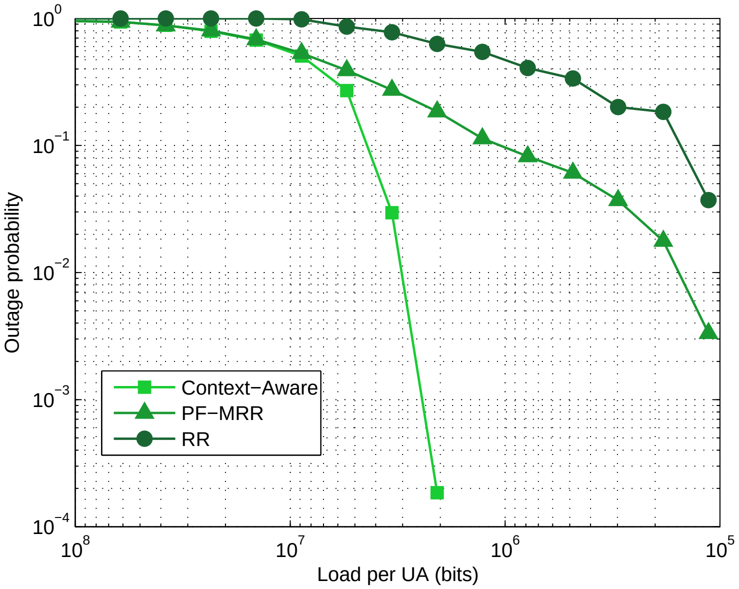


Figure 3.9: Performance comparison between scheduling approaches versus the required bit per UA b_a. Parameters $M = 30$ and $\rho_a = 1$ for all $a \in \mathcal{A}$ are used.

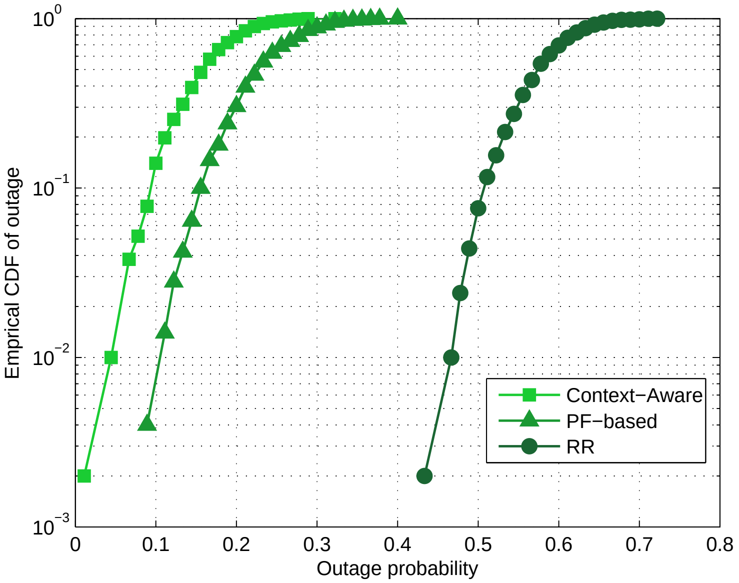
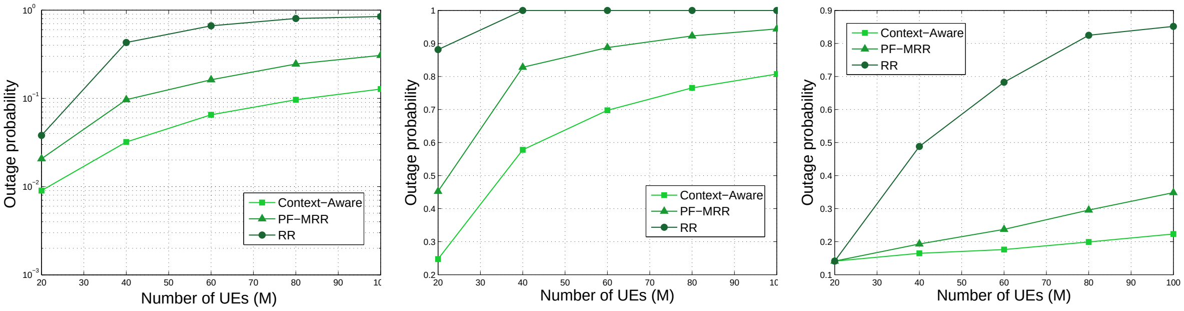


Figure 3.10: The CDF of the outage probability for $M = 30$ UAs and $b_a = 1$ Mbits. For the cell edge UEs, ρ_a is sampled randomly from $[0, 1]$ and for others $\rho_a = 1$.

3.5.5 Statistics of the outage probability

The empirical CDF of the outage probability is shown in Fig. 3.10 for $M = 30$ UAs with $b_a = 1$ Mbits required load. From Fig. 3.10, we can see that the proposed context-aware approach substantially improves the statistics of the outage, compared with PF-MRR and RR approaches.



(a) Dual-mode (b) Only μW (c) Only mmW

Figure 3.11: Performance comparison between scheduling approaches versus the number of UEs, M for $b_a = 0.1$ Mbits. ρ_a is sampled randomly from $[0, 1]$ for the half of UAs.

For example, the probability that P_{out} be less than 0.2 is only 30% for PF-MRR approach, while this value is 80% for the proposed approach.

3.5.6 Dual-mode vs single-mode scheduling

Fig. 3.11 shows the performance of the scheduling algorithms for three scenarios: a) with dual-mode communication in presence of both mmW and μW frequency resources, b) with only μW band being available, and c) with only mmW band being available[1]. The results in Fig. 3.11 show the key impact of the proposed dual-mode communication on maximizing QoS, compared with single-mode scenarios. In fact, Fig. 3.11b shows that, without mmW communications, the outage probability is significantly high across all network sizes. This is due to the fact that the requested traffic load by UEs falls beyond the available capacity of the network over μW band. Moreover, Fig. 3.11c shows that even for small network sizes, e.g. $M = 20$ UEs, the outage probability is greater than 10% which is significantly high for practical cellular networks. That is because the blockage is likely to happen for the subset of UAs with small ρ_a values. Therefore, to address high traffic loads on the one hand, and guarantee high QoS on the other hand, joint usage of mmW-μW resources is imperative. Indeed, Fig. 3.11a shows that the proposed dual-mode scheduling scheme will yield outage probabilities as low as 1%, while managing very large network sizes up to 300 UAs, with a reasonably small outage probability.

The average transmitted loads to UAs over mmW and μW frequency bands are shown, respectively, in Figs. 3.12a and 3.12b. In fact, Fig. 3.12 shows the average load per RAT at each time slot. We can observe that the transmitted traffic over the mmW RAT is significantly larger than the μW RAT. That is clearly due to the larger available bandwidth at the mmW band. Moreover, the transmitted load is lower at last time slots, since by that time, most of the UAs would have

[1] We note that, in our model, the μW mode does not employ advanced techniques, such as multi-antenna schemes (e.g., beamforming) or carrier aggregation to achieve higher data rates. Performance evaluation of such advanced μW systems (e.g. LTE-Advanced) can be considered in future work.

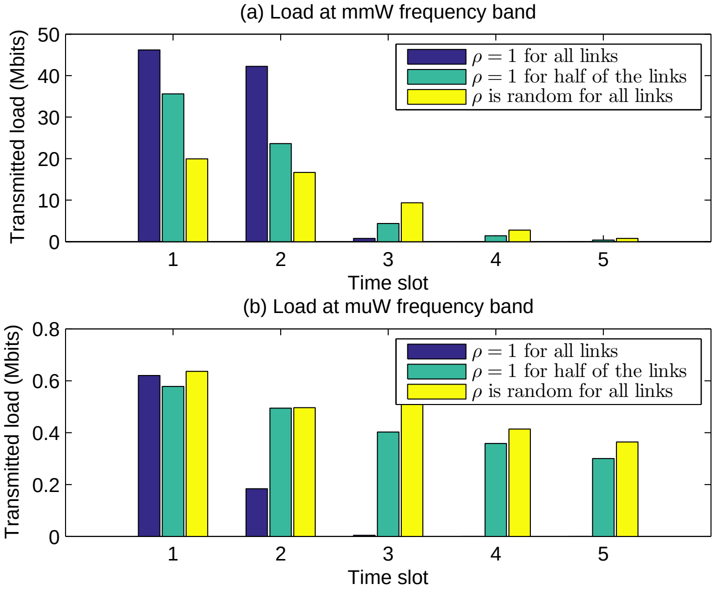


Figure 3.12: Average transmitted load to UAs at different time slots for $A = 90$ UAs and $b_a = 1$ Mbits.

already received their requested traffic. In fact, available bandwidth at the mmW band will allow to serve the LoS UAs prior to their due time slot. Clearly, as the link state becomes random for a higher number of UAs, more mmW links will be blocked and, thus, the traffic over the mmW band decreases. Given the results in Figs. 3.11 and 3.12, it is interesting to observe the critical role of exploiting μW resources, despite the significantly larger traffic at the mmW band. In fact, the joint exploitation of mmW-μW resources allows to leverage mmW resources for the UAs that are less likely to experience outage, which ultimately decreases traffic at the μW band in subsequent time slots.

3.5.7 Effect of beam training overhead

In Fig. 3.13, the effect of the beam training overhead on the outage probability is shown. Here, we observe that τ' will significantly affect the performance. From Fig. 3.13, we can clearly see that as τ' increases, the remaining time for data transmissions to UAs decreases which results in a higher outage probability. Fig. 3.13 shows that, in the absence of beam training overhead, the outage probability is always less than 0.35. However, for $\tau' = 0.8$ ms, the outage probability will always be less than 0.55.

3.5.8 Number of iterations

Fig. 3.14 shows number of iterations resulting from the proposed scheduling approach as the number of UEs varies for different number of UAs per UE. Clearly, the number of iterations increases almost linearly with the number of UEs. From this figure, we can see that even for large network size up to 30 UEs and 60 UAs, the proposed framework is relatively fast, as it converges within

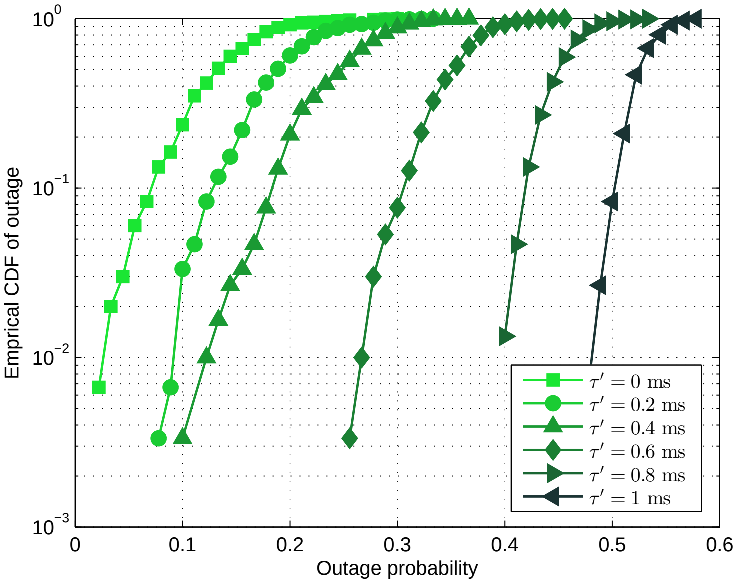


Figure 3.13: The CDF of the outage probability for $M = 30$ UAs and $b_a = 1$ Mbits. For the cell edge UEs, ρ_a is sampled randomly from $[0, 1]$ and for others $\rho_a = 1$.

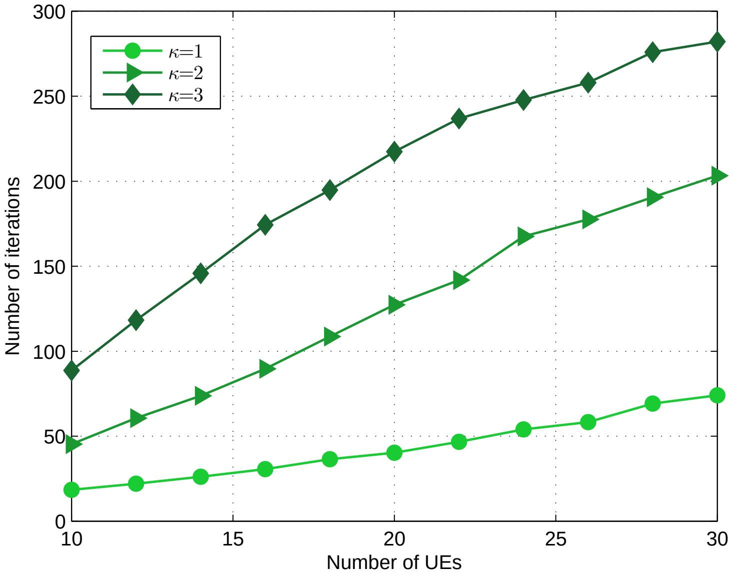


Figure 3.14: The number of iterations versus the number of UEs for $b_a = 0.1$ Mbits and $\kappa = 1, 2, 3$. For the cell edge UEs, ρ_a is sampled randomly from $[0, 1]$ and for others $\rho_a = 1$.

205 number of iterations.

3.6 Summary

In this chapter, we have proposed a novel context-aware scheduling framework for dual-mode small base stations operating at mmW and μW frequency bands. To this end, we have developed a two-stage UA selection and scheduling framework that takes into account various network and UA specific context information to make scheduling decisions. Over the μW band, we have formulated the context-aware scheduling problem as a one-to-many matching game. To solve this game, we have proposed a novel algorithm for joint UA selection and resource allocation and we have shown that it yields a two-sided stable matching between μW resources and UAs. Next, we have proposed a joint UA selection and scheduling to allocate mmW resources to the unscheduled UAs. The scheduling problem over mmW band is formulated as a 0-1 Knapsack problem and solved using a suitable algorithm. Moreover, we have proved that the proposed two-stage dual-mode scheduling framework can solve the problem in a polynomial time. Simulation results have shown the various merits and performance advantages of the proposed context-aware scheduling compared to the PF-MRR and RR approaches.

3.7 Appendix A

A.1 Proof of Theorem 1

The convergence of the Algorithm 1 at each slot is guaranteed, since a UA never applies for a certain RB twice. Hence, at the worst case scenario, all UAs will apply for all RBs once, which yields $\mathcal{K}_a = \emptyset, \forall a \in \mathcal{A}$. Next, we show that, once the algorithm converges, the resulting matching between UAs and RBs is two-sided stable. Assume that there exists a pair $(a, k) \notin \pi_{t,1}$ that blocks $\pi_{t,1}$. Since the algorithm has converged, we can conclude that at least one of the following cases is true about a: $\mathbb{1}_a(\boldsymbol{x}) = 1$, or $\mathcal{K}_a = \emptyset$.

The first case, $\mathbb{1}_a(\boldsymbol{x}) = 1$ implies that a does not need to add more RBs to $\pi_{t,1}(a)$. In addition, a would not replace any of $k' \in \pi_t(a)$ with k, since $k' \succ_a k$. Otherwise, a would apply earlier for k. If a has applied for k and got rejected, this means $\pi_{t,1}(k) \succ_k a$, which contradicts (a, k) to be a blocking pair. Analogous to the first case, $\mathcal{K}_a = \emptyset$ implies that a has got rejected by k, which means $\pi_{t,1}(k) \succ_k a$ and (a, k) cannot be a blocking pair. This proves the theorem.

A.2 Proof of Proposition 1

We can write $\lambda_{t,2}$ as the sum of Bernoulli random variables ζ_{at}, i.e., $\lambda_{t,2}(\boldsymbol{\tau}) = \sum_{a\in\mathcal{G}_{t,2}} \zeta_{at}$. Hence, using Le Cam's theorem, the distribution of $\lambda_{t,2}$ follows Poisson distribution, i.e.,

$$\mathbb{P}\left[\lambda_{t,2}(\boldsymbol{\tau}_t) = k\right] \approx \frac{\lambda_{ave}^k e^{-\lambda_{ave}}}{k!}, \tag{3.28}$$

where $\mathbb{E}\left[\lambda_{t,2}(\boldsymbol{\tau}_t)\right]$ is the sum of expected values of ζ_{at} for selected UAs in $\mathcal{G}_{t,2}$ as given by (3.17). Moreover, the approximation error is bounded by

$$\sum_{k=0}^{\infty}\left|\mathbb{P}\left[\lambda_t = k\right] - \frac{\lambda_{ave}^{k}e^{-\lambda_{ave}}}{k!}\right| < 2\sum_{a\in\mathcal{G}_{t,2}}\rho_a^2, \tag{3.29}$$

where $\lambda_t = \lambda_{t,1} + \lambda_{t,2}$. Next,

$$F_{P_{out,t}}(P_{th}) = \mathbb{P}\left(P_{out,t} \leq P_{th}\right) = \mathbb{P}\left(1 - \frac{\lambda_{t,1}+\lambda_{t,2}}{A_t} \leq P_{th}\right) \tag{3.30}$$

$$= 1 - \mathbb{P}\left(\lambda_{t,2} \leq \lfloor(1-P_{th})A_t - \lambda_{t,1}\rfloor\right) \tag{3.31}$$

$$\approx 1 - \frac{\mathit{\Gamma}\left(\lfloor(1-P_{th})A_t - \lambda_{t,1} + 1\rfloor, \lambda_{ave}\right)}{\lfloor(1-P_{th})A_t - \lambda_{t,1}\rfloor!}, \tag{3.32}$$

where (3.32) follows the CDF of the Poisson distribution.

A.3 Proof of Theorem 2

First, we analyze the complexity of Algorithm 1. For each slot t, let $A_{j\geq t} = |\mathcal{A}_{j\geq t}|$ be the number of UAs that can be selected to be scheduled at μW band. At most, the algorithm must find the solution for $A_{j\geq t}$ number of matchings. In addition, each matching has the complexity of $O(K_1)$, since in the worst case, each UA must be re-allocated to K_1 RBs by SBS. Hence, the complexity of Algorithm 1 at each slot t is $O(K_1 A_{j\geq t})$ and the total complexity from slot $t=1$ to $t=J$ is $O(K_1\sum_{j=1}^{J} jA_j)$.

Next, we analyze the complexity of Algorithm 2. At each slot t, there are at most $A_{j\geq t}$ UAs to be scheduled at mmW ($\mathcal{G}_{t,1} = \emptyset$). Therefore, the Algorithm 2 must converge after $A_{j\geq t}$ resource allocations, where each allocation is a special case of the 0-1 Knapsack problem. Hence, the total complexity of the Algorithm 2 from slot $t=1$ to slot $t=J$ is $O(\sum_{j=1}^{J} jA_j)$. From the above results, the overall complexity of the proposed long-term scheduling is $O\left((K_1+1)\sum_{j=1}^{J} jA_j\right)$.

Chapter 4

Inter-Operator Resource Management for Millimeter Wave, Multi-Hop Backhaul Networks

4.1 Background, Related Works, and Summary of Contributions

Network densification based on the concept of SCNs is seen as the most promising solution to cope with the increasing demand for wireless capacity [100]. SCNs are built on the premise of a viral and dense deployment of SBSs over large geographical areas so as to reduce the coverage holes and improve the spectral efficiency [101]. However, such a large-scale deployment of SBSs faces many challenges in terms of resource management, network modeling, and backhaul support [101].

In particular, providing backhaul support for a large number of SBSs that can be deployed at adverse locations within a geographical area has emerged as one of the key challenges facing the effective operation of future heterogeneous SCNs [11]. In particular, due to the density of SCNs, MNOs will not be able to maintain an expensive and costly deployment of fiber backhauls to service SBSs as shown in [9] and [11]. Instead, MNOs are moving towards the adoption of wireless backhaul solutions that are viewed as an economically viable approach to perform backhauling in dense SCNs. In fact, MNOs expect that 80% of SBSs will connect to the core network via wireless backhaul as detailed in [11] and [12].

4.1.1 Related works

The authors in [102] propose a fair resource allocation for the out-band relay backhaul links. The proposed approach developed in [102] aims to maximize the throughput fairness among backhaul and access links in LTE-Advanced relay system. In [103], a backhaul resource allocation approach is proposed for LTE-Advanced in-band relaying. This approach optimizes resource partitioning between relays and macro users, taking into account both backhaul and access links quality. Dynamic backhaul resource provisioning is another important problem in order to avoid outage in peak traffic hours and under-utilizing frequency resources in low traffic scenarios. In this regard, in [104], a dynamic backhaul resource allocation approach is developed based on evolutionary game theory. Instead of static backhaul resource allocation, the authors take into account the dynamics of users' traffic demand and allocate sufficient resources to the base stations, accordingly. Although interesting, the body of work in [102–104] does not consider the potential deployment of millimeter wave communication at the backhaul network and is primarily focused on modeling rather than resource management and multi-hop backhaul communication.

Providing wireless backhaul links for SBSs over mmW frequencies has recently been dubbed as one of the most attractive technologies for sustaining the backhaul traffic of SCNs [7–15], due to the following promising characteristics, among others: 1) The mmW spectral band that lies within the range 30-300 GHz will deliver high-capacity backhaul links by leveraging up to 10 GHz of available bandwidth which is significantly larger than any ultra-wideband system over sub-6 GHz frequency band. In addition, high beamforming gains are expected from mmW antenna arrays, with large number of elements, to overcome path loss [13], 2) more importantly, mmW backhaul links will not interfere with legacy sub-6 GHz communications in either backhaul or access links, due to operating at a different frequency band. Even if the access network operates over the mmW frequency band such as in self-backhauling architectures, mmW communications will generally remain less prone to interference, due to the directional transmissions, short-range links, as well as susceptibility to the blockage [96], and 3) over the past few years, research for utilizing mmW frequencies for wireless backhaul networks has become an interesting field that attracted a lot of attention in both academia and industry [7–16]. As an example, in 2014, a total of 15 telecom operators, vendors, research centers, and academic institutions (including Nokia, Intel, and operators Orange and Telecom Italia) have launched a collaborative project in Europe, called *MiWaveS*, to develop mmW communications for 5G backhaul and access networks [16].

However, compared to existing ultra-dense networks over sub-6 GHz band, the major *challenges* of mmW backhaul networks can be listed as follows: 1) MmW backhaul links will typically operate over much shorter range than their sub-6 Ghz counterparts (usually do not exceed 300 meters [28, 105]), and, thus, more SBSs will be required to provide backhaul support for the users within a certain geographical area. Therefore, mmW SBS deployments are expected to be even denser, compared to the already dense sub-6 GHz networks [106]. Such ultra dense network will require fast and efficient network formation algorithms to establish a multi-hop backhaul link between the core network and each demanding SBS, 2) the backhaul network must be significantly reliable. However, the received signal power of mmW signals may significantly degrade if the backhaul link is blocked by an obstacle. For SBSs that are deployed in adverse locations, such as urban furniture, the received signal power may degrade due to rain or blockage by large vehicles. One solution is to increase the density of SBSs such that if a backhaul link between two SBSs is blocked, the demanding SBS can establish a reliable link with another SBS. However, this solution will increase the cost of the backhaul network for the MNO. In our work, we have motivated the use of cooperation between MNOs to achieve a robust and economically efficient backhaul solution, and 3) due to the directional transmissions of the mmW signals, broadcast control channels can lead to a mismatch between the control and data planes at mmW frequency bands [107]. Therefore, fully centralized approaches that rely on receiving control signals from a central station over broadcast channels may not be practical, thus, motivating the adoption of suitable distributed algorithms for an effective resource management.

Several recent studies have studied the viability of mmW as a backhaul solution as presented in [14] and [105, 108–112]. For instance, the work in [14] proposes a model based on stochastic geometry to analyze the performance of the self-backhauled mmW networks. The work in [109] analyzes the performance of a dual-hop backhaul network for mmW small cells. In [105], the authors perform channel measurements and provide insights for the mmW small cell backhaul links. In [110], the performance of adaptive and switching beamforming techniques are investigated and evaluated for mmW backhaul networks. Moreover, the impact of diffraction loss in mmW backhaul network is analyzed in [111]. The authors in [112] propose a multi-objective optimization framework for joint deployment of small cell base stations and wireless backhaul links. In [108], the authors propose an autonomous beam alignment technique for self-organizing multi-hop mmW backhaul networks. In [106], the authors have motivated the use of a multi-hop mmW backhaul as a viable solution for emerging 5G networks and they analyzed the impact of the deployment density on the backhaul network capacity and power efficiency. Moreover, in [113], the authors have proposed a multi-hop backhaul solution with a TDMA MAC protocol for WiMAX.

The body of work in [14] and [105, 108–112] solely focuses on physical layer metrics, such as links' capacity and coverage. In addition, it is focused only on single-hop or dual-hop backhaul networks, while new standards such as IEEE 802.11ay envision fully multi-hop networks. The work presented in [106] does not provide any algorithm to determine how SBSs must form a multi-hop mmW backhaul network. Moreover, the proposed model in [106] is too generic and does not capture specific characteristics of a mmW network, such as susceptibility to blockage and directional transmissions. Last but not the least, no specific analysis or algorithm is provided for resource management in multi-hop mmW backhaul networks. The solution presented in [113] is not directly applicable to the mmW backhaul networks, as mmW is substantially different from WiMAX systems. In fact, authors in [113] focus primarily on the routing and link activation protocols in order to minimize the interference among active links. Such a conservative approach will yield an inefficient utilization of the mmW frequency resources, since interference scenario in WiMAX systems is completely different with directional mmW communications.

Furthermore, the body of work in [14], [105, 108–112], [106], and [113] does not account for the effect of backhaul cost in modeling backhaul networks. In fact, these existing works typically assume that all infrastructure belong to the same MNO which may not be practical for dense SCNs. In wireless networks, the backhaul cost constitutes a substantial portion of the total cost of ownership (TCO) for MNOs as indicated in [9] and [11]. In fact, it is economically inefficient for an individual MNO to afford the entire TCO of an independent backhaul network as demonstrated in [9], [11], and [114]. The main advantages of inter-operator backhaul sharing is to reduce the number of required sites/RAT interfaces per MNO to manage backhaul traffic, site rent, capital expenditures (CAPEX) by avoiding duplicate infrastructure, site operating expenditures (OPEX), and electricity costs [115]. Moreover, inter-operator mmW backhaul architectures are more robust against the blockage and link quality degradation compared to the schemes in which operators act independently and non-cooperatively [7]. This stems from the fact that cooperation increases flexibility to establish new backhaul links that can easily bypass obstacles. Therefore, MNOs will need to share their backhaul network resources with other MNOs that require backhaul support for their SBSs [114]. Hence, beyond the technical challenges of backhaul management in SCNs, one must also account for the cost of sharing backhaul resources between MNOs.

To address such economic challenges, a number of recent works have emerged in [15] and [114, 116–118]. The work in [114] motivates a business model for an SCN where multiple MNOs share the SBSs that are deployed on the street lights of dense urban areas. In [116] an economic framework is developed to lease the frequency resources to different MNOs by using novel pricing mechanisms. In [117], the authors propose a cost evaluation model for small cell backhaul networks. This work highlights the fact that integrating heterogenous backhaul technologies is mandatory to achieve a satisfactory performance in a backhaul network. Moreover, they show that the TCO of an SCN is much higher than conventional cellular networks. Therefore, it is more critical to consider backhauling cost in small cell backhaul network design. The authors in [118] propose a model where MNOs buy energy from the renewable power suppliers for their mmW backhaul network and solve the problem as a Stackelberg game between MNOs and power suppliers. In [15], we studied the problem of resource management for the mmW-microwave backhaul networks with multiple MNOs. The approach in [15] considers both cost and the CSI to allocate backhaul resources to the SBSs. The provided solutions in [114, 116–118] focus solely on the economic aspects of the backhaul network, while a suitable backhaul network model must integrate the cost constraints with the physical constraints of the wireless network. In addition, [15] does not consider multi-hop backhaul networks. Moreover, the backhaul model studied in [15] is restricted to the case in which only two MNOs are in the network.

4.1.2 Summary of contributions

The main contribution of this chapter is to propose a novel framework to model and analyze resource management and pricing for facilitating inter-operator sharing of multi-hop, mmW backhaul infrastructure in dense SCNs. In particular, the proposed framework is formulated using suitable techniques from matching theory [90] so as to provide a distributed solution for managing the resources over multi-hop backhaul links. In the formulated model, the SBSs of one MNO can act as *anchored BSs (A-BSs)* to provide backhaul support to other, *demanding BSs (D-BSs)* that may belong to other MNOs. The proposed framework is composed of two highly-interrelated matching games: a network formation game and a resource management game. The goal of the network formation game is to associate the D-BSs to A-BSs for every hop of the backhaul links. This game is shown to exhibit peer effects thus mandating a new algorithmic approach that differs from classical matching works in [90] and [76]. To solve this game, we propose a distributed algorithm that

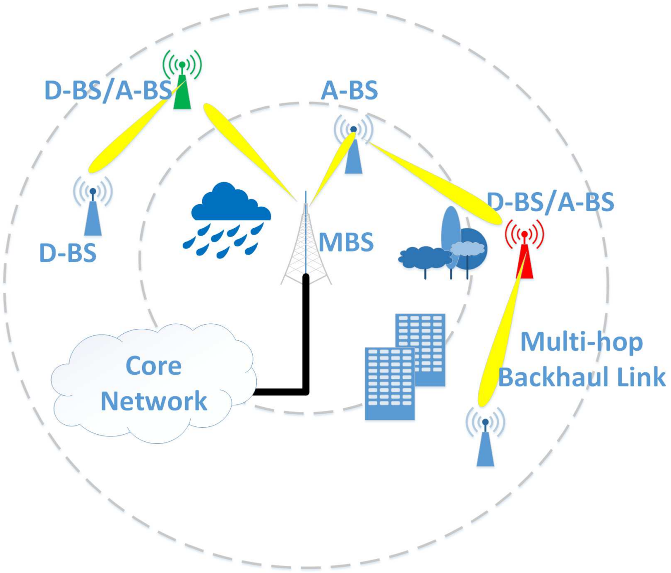


Figure 4.1: An example of mmW-MBN with multiple MNOs. SBSs with the same color belong to the same MNO.

is guaranteed to converge to a two-sided stable and Pareto optimal matching between the A-BSs and the D-BSs. Once the stable and optimal network formation solution is found, we propose a second matching game for resource management that allocates the sub-channels of each A-BS to its associated D-BSs, determined by the first matching game. The proposed approach considers the cost of the backhaul jointly with the links' achievable rates to allocate the sub-channels to the D-BSs. To solve this resource management matching game with peer effects, we propose a novel distributed algorithm that yields a two-sided stable and Pareto optimal matching between the sub-channels and the D-BSs. We compare the performance of the proposed *cooperative mmW multi-hop backhaul network* (mmW-MBN) and compare the results with non-cooperative mmW-MBN. Simulation results show that MNOs cooperation provides significant gains in terms of network's average backhaul sum rate, reaching up to 30%, compared to the non-cooperative mmW-MBN. The results also show that the cooperation among MNOs will significantly improve the statistics of the backhaul rate per SBS.

The rest of this chapter is organized as follows. Section 4.2 describes the system model and formulates the problem. Section 4.3 presents our distributed approach to solve the network formation problem. Section 4.4 provides the proposed solution to solve the resource allocation problem. Section 4.5 provides the simulation results and Section 4.6 concludes the chapter.

4.2 System Model

Consider a mmW-MBN that is used to support the downlink transmissions of M SBSs within the set $\mathcal{M}$. Each SBS belong to one of N MNOs within the set $\mathcal{N}$. The set $\mathcal{M}$ can be decomposed into N subsets $\mathcal{M}_n$, with $\bigcup_{n\in\mathcal{N}}\mathcal{M}_n = \mathcal{M}$ and $\bigcap_{n\in\mathcal{N}}\mathcal{M}_n = \emptyset$, where $\mathcal{M}_n$ represents the subset of SBSs belonging to MNO n. The SBSs are distributed uniformly in a planar area with radius d_{max} around an MBS, m_0, located at $(0,0) \in \mathbb{R}^2$. The MBS is connected to the core network over a broadband fiber link, as shown in Fig. 4.1, and is shared by all MNOs. The SBSs can be connected to the MBS via a *single-hop or a multi-hop mmW* link. The mmW-MBN can be represented as a directed graph $G(\mathcal{M},\mathcal{E})$, in which the SBSs are the vertices and $\mathcal{E}$ is the set of edges. Each edge, $e(m',m) \in \mathcal{E}$, represents a mmW backhaul link from SBS m' to m. Hereinafter, for any link, the transmitting and the receiving SBSs (over the backhaul) will be referred to, respectively, as the A-BSs and the D-BSs.

Thus, in our model, an SBS can be either a D-BS or an A-BS. Each A-BS m will serve up to Q_m D-BSs, while each D-BSs will be connected to one A-BS.

To show that an arbitrary D-BS m is connected to an A-BS m', we use the following binary variable

$$\epsilon_e(m',m) = \begin{cases} 1 & \text{if } e(m',m) \in \mathcal{E}, \\ 0 & \text{otherwise}, \end{cases} \tag{4.1}$$

where $\epsilon_e(m',m) = 0$ implies that no backhaul link exists from SBS m' to m. Finally, we denote by $\mathcal{M}_{m'}^{\text{D-BS}}$ the subset of SBSs for whom SBS m' serves as an A-BS. In other words, $\mathcal{M}_{m'}^{\text{D-BS}} = \{m \in \mathcal{M} |\, \epsilon_e(m',m) = 1\}$. The backhaul links are carried out over a mmW frequency band, composed of K sub-channels, within the set $\mathcal{K}$, each of a bandwidth w. A summary of our notation is provided in Table 4.1.

4.2.1 Channel model

The state of a backhaul link is defined as a Bernoulli random variable $\zeta_{m'm}$ with success probability $\rho_{m'm}$ to determine if the link is LoS or NLoS. In fact, $\zeta_{m'm} = 1$, if $e(m',m)$ is LoS, otherwise, $\zeta_{m'm} = 0$. Based on the field measurements carried out in [10] and [119–121], the large-scale path loss of the link $e(m',m)$, denoted by $L_{\text{dB}}\,(m',m)$ in dB, is given by

$$\begin{aligned} L_{\text{dB}}(m',m) &= 10\log_{10}(l(m',m)), \\ &= 20\log_{10}\left(\frac{4\pi d_0}{\lambda}\right) + 10\alpha\log_{10}\left(\frac{\|\boldsymbol{y}_m - \boldsymbol{y}_{m'}\|}{d_0}\right) + \chi, \quad d \geq d_0, \end{aligned} \tag{4.2}$$

where λ is the wavelength at carrier frequency $f_c = 73$ GHz, d_0 is the reference distance, and α is the path loss exponent. Moreover, $\|\boldsymbol{y}_m - \boldsymbol{y}_{m'}\|$ is the Euclidean distance between SBSs m and m', located, respectively, at $\boldsymbol{y}_m \in \mathbb{R}^2$ and $\boldsymbol{y}_{m'} \in \mathbb{R}^2$. In addition, χ is a Gaussian random variable with zero mean and variance ξ^2. Path loss parameters α and ξ will naturally have different values, depending on the state of the link. In fact, depending on whether the link is LoS or NLoS, these values can be chosen such that the path loss model in (4.2) will provide the best linear fit with the field measurements carried out in [10]. The benefit of the free space path loss model used

Table 4.1: Variables and notations

Notation	**Description**	**Notation**	**Description**
M	Number of SBSs	$\mathcal{M}$	Set of SBSs
N	Number of MNOs	$\mathcal{N}$	Set of MNOs
K	Number of sub-channels	$\mathcal{K}$	Set of sub-channels
$e(m',m)$	Backhaul link from m' to m	$\mathcal{E}$	Set of backhaul links
$\mathcal{M}_n$	Set of SBSs belonging to MNO n	$\mathcal{M}_m^d$	Set of SBSs of distance d from m
$\mathcal{M}_{m'}^{\text{D-BS}}$	SBSs for whom SBS m' serves as an A-BS	w	Bandwidth of each sub-channel
$\zeta_e \in \{0,1\}$	State of link e	ρ_e	Expected value of ζ_e
$r_m(k,m';\zeta)$	Rate for D-BS m over sub-channel k	$r_m(m',\boldsymbol{x})$	Rate of D-BS m from A-BS m', given $\boldsymbol{x}$
$\succ_m^D$	Preference profile of D-BSs over A-BSs	$\succ_m^A$	Preference profile of A-BSs over D-BSs
P_m^D	Preference profile of D-BSs over sub-channels	P_k^K	Preference profile of sub-channels over D-BSs
π_j	Network formation matching for j-th hop	μ_j	Resource allocation matching for j-th hop
Q_m	Quota of A-BS m	$r_{m,\text{th}}$	Backhaul minimum rate requirement for m
$\mathbb{1}_{mm'}$	Indicates if m and m' belong to same MNO	$\bar{r}_m(m')$	Average rate for m over all sub-channels

in (4.2), compared with other models such as the alpha-plus-beta model, is that it is valid for all distances above the reference distance d_0 and the model parameters α and χ have concrete physical interpretations.

In addition, the field measurements in [122–124] show that the mmW channel delay spread can be large, reaching up to more than 100 ns, for the outdoor deployment of mmW SBSs in urban areas. To this end, for any link $e(m',m)$, a slow-varying frequency flat fading channel $h_{m'km}$ is considered over sub-channel k. Hence, conditioned to the link state ζ, the achievable rate for a given link $e(m',m)$ over sub-channel k will be given by

$$r_m(k,m';\zeta) = w\log_2\left(1+\frac{p_{m',k}\psi(m',m)l\,(m',m)\,|h_{m'km}|^2}{\sum_{m''\neq m,m'} p_{m'',k}\psi(m'',m)l\,(m'',m)\,|h_{m''km}|^2+\sigma^2}\right), \tag{4.3}$$

where $p_{m',k}$ and σ^2 denote, respectively, the transmit power of A-BS m' over sub-channel k and the noise power. To strike a balance between system performance and complexity, uniform power allocation is assumed. Here, we assume that total transmit power $p_{t,m'}$ is distributed uniformly over all sub-channels, such that $p_{m',k} = p_{t,m'}/K$ [84, 125–127]. The uniform power allocation assumption is also due to the fact that at a high SNR/SINR regime, as is expected in a mmW network with relatively short-range links and directional transmissions, it is well known that optimal power allocation policies such as the popular water-filling algorithm will ultimately converge to the uniform power allocation [84]. Moreover, $\psi(m',m)$ represents the combined transmit and receive antenna gains. The antenna gain pattern for each BS is assumed to be sectorized and is given by [14]:

$$G(\theta) = \begin{cases} G_{\text{max}}, & \text{if} \quad \theta < |\theta_m|, \\ G_{\text{min}}, & \text{otherwise}, \end{cases} \tag{4.4}$$

where θ and θ_m denote, respectively, the azimuth angle and the antennas' main lobe beamwidth. Moreover, G_{max} and G_{min} denote, respectively, the antenna gain of the main lobe and side lobes. It is assumed that for a desired link between A-BS m' and D-BS m, $\psi(m',m) = G_{\text{max}}^2$. Moreover, $\psi(m'',m)$ of an interference link from A-BS m'' to the target D-BS m is assumed to be random. Using (4.3), we can write the achievable rate for the link $e(m',m)$ over the allocated sub-channels as follows:

$$r_m(m';\boldsymbol{x}) = \sum_{k\in\mathcal{K}} r_m(k,m';\zeta)x_{m'km}, \tag{4.5}$$

where $\boldsymbol{x}$ is the resource allocation vector with elements $x_{m'km} = 1$, if SBS m' transmits to m over sub-channel k, otherwise, $x_{m'km} = 0$. In (4.5), we remove the dependency on ζ in the left-hand side to simplify the notations. Here, considering a decode-and-forward scheme, we note that, if an SBS m is connected to the MBS via a multi-hop link of length n, then $r_m(m';\boldsymbol{x})$ will be limited by $1/n$ times the minimum (bottleneck) of all link rates over the multi-hop connection [128]. In addition, by averaging with respect to ζ, the average achievable rate over all sub-channels for D-BS m assigned to A-BS m' will be

$$\begin{aligned}
\bar{r}_m(m') &= \mathbb{E}\left[\sum_{k\in\mathcal{K}} r_m(m',k;\zeta)\right], && (4.6)\\
&= \mathbb{P}(\zeta_{m',m}=1)\sum_{k\in\mathcal{K}} r_m(m',k;\zeta)x_{m'km} + \mathbb{P}(\zeta_{m',m}=0)\sum_{k\in\mathcal{K}} r_m(m',k;\zeta)x_{m'km}, && (4.7)\\
&= \rho_{m'm}\sum_{k\in\mathcal{K}} r_m(m',k;\zeta=1)x_{m'km} + (1-\rho_{m'm})\sum_{k\in\mathcal{K}} r_m(m',k;\zeta=0)x_{m'km}. && (4.8)
\end{aligned}$$

For dense urban areas, the number of obstacles blocking an arbitrary link $e(m',m)$ increases as $\|\boldsymbol{y}_m - \boldsymbol{y}_{m'}\|$ increases. Such severe shadowing will significantly reduce the received signal power, particularly, for street-level deployment of mmW SBSs over urban furniture such as lamp posts. Therefore, the communication range of each SBS will be limited to a certain distance d, where d depends on the density of the obstacles, as suggested in [105] and [28]. To this end, we define $\mathcal{M}_m^d$ as

$$\mathcal{M}_m^d = \{m' \in \mathcal{M}, m' \neq m \big| \; \|\boldsymbol{y}_m - \boldsymbol{y}_{m'}\| \leq d\}, \tag{4.9}$$

which effectively represents the set of SBSs with which m is able to communicate over an LoS or an NLoS link.

4.2.2 Network formation and resource allocation in mmW-MBN with multiple MNOs

We consider a cooperative, inter-operator mmW-MBN in which, under proper pricing incentives, the SBSs of each MNO may act as A-BSs for other SBSs belonging to other MNOs. We let q_m be a unit of price per sub-channel of SBS $m \in \mathcal{M}_n$, as determined by MNO n. That is, if $x_{mkm'} = 1$ for $m \in \mathcal{M}_n$ and $m' \in \mathcal{M}_{n'}$, where $n \neq n'$, MNO n' will have to pay q_m to MNO n. To solve the resource management problem for the proposed mmW-MBN, we need to first determine the backhaul links, $\epsilon_e(m',m)$, and then specify the rate over each link, $r_m(m',\boldsymbol{x})$. To this end, as

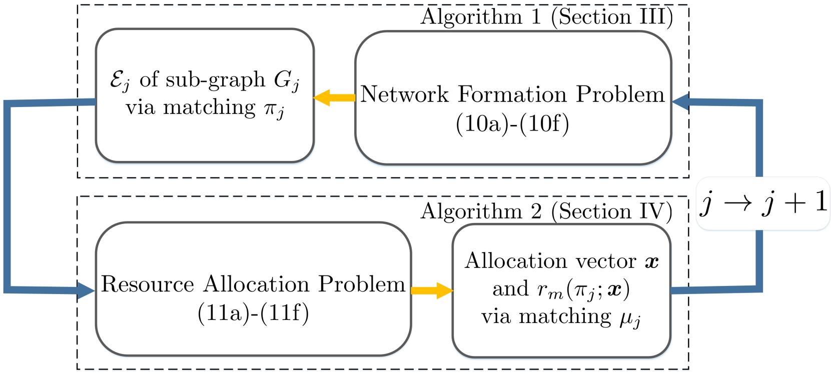

Algorithm 1 (Section III)
$\mathcal{E}_j$ of sub-graph G_j via matching π_j
Network Formation Problem (10a)-(10f)
$j \rightarrow j+1$
Algorithm 2 (Section IV)
Resource Allocation Problem (11a)-(11f)
Allocation vector $\boldsymbol{x}$ and $r_m(\pi_j;\boldsymbol{x})$ via matching μ_j

Figure 4.2: Proposed multi-stage framework for joint backhaul network formation and resource allocation.

illustrated in Fig. 4.2, we must solve two interrelated problems: 1) *network formation* problem that determines $\mathcal{E}$, and 2) *resource allocation* problem to assign sub-channels of each A-BS m to their corresponding D-BSs in $\mathcal{M}_m^{\text{D-BS}}$.

The network formation problem can be formulated as follows:

$$\begin{aligned}
&\underset{\mathcal{E}}{\operatorname{argmax}} \sum_{m\in\mathcal{M}} \sum_{m'\in\mathcal{M}_m^d} \epsilon_e(m',m)\bar{r}_m(m') - \kappa_m q_{m'}^t \mathbb{1}_{mm'}, && && (4.10a)\\
&\text{s.t.} \quad \epsilon_e(m',m)+\epsilon_e(m,m'')+\epsilon_e(m'',m') \le 2, && \forall m,m',m'' \in \mathcal{M}, && (4.10b)\\
&\sum_{m'\in\mathcal{M}_m^d} \epsilon_e(m',m) \le 1, && \forall m \in \mathcal{M}, && (4.10c)\\
&\sum_{m\in\mathcal{M}_{m'}^d} \epsilon_e(m',m) \le Q_{m'}, && \forall m' \in \mathcal{M}, && (4.10d)\\
&\epsilon_e(m',m)+\epsilon_e(m,m') \le 1, && \forall m,m' \in \mathcal{M}, && (4.10e)\\
&\epsilon_e(m',m) \in \{0,1\}, && \forall m,m' \in \mathcal{M}, && (4.10f)
\end{aligned}$$

where $\mathbb{1}_{mm'} = 1$, if both SBSs m and m' belong to different MNOs, otherwise, $\mathbb{1}_{mm'} = 0$. In addition, κ_m is a weighting scalar that scales the cost of a link with respect to its rate. The total cost of a link $e(m',m)$ for m is $q_{m'}^t = q_{m'}\sum_{k\in\mathcal{K}} x_{m'km}$. Constraint (4.10b) is to avoid any cycles. In addition, (4.10c) indicates that each D-BS must be assigned to at most one A-BS. Moreover, (4.10d) indicates that each A-BS m' can be assigned to up to $Q_{m'}$ D-BSs. Constraint (4.10e) ensures that all links are directional. That is, an SBS m may transmit to m' or receive its traffic from m', however, cannot do both simultaneously.

The solution of problem (4.10a)-(4.10f) yields $\mathcal{E}$ for the mmW-MBN graph $G(\mathcal{M},\mathcal{E})$ which also determines $\mathcal{M}_{m'}^{\text{D-BS}}$ for all $m' \in \mathcal{M}$. Next, the sub-channels of each A-BS m' must be allocated to its assigned D-BSs in $\mathcal{M}_{m'}^{\text{D-BS}}$. Each MNO n seeks to minimize the cost of its backhaul network, while maximizing the rate for each one of its SBSs $m \in \mathcal{M}_n$. To this end, the cooperative backhaul resource allocation problem can be formulated at each D-BS m with $\epsilon_e(m',m)=1$ as follows:

$$\underset{\boldsymbol{x}}{\operatorname{argmax}} \sum_{k\in\mathcal{K}} \left[r_m(k,m';\zeta) - \kappa_m q_{m'} \mathbb{1}_{m'm}\right] x_{m'km}, \tag{4.11a}$$

$$\text{s.t.} \quad r_m(m';\boldsymbol{x}) \leq \frac{1}{|\mathcal{M}^{\text{D-BS}}_{m'}|+1} r_{m'}(m'';\boldsymbol{x}), \qquad m' \in \mathcal{M}^{\text{D-BS}}_{m''} \tag{4.11b}$$

$$r_m(m';\boldsymbol{x}) \geq r_{m,\text{th}}, \tag{4.11c}$$

$$\sum_{k\in\mathcal{K}} x_{m'km} \leq K, \tag{4.11d}$$

$$\sum_{m\in\pi(m')} x_{m'km} \leq 1, \tag{4.11e}$$

$$x_{m'km} \in \{0,1\}, \tag{4.11f}$$

where $|.|$ denotes the set cardinality and $r_{m,\text{th}}$ denotes the minimum required rate for SBS m which is typically determined by the traffic that is circulating over the downlink of the radio access network. Constraint (4.11b) ensures that the backhaul capacity of each A-BS m' will be shared between its all assigned D-BSs in $\mathcal{M}^{\text{D-BS}}_{m'}$ as well as m''s traffic. That is why $|\mathcal{M}^{\text{D-BS}}_{m'}|$ is increased by one in (4.11b). This scheme allows every A-BS receiving its traffic from the core network in addition to the traffic of the associated D-BSs.

Prior to solving the proposed resource management problem, in (4.10a)-(4.10f) and (4.11a)-(4.11f), we note that the solution of network formation problem will depend on the resource allocation and vice versa. That is because for any multi-hop backhaul connection, the rate of a backhaul link $e(m',m)$, $r_m(m',\boldsymbol{x})$, depends on the network formation $\mathcal{M}^{\text{D-BS}}_{m'}$, as shown in (4.11b). Moreover, to associate a D-BS m to an A-BS in $\mathcal{M}^d_m$, the backhaul rates for A-BSs must be considered. Following, we propose a novel approach that allows to jointly solve these two problems.

4.3 Matching Theory for Multi-Hop Backhaul Network Formation

The problems in (4.10a)-(4.10f) and (4.11a)-(4.11f) are 0-1 integer programming which do not admit closed-form solutions and have exponential complexity [89]. To solve these problems, we propose a novel approach based on matching theory to derive a decentralized solution with tractable complexity [76, 90, 129].

To jointly solve the network formation and resource allocation problems, we propose a multi-stage framework, as shown in Fig. 4.2, using which the mmW-MBN can be formed as follows:

$$G_1(\mathcal{A}_1 \cup \mathcal{D}_1, \mathcal{E}_1) \rightarrow G_2(\mathcal{A}_2 \cup \mathcal{D}_2, \mathcal{E}_2) \rightarrow \cdots \rightarrow G_J(\mathcal{A}_J \cup \mathcal{D}_J, \mathcal{E}_J), \tag{4.12}$$

where the arrows in (4.12) indicate the transformation from sub-graph G_j to G_{j+1}, where $\mathcal{A}_{j+1} = \mathcal{D}_j$ and $\mathcal{D}_{j+1} = \left\{ m \in \bigcup_{m'\in\mathcal{A}_{j+1}} \mathcal{M}^d_{m'} \middle| m \notin \bigcup_{j'=1}^{j+1} A_{j'} \right\}$. Each sub-graph $G_j(\mathcal{A}_j \cup \mathcal{D}_j, \mathcal{E}_j)$ is defined as a directed graph from the set of A-BSs $\mathcal{A}_j$ to the set of D-BSs $\mathcal{D}_j$ via directed links in $\mathcal{E}_j$. Initially, $\mathcal{A}_1 = \{m_0\}$, and $\mathcal{D}_1 = \mathcal{M}^d_{m_0}$. Each stage j corresponds to the formation and resource management of j-th hop of the backhaul links. In fact, at each stage j, we address the following two problems: 1) in Subsections 4.3.1-4.3.2, we find $\mathcal{E}_j$ of sub-graph G_j that solves problem (4.10a)-(4.10f), given the rate of each backhaul link from the previous stages, and 2) in Section 4.4, we solve (4.11a)-(4.11f) for sub-graph G_j to allocate the sub-channels of each A-BS in $\mathcal{A}_j$ to its associated D-BSs. The variable J, resulting from the proposed solution, will yield the maximum number of hops for the multi-hop backhaul link from the MBS to SBSs. The final graph $G(\mathcal{M}, \mathcal{E}^*)$ is the overlay of all sub-graphs in (4.12), such that $\mathcal{E}^* = \bigcup_{j=1}^{J} \mathcal{E}_j$.

4.3.1 Multi-hop backhaul network formation problem as a matching game

At each stage j, the backhaul network formation problem can be cast as a one-to-many matching game [74] which is defined next.

Definition 5. Given two disjoint sets $\mathcal{A}_j$ and $\mathcal{D}_j$, the network formation policy π_j can be defined as a *one-to-many matching relation*, $\pi_j : \mathcal{A}_j \cup \mathcal{D}_j \to \mathcal{A}_j \cup \mathcal{D}_j$, such that

1) $\forall m \in \mathcal{D}_j$, if $\pi_j(m) \neq m$, then $\pi_j(m) \in \mathcal{A}_j$,
2) $\forall m' \in \mathcal{A}_j$, if $\pi_j(m') \neq m'$, then $\pi_j(m') \subseteq \mathcal{D}_j$,
3) $\pi_j(m) = m'$, if and only if $m \in \pi_j(m')$,
4) $\forall m' \in \mathcal{A}_j$, $|\pi_j(m')| \leq Q_{m'}$,

where $\pi_j(m) = m$ indicates that SBS m is unmatched.

The quota of A-BS m', $Q_{m'}$, represents the maximum number of D-BSs that can be assigned to m'. The relationship of the matching π_j with the link formation $\mathcal{E}_j$ is such that $m \in \pi_j(m')$ is equivalent to $\epsilon_e(m', m) = 1$. In addition, the matching policy π_j by definition satisfies the constraints in (4.10c)-(4.10f).

To complete the definition of the matching game, we must introduce suitable utility functions that will subsequently be used to define the preference profiles of all players. In the proposed mmW-MBN, in addition to the achievable rate, the cost of cooperation among MNOs must be considered in the preference relations of the SBSs. Here, we define the utility of D-BS $m \in \mathcal{D}_j$ that seeks to evaluate a potential connection to an A-BS $m' \in \mathcal{A}_j$, $U_m(m')$, as

$$U_m(m') = \min\left(\bar{r}_m(m'), r_{m'}(\pi_{j-1}(m'), \boldsymbol{x})\right) - \kappa_m q_{m'} \mathbb{1}_{mm'}, \tag{4.13}$$

where $\bar{r}_m(m')$ is given by (4.8). Here, we note that $r_{m'}(\pi_{j-1}(m'), \boldsymbol{x})$ is determined at stage $j-1$. If $j = 1$, then SBS m is directly connected to the MBS. The first term in (4.13) captures the fact that achievable rate for D-BS m is bounded by the backhaul rate of A-BS m'. The second term indicates that D-BS $m \in \mathcal{M}_n$ considers the cost of the backhaul link, if the A-BS does not belong to MNO n. However, if the A-BS belongs to MNO n, the cost will naturally be zero.

Furthermore, the utility of an A-BS $m' \in \mathcal{A}_j$ that evaluates the possibility of serving a D-BS $m \in \mathcal{D}_j$, $V_{m'}(m)$ will be:

$$V_{m'}(m) = \bar{r}_{m'}(m) + \kappa_{m'} q_{m'} \mathbb{1}_{mm'}. \tag{4.14}$$

In fact, (4.14) implies that A-BS m' aims to maximize the backhaul rate, while considering the revenue of providing backhaul support, if $\mathbb{1}_{mm'} \neq 0$. Based on the utilities in (4.13) and (4.14), the preference profiles of D-BSs and A-BSs will be given by:

$$m'_1 \succ^D_m m'_2 \iff U_m(m'_1) > U_m(m'_2), \tag{4.15}$$

$$m_1 \succ^A_{m'} m_2 \iff V_{m'}(m_1) > V_{m'}(m_2), \tag{4.16}$$

where $\succ^D$ and $\succ^A$ denote, respectively, the preference relations for D-BSs and A-BSs. Here, we assume that if $U_m(m') \leq 0$, A-BS $m' \in \mathcal{M}_{n'}$ will not be acceptable to D-BS $m \in \mathcal{M}_n$. This allows MNO n to choose the control parameter κ_m in (4.13), to prevent the formation of any link between a given D-BS m and any A-BS m' that is charging a high price for using its sub-channels. Given this formulation, we next propose an algorithmic solution for the proposed matching game that will allow finding suitable network formation policies.

4.3.2 Proposed mmW-MBN formation algorithm

To solve the formulated game and find the suitable network formation policy π_j for stage j, we consider two important concepts: *two-sided stability* and *Pareto optimality*. A two-sided stable matching is essentially a solution concept that can be used to characterize the outcome of a matching game. In particular, two-sided stability is defined as follows [90]:

Definition 6. A pair of D-BS $m \in \mathcal{D}_j$ and A-BS $m' \in \mathcal{A}_j$ in network formation policy π_j, $(m', m) \in \pi_j$, is a *blocking pair*, if and only if $m' \succ^D_m \pi_j(m)$ and $m \succ^A_{m'} m''$ for some $m'' \in \pi_j(m')$. A matching policy π_j is said to be *two-sided stable*, if there is no blocking pair.

The notion of two-sided stability ensures fairness for the SBSs. That is, if a D-BS m prefers the assignment of another D-BS m'', then m'' must be preferred by the A-BS $\pi_j(m'')$ to m, otherwise, π_j will not be two-sided stable. While two-sided stability characterizes the stability and fairness of a matching problem, the notion of Pareto optimality, defined next, can characterize the efficiency of the solution.

Definition 7. A matching policy π_j is said to be *Pareto optimal* (PO), if there is no other matching π'_j such that π'_j is equally preferred to π_j by all D-BSs, $\pi'_j(m) \succeq^D_m \pi_j(m)$, $\forall m \in \mathcal{D}_j$, and strictly preferred over π_j, $\pi'_j(m) \succ^D_m \pi_j(m)$ for some D-BSs.

To find the stable policy π_j, the *deferred acceptance* (DA) algorithm, originally introduced in [130], can be adopted. Hence, we introduce Algorithm 3 based on the DA algorithm which

Algorithm 3 Millimeter-Wave Mesh Backhaul Network Formation Algorithm

```
Inputs: $\mathcal{A}_j, \mathcal{D}_j, \succ^{A}_{m'}, \succ^{D}_{m}$.
Output: $\pi_j$.
Initialize: Temporary set of the rejected D-BSs $\mathcal{D}^r = \mathcal{D}_j$. Tentative set $\mathcal{A}^a_{m'} = \emptyset$ of accepted D-BSs
by A-BS $m'$, $\forall m' \in \mathcal{A}_j$. Let $\mathcal{S}_m = \mathcal{A}_j \cap \mathcal{M}^d_m, \forall m \in \mathcal{D}_j$.
while $\mathcal{D}^r \neq \emptyset$ do
    For each D-BS $m \in \mathcal{D}^r$, find the most preferred A-BS, $m'^* \in \mathcal{S}_m$, based on $\succ^D_m$. Each D-BS $m$ sends
    a link request signal to its corresponding $m'^*$.
    Add $m$ to $\mathcal{A}^a_{m'^*}$ and remove $m'^*$ from $\mathcal{S}_m$. If $\mathcal{S}_m = \emptyset$, remove $m$ from $\mathcal{D}^r$.
    Each A-BS $m' \in \mathcal{A}_j$ receives the proposals, tentatively accepts $Q_{m'}$ of the most preferred applicants
    from $\mathcal{A}^a_{m'}$, based on $\succ^A_{m'}$ and reject the rest.
    Remove rejected D-BSs from $\mathcal{A}^a_{m'}$ for every A-BS $m'$ and add them to $\mathcal{D}^r$. Remove accepted D-BSs
    from $\mathcal{D}^r$.
end while
```

proceeds as follows. Initially, no D-BS in $\mathcal{D}_j$ is assigned to an A-BS in $\mathcal{A}_j$. The algorithm starts by D-BSs sending a link request signal to their most preferred A-BS, based on their preference relation $\succ^D_m$. Next, each A-BS m' receives the request signals and approves up to $Q_{m'}$ of the most preferred D-BSs, based on $\succ^A_{m'}$ and rejects the rest of the applicants. The algorithm follows by rejected D-BSs applying for their next most preferred A-BS. Algorithm 3 converges once each D-BS m is assigned to an A-BS or is rejected by all A-BS in $\mathcal{A}_j \cap \mathcal{M}^d_m$. Since it is based on a variant of the DA process, Algorithm 3 is guaranteed to converge to a stable matching as shown in [130]. Moreover, among the set of all stable solutions, Algorithm 3 yields the solution that is PO for the D-BSs. Here, we note that the role of an SBS will change dynamically according to the changes of the CSI. However, due to the slow-varying channels, the CSI will remain relatively static within the channel coherence time (CCT), and consequently, the role of SBSs can be considered fixed within one CCT. The proposed distributed solution in Algorithm 3 allows the SBSs to update their preference profiles, which depend on the CSI, and accordingly their role, after each CCT period.

Given π_j resulted from Algorithm 3, the sub-channels of each A-BS $m' \in \mathcal{A}_j$ must be allocated to the D-BSs in $\pi_j(m')$. To this end, we next propose a distributed solution to solve the backhaul resource allocation problem.

4.4 Matching Theory for Distributed Backhaul Resource Management

To solve the problem in (4.11a)-(4.11f), centralized approaches will require MNOs to share the information from their SBSs with a trusted control center. Therefore, centralized approaches will not be practical to perform inter-operator resource management. To this end, we formulate the problem in (4.11a)-(4.11f) in each stage j as a second matching game and propose a novel distributed algorithm to solve the problem. The resulting resource allocation over stage j will determine the rate for each backhaul link in j-th hop. As discussed in the previous section, this information will be used to find the network formation policy π_{j+1} of the next stage $j+1$.

4.4.1 Resource management as a matching game

To allocate the sub-channels of an A-BS $m' \in \mathcal{A}_j$ to its associated D-BSs in $\pi_j(m')$, we consider a one-to-many matching game μ_j composed of two disjoint sets of mmW sub-channels, $\mathcal{K}$, and the D-BSs in $\pi_j(m')$ associated to A-BS m'. The matching μ_j can be formally defined, similar to the network formation matching π in Definition 5. However, unlike in the network formation matching game, here, we do not introduce any quota for the D-BSs. There are two key reasons for not considering quotas in our problem which can be explained as follows. First, for a resource allocation problem with minimum rate requirement, as presented in (4.11b) and (4.11c), a quota cannot be determined a priori for an SBS. That is because the number of sub-channels required by a given SBS is a function of the CSI over all sub-channels between an A-BS m' and all D-BSs assigned to m'. Therefore, sub-channel allocation for one D-BS will affect the number of required sub-channels by other D-BSs. This is a significant difference from classical solutions based on matching theory such as in [74, 76, 129], and [131]. The second practical reason for not using a fixed quota in our resource allocation problem is that there is no clear approach to determine the suitable quota values as a function of the various system metrics, such as CSI. On the other hand, with no constraint on the maximum number of sub-channels to be allocated to a D-BS, a distributed matching algorithm may assign all the sub-channels to a few of D-BSs, resulting in an inefficient allocation. Therefore, considering the significant impact of quota on the resource allocation, it is more practical to limit the number of allocated sub-channels by the natural constraint of the system, as presented in constraint (4.11b).

Therefore, we let a D-BS m assign a utility $\Psi_m(k;\mu_j)$ to a sub-channel k, where $(m,k) \notin \mu_j$, only if

$$\sum_{k' \in \mu_j(m)} r_m(k', \pi_j(m)) < \frac{1}{|\mathcal{M}^{\text{D-BS}}_{\pi_j(m)}| + 1} r_{\pi_j(m)}(m''; \boldsymbol{x}), \tag{4.17}$$

where m'' is the A-BS that serves $\pi_j(m)$, i.e., $m'' = \pi_{j-1}(\pi_j(m))$. Otherwise, $\Psi_m(k;\mu_j) = -\infty$, meaning that sub-channel k is not acceptable to D-BS m, given the current matching μ_j. In fact, (4.17) follows the rate constraint in (4.11b) and prevents the D-BS m from being allocated to unnecessary sub-channels. With this in mind, we define the utilities and preferences of sub-channels and D-BSs, considering the rate constraints, CSI, and the cost of each sub-channel. For any D-BS m, the utility that m achieves when being matched to a sub-channel k will be given by:

$$\Psi_m(k;\mu_j) = \begin{cases} r_m(k, \pi_j(m)), & \text{if (4.17) is held,} \\ -\infty, & \text{otherwise.} \end{cases} \tag{4.18}$$

Here, note that (4.18) does not include the price of the sub-channels. That is because the sub-channel price $q_{\pi_j(m)}$ is equal for all sub-channels and will not affect the preference of the D-BS.

The utility of sub-channels is controlled by their corresponding A-BS. The utility that is achieved by sub-channel k when being matched to a D-BS m will be:

$$\Phi_k(m) = r_m(k, \pi_j(m)) + \kappa_{\pi_j(m)} q_{\pi_j(m)} \mathbb{1}_{m\pi_j(m)}. \tag{4.19}$$

In (4.19), the second term represents the revenue obtained by A-BS $\pi_j(m)$ for providing backhaul support to D-BS m over sub-channel k. The scaling factor $\kappa_{\pi_j(m)}$ enables the A-BS to balance between the achievable rate and the revenue. In fact, as $\kappa_{\pi_j(m)}$ increases, a given A-BS will tend to assign more resources to D-BSs of other MNOs. Similar to (4.15) and (4.16), the preference profiles of sub-channels and D-BSs are given by

$$k_1 P_m^D k_2 \iff \mathit{\Psi}_m(k_1) > \mathit{\Psi}_m(k_2), \tag{4.20}$$

$$m_1 P_k^K m_2 \iff \mathit{\Phi}_k(m_1) > \mathit{\Phi}_k(m_2), \tag{4.21}$$

where P_m^D and P_k^K denote, respectively, the preference profiles of D-BS m and sub-channel k.

4.4.2 Proposed resource allocation algorithm for mmW-MBNs

Here, our goal is to find a two-sided stable and efficient PO matching μ_j between the sub-channels of A-BS m' and the D-BSs in $\pi_j(m')$, for every A-BS $m' \in \mathcal{A}_j$. From (4.18), we observe that the utility of a D-BS and, consequently, its preference ordering depend on the matching of the other D-BSs. This type of game is known as a *matching game with peer effect* [132]. This is in contrast with the traditional matching games in which players have strict and non-varying preference profiles. For the proposed matching game, we can make the following observation.

Proposition 2. Under the mmW-MBN specific utility functions in (4.18) and (4.19), the conventional DA algorithm is not guaranteed to yield a stable solution.

Proof. See Appendix B.1. □

Thus, we cannot directly apply the DA algorithm to our problem and we need to adopt a novel algorithm that handles blocking pairs and achieves a stable solution. To this end, we proposed a distributed resource allocation scheme in Algorithm 4. The algorithm proceeds as follows. For every A-BS $m' \in \mathcal{A}_j$, the initial set of rejected sub-channels is $\mathcal{K}^r = \mathcal{K}$. The algorithm initiates by each sub-channel $k \in \mathcal{K}^r$ sending a request signal to its most preferred D-BS, based on (4.21). The D-BSs receive the requests and accept a subset of most preferred sub-channels, based on (4.20), that satisfy their minimum rate requirement and reject the rest. The rejected sub-channels are added to $\mathcal{K}^r$. Accepted sub-channels and the sub-channels that are rejected by all D-BSs in $\pi_j(m')$ are removed from $\mathcal{K}^r$. Moreover, in step 9, D-BSs update their preferences P_m^D. The rejected sub-channels apply for their next most preferred D-BS from their preference profile. Algorithm 4 proceeds until $\mathcal{K}^r$ is an empty set. Next, any unmatched sub-channel $k \in \mathcal{K}$ is assigned to the most preferred D-BS m with $\mathbb{1}_{mm'} = 0$. The algorithm converges once all sub-channels are matched. Throughout this algorithm, we note that the corresponding A-BS sends the matching requests to D-BSs on the behalf of its sub-channels. The proposed Algorithm 4 exhibits the following properties.

Theorem 3. *Algorithm 4 is guaranteed to converge to a two-sided stable matching μ_j between sub-channels and D-BSs. Moreover, the resulting solution, among all possible stable matchings, is Pareto optimal for sub-channels.*

Algorithm 4 Backhaul Resource Allocation Algorithm

```
Inputs: $\mathcal{A}_j$, $\pi_j$, $r_{\text{th}}$.
Output: $\mu_j$.
for $i = 1, i \leq |\mathcal{A}_j|, i++$ do
    Initialize: Set A-BS $m'$ to $i$-th element of $\mathcal{A}_j$. Temporary set of the rejected sub-channels $\mathcal{K}^r = \mathcal{K}$. Tentative sets $\mathcal{D}_j^m = \emptyset$ and $r_m(m') = 0$ for each D-BS $m \in \pi_j(m')$. For each sub-channel $k$, let $\mathcal{C}_k = \pi_j(m')$.
    while $\mathcal{K}^r \neq \emptyset$ do
        For each sub-channel $k \in \mathcal{K}^r$, find the most preferred D-BS, $m^* \in \pi_j(m')$, based on $P_k^K$. A-BS sends a link request signal to the corresponding $m^*$ for each sub-channel. Add $k$ to $\mathcal{D}_j^{m^*}$ and remove $m^*$ from $\mathcal{C}_k$.
        If $\mathcal{C}_k = \emptyset$, remove $k$ from $\mathcal{K}^r$.
        Each D-BS $m \in \pi_j(m')$ receives the proposals and tentatively accepts the most preferred sub-channel from $\mathcal{D}_j^m$, based on $P_m^D$ and adds the corresponding rate of the accepted sub-channels to $r_m(m')$.
        If (4.17) is not met, add the next most preferred sub-channel and update $r_m(m')$, otherwise, reject the rest of sub-channels.
        Remove rejected sub-channels from $\mathcal{D}_j^m$ for every D-BS $m$ and add them to $\mathcal{K}^r$. Remove accepted sub-channels from $\mathcal{K}^r$.
        Update $\mu_j$ and $P_m^D$ for every D-BS.
    end while
    Based on current allocation, update $\Psi_m(k; \mu_j)$ for D-BSs with $\mathbb{1}_{mm'} = 0$.
    Let $\mathcal{K}'^r = \{k \in \mathcal{K} | \mu_j(k) = k\}$.
    while $\mathcal{K}'^r \neq \emptyset$ do
        Remove an arbitrary sub-channel $k$ from $\mathcal{K}'^r$ and allocate it to its most preferred D-BS $m$ from $\pi_j(m')$, with $\mathbb{1}_{mm'} = 0$, if (4.17) is held.
        Update $\mu_j$, $P_m^D$, and $r_m(m')$.
    end while
end for
```

Proof. See Appendix B.2. □

Here, we note that the proposed solution is Pareto optimal within each subgraph, corresponding to each stage, and it is assumed that network formation in subsequent subgraphs will not affect the utility functions in (4.13) and (4.14) for the SBSs in previous subgraphs. This assumption is valid, since a given D-BS will experience random interference from the interfering A-BSs. Given that the number of interfering A-BSs is large, which is true for backhaul networks that are supporting many SBSs, the average interference power in (4.13) and (4.14) will not depend on the network formation in subsequent subgraphs. Hence, the preference profiles of the D-BSs and A-BSs, and the consequent matching within each subgraph will be independent of the other subgraphs. Therefore, given that matching within each subgraph is Pareto optimal and is not affected by other subgraphs, the overall network formation is Pareto optimal in terms of maximizing the sum-rate.

4.4.3 Complexity analysis of the proposed multi-stage solution

First, we analyze the network formation complexity of an arbitrary stage j from Algorithm 3. For the purpose of complexity analysis, we consider the maximum number of requesting signals that D-BSs in $\mathcal{D}_j$ will send to the A-BSs in $\mathcal{A}_j$ before Algorithm 3 converges. In the worst-case scenario, i.e., the scenario with the highest conflict among D-BSs, all D-BSs in $\mathcal{D}_j$ have the same preference ordering. Let

$$m'_1 \succ_m m'_2 \succ_m \cdots \succ_m m'_{|\mathcal{A}_j|-1} \succ_m m'_{|\mathcal{A}_j|}, \tag{4.22}$$

be the preference ordering of all D-BSs $m \in \mathcal{D}_j$, where $\mathcal{A}_j = \{m'_1, m'_2, \cdots, m'_{|\mathcal{A}_j|}\}$. Hence, only $Q_{m'_i}$ D-BSs will be accepted by the A-BS m'_i during the i-th iteration of Algorithm 3. Moreover, the number of iterations I is an integer that satisfies

$$\sum_{i=1}^{I-1} Q_{m'_i} < |\mathcal{D}_j| \leq \sum_{i=1}^{I} Q_{m'_i}. \tag{4.23}$$

Therefore, the total number of requests sent by D-BSs will be

$$\begin{aligned}&|\mathcal{D}_j| + \left(|\mathcal{D}_j| - Q_{m'_1}\right) + \left(|\mathcal{D}_j| - Q_{m'_1} - Q_{m'_2}\right) + \cdots + \left(|\mathcal{D}_j| - Q_{m'_1} - \cdots - Q_{m'_{I-1}}\right) = \\ &I|\mathcal{D}_j| - \sum_{i=1}^{I-1}(I-i)Q_{m'_i} = I|\mathcal{D}_j| - I\sum_{i=1}^{I} Q_{m'_i} + \sum_{i=1}^{I} iQ_{m'_i} \leq \sum_{i=1}^{I} iQ_{m'_i},\end{aligned} \tag{4.24}$$

where (4.23) is used to derive the inequality in (4.24). For the special case in which $Q_{m'_i} = Q, \forall m'_i \in \mathcal{A}_j$, (4.23) implies that $(I-1)Q < |\mathcal{D}_j|$. Hence, (4.24) can be simplified to

$$\sum_{i=1}^{I} iQ_{m'_i} = \frac{1}{2}Q(I)(I+1) < \frac{1}{2}Q\left(\frac{|\mathcal{D}_j|}{Q} + 1\right)\left(\frac{|\mathcal{D}_j|}{Q} + 2\right). \tag{4.25}$$

Therefore, the complexity of stage j in Algorithm 3 is $\mathcal{O}(|\mathcal{D}_j|^2)$, which admits a second-order polynomial relation with respect to the number of D-BSs.
Similarly, for Algorithm 4, the worst case scenario is when all sub-channels of A-BS m' have the same preference ordering for D-BSs in $\pi_j(m')$ and only one sub-channel is accepted over each iteration. Therefore, the number of requesting signals sent from A-BS m' to its associated D-BSs in $\pi_j(m')$ will be at most

$$\begin{aligned}K + (K-1) + \cdots + (K - |\pi_j(m')| + 1) &= |\pi_j(m')|K - \frac{1}{2}\left(|\pi_j(m')|\right)\left(|\pi_j(m')| + 1\right) \\ &< KQ_{m'},\end{aligned} \tag{4.26}$$

where the inequality in (4.26) results from having $0 \leq |\pi_j(m')| \leq Q_{m'}$. Therefore, the total number of requesting signals is $\sum_{m' \in \mathcal{A}_j} KQ_{m'}$. For $Q_{m'_i} = Q, \forall m'_i \in \mathcal{A}_j$, the complexity of Algorithm 4 in stage j is $\mathcal{O}(KQ|\mathcal{A}_j|)$. Thus, the overall complexity of an arbitrary stage j of the proposed distributed solution is $\mathcal{O}(|\mathcal{D}_j|^2 + |\mathcal{A}_j|)$. This result implies that the complexity of the proposed distributed solution, composed of Algorithms 3 and 4, is bounded by a second-order polynomial with respect to the network size. This result shows that the proposed approach yields a solution with a manageable complexity for the two interrelated integer programming problems in (4.10a)-(4.10f) and (4.11a)-(4.11f).

Table 4.2: Simulation parameters

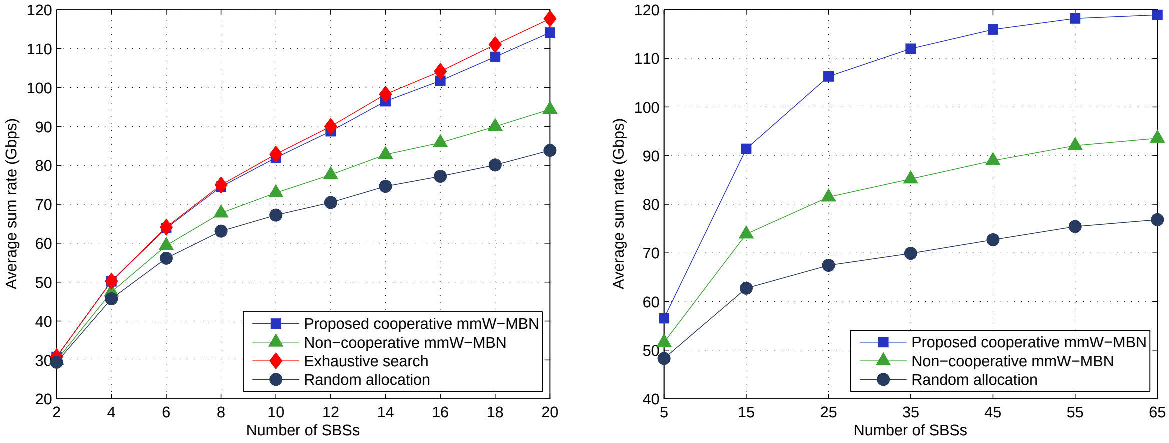
Notation	**Parameter**	**Value**
f_c	Carrier frequency	73 GHz
p_{t,m_0} and $p_{t,m}$	Total transmit power for MBS and SBSs	40 and 30 dBm
M	Total number of SBSs	3 to 65
N	Number of MNOs	3 to 5
Ω	Available Bandwidth	5 GHz
K	Number of sub-channel	50
(ξ_{LoS},ξ_{NLoS})	Standard deviation of mmW path loss	$(4.2, 7.9)$ [10]
(α_{LoS},α_{NLoS})	Path loss exponent	$(2, 3.5)$ [10]
d_0	Path loss reference distance	1 m [10]
G_{max}	Antenna main lobe gain	18 dB [14]
G_{min}	Antenna side lobe gain	-2 dB [14]
θ_m	beam width	10° [14]
σ^2	Noise power	-174 dBm/Hz + $10\log_{10}\frac{\Omega}{K}$
d_{max}	Radius of simulation area	400 m
d	SBSs communication range	200 m [28, 105]
r_{th}	Minimum required rate per SBS	1 Mbps
q	Unit of price per sub-channel	$1

4.5 Simulation Results

For our simulations, we consider a mmW-MBN with an MBS located at $(0, 0) \in \mathbb{R}^2$ and up to $M = 65$ SBSs distributed uniformly and randomly within a planar area with radius $d_{\text{max}} = 400$ m. The simulation parameters are summarized in Table 4.2. Moreover, the number of SBSs is considered to be equal for all MNOs. We compare our proposed approach with the following three other approaches:

1) *Optimal solution* obtained via an exhaustive search which finds the resource allocation that maximizes the backhaul sum rate. In fact, this benchmark explores all the possibilities for sub-channel allocation with uniform transmission power.
2) *Non-cooperative scheme* which follows the proposed algorithms for both network formation and resource allocation, however, cooperation among MNOs is not allowed. That is, the SBSs of an MNO do not provide backhaul support to the SBSs of other MNOs.
3) *Random allocation* that assigns D-BSs randomly to an A-BS within their communication range, subject to the constraints in (4.10b)-(4.10f). In addition, each A-BS randomly allocates sub-channels to its assigned D-BSs, subject to the constraints in (4.11b)-(4.11f).

All statistical results are averaged over a large number of independent runs.



(a) (b)

Figure 4.3: Average sum rate resulting from the proposed cooperative mm-MBN approach, non-cooperative scheme, random allocation, and the optimal solution as the number of SBSs varies.

4.5.1 Achievable backhaul sum rate of the mmW-MBN

Fig. 4.3a shows a performance comparison between the proposed framework with the optimal solution, non-cooperative, and random allocation approaches, for a mmW-MBN with $K = 7$ sub-channels, up to $M = 20$ SBSs, and $N = 2$ MNOs. Due to the computational complexity of the exhaustive search, for this comparison figure, a relatively small network size is considered. In Fig. 4.3a, the optimal solution and the random allocation provide, respectively, an upper and lower bound on the achievable sum-rate of the given network. Fig. 4.3a shows that the proposed cooperative framework based on matching theory yields a promising performance comparable with results from the optimal solution. In fact, the performance gap will not exceed 3.2% with the network size up to $M = 20$ SBSs. In addition, the results in Fig. 4.3a show that the proposed solution improves the sum-rate up to 21% and 36% compared to, respectively, the non-cooperative and the random allocation scheme.

In Fig. 4.3b, the average sum rate resulting from the proposed cooperative approach is compared with both the non-cooperative and random allocation schemes, for a dense mmW-MBN with $N = 5$ MNOs and up to $M = 65$ SBSs. From Fig. 4.3b, we can see that the average sum rate increases as the number of SBSs increases. This is due to the fact that more SBSs will be able to connect to the MBS via a multi-hop backhaul link. Fig. 4.3b shows that, the proposed approach outperforms both the non-cooperative and random allocation schemes for all network sizes. In fact, the proposed framework increases the average sum rate by 27% and 54%, respectively, compared to the non-cooperative and random allocation schemes, for $M = 65$ SBSs.

In Fig. 4.4, we compare the average sum rate for the proposed approach with the non-cooperative and random allocation schemes, versus the average LoS probability. The primary goal here is to analyze the severe impact of the blockage on the network performance. In particular, Fig. 4.4

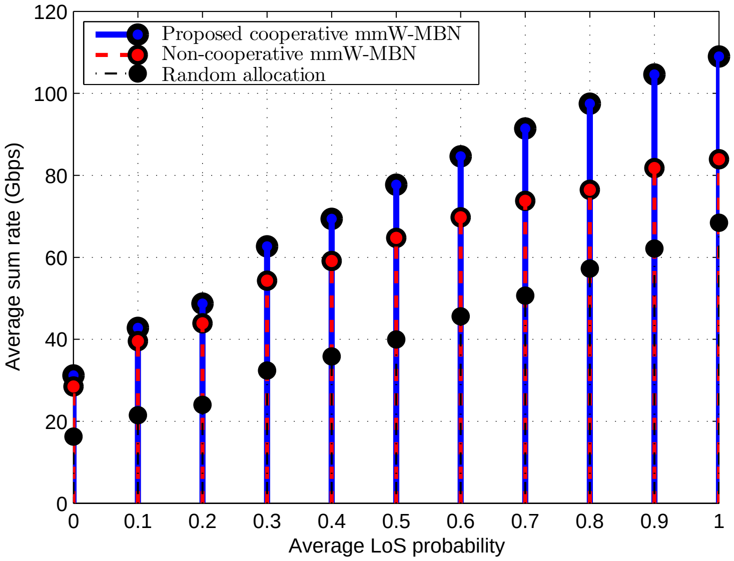


Figure 4.4: Average sum rate versus the average LoS probability ρ.

shows that blockage degrades the sum rate up to six times, when ρ decreases from 1 to 0. However, the results in Fig. 4.4 show that the proposed approach is more robust against blockage, compared to the non-cooperative and random allocation schemes. In fact, the proposed approach yields up to 25% and 42% performance gains for $\rho = 1$, respectively, compared to the non-cooperative and random allocation schemes. We note that for extreme blockage scenarios, e.g., $\rho = 0, 0.2$, it is expected that the gains will be small, since the achievable rate for most of the links is degraded by blockage. However, we can observe that as more LoS backhaul links become available, the performance gap increases. The main reason for this trend is that the backhaul rate for each A-BS increases, as ρ increases, which can support higher rates for its associated D-BSs. The average sum rate increases by 27% for the proposed cooperative mmW-MBN, as ρ increases from 0.6 to 1.

4.5.2 Statistics of the achievable rate for the mmW-MBN

Fig. 4.5 shows the cumulative distribution function (CDF) of the average sum rate for the proposed cooperative approach, compared to the non-cooperative and random allocation schemes, for $N = 5$ MNOs and $M = 60$ SBSs. The results show that the proposed cooperative approach substantially improves the statistics of the average sum rate. For example, Fig. 4.5 shows that the probability of achieving a 80 Gbps target sum rate is 90%, 36%, and 20%, respectively, for the proposed approach, the non-cooperative scheme, and the random allocation.

Fig. 4.6 shows the empirical CDF of the average sum rate for different average LoS probabilities, for $N = 5$ MNOs and $M = 20$ SBSs. From this figure, we can see that severe blockage with small ρ significantly degrades the performance of the mmW-MBN. Interestingly, we can observe that with $\rho = 1$, the average sum rate does not fall below the 40 Gbps. However, as the probability of

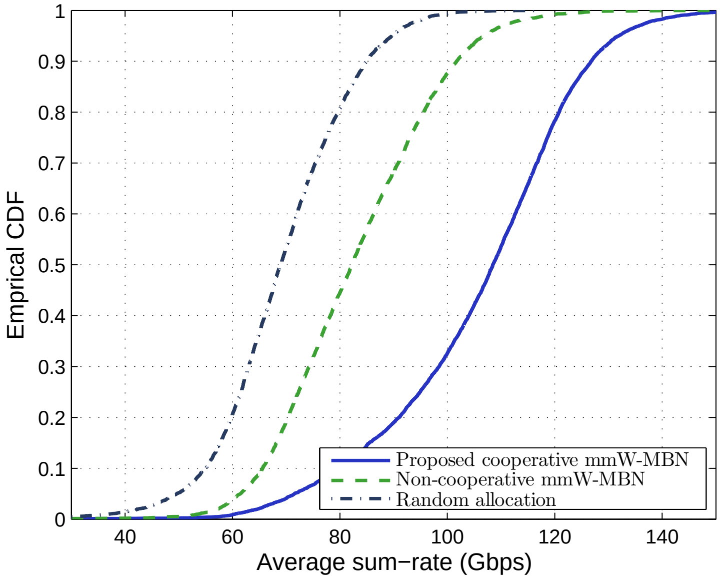


Figure 4.5: The CDF of the average sum rate resulting from the proposed cooperative mmW-MBN, the non-cooperative baseline, and the random allocation approach.

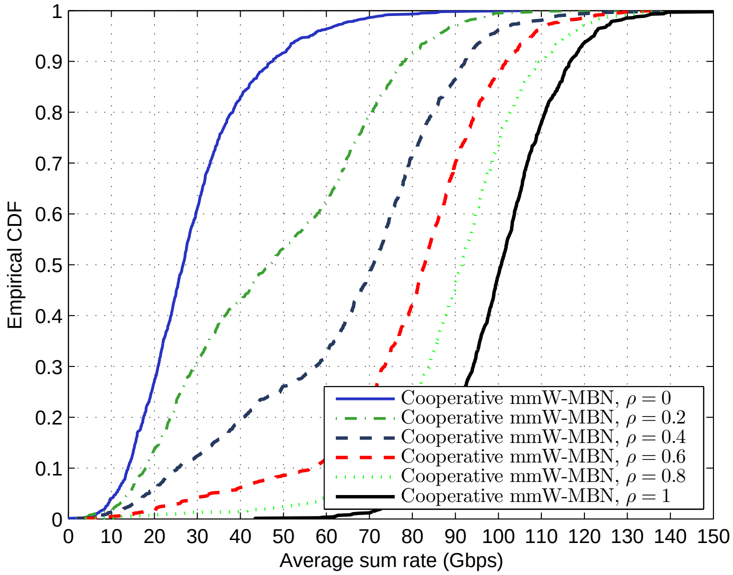


Figure 4.6: The empirical CDF of the sum rate for different average LoS probabilities.

LoS decreases to 0.2, the probability of the average sum rate be less than 40 Gbps is 42%.

4.5.3 Economics of the proposed mmW-MBN framework

Fig. 4.7 provides a design guideline to manage pricing and the cost of the cooperative mmW-

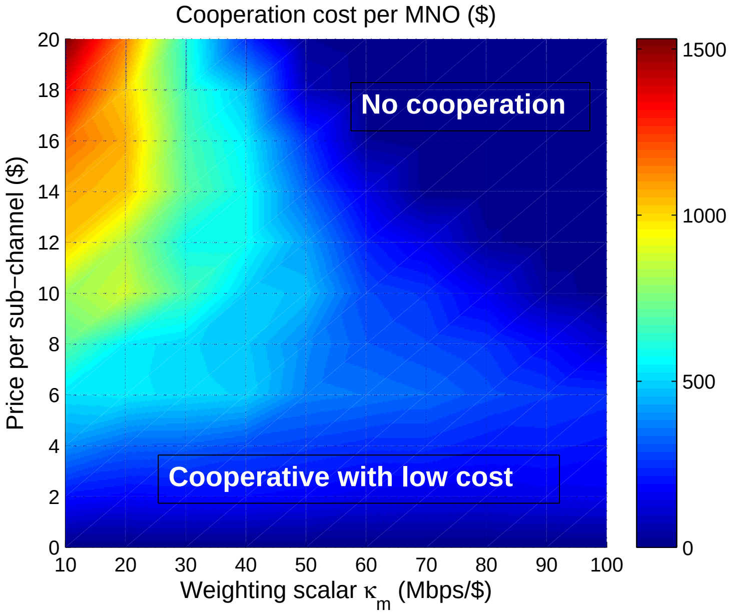
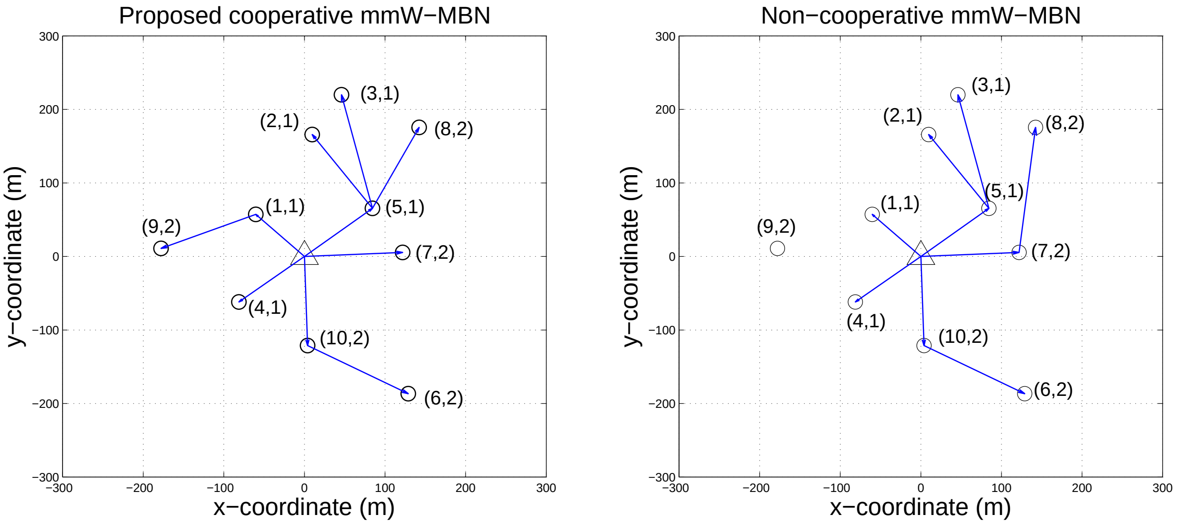


Figure 4.7: Cooperation cost per MNO as a function of both sub-channel price and the weighting parameter κ_m.

MBN for the MNOs. In this figure, the cost of cooperation per MNO is shown as the price per sub-channel q and the weighting parameter κ vary, for $N = 3$ MNOs and $M = 15$ SBSs. The weighting parameter κ_m, which is first defined in (4.10a) allows each MNO to control the cost of its backhaul network, with respect to q that is determined by other MNOs. *Here, we explicitly define the backhaul cost for an MNO $n \in \mathcal{N}$, as the total money that MNO n must pay to other MNOs for receiving backhaul support to the SBSs in $\mathcal{M}_n$.* A larger κ_m implies that the MNO has less incentive to cooperate with other MNOs. Hence, as shown in Fig. 4.7, no cooperation will happen between MNOs, as both q and κ_m increase, labeled as the *no cooperation* region. As an example, if the budget of an arbitrary MNO n is $\$500$ and $q = \$10$ is chosen by other MNOs, from Fig. 4.7, we can see that MNO n must choose $\kappa_m \geq 40\,\text{Mbps}/\$$ in order to keep the cost less than its budget. In addition, Fig. 4.7 will provides a systematic approach to determine a suitable pricing mechanism for an MNO, if the model parameter κ_m and the budget for other MNOs are known. This initial result can be considered as a primary step towards more complex models, which may consider dynamic pricing policies and competing strategies for MNOs.

Moreover, we note that the economic gains of the proposed cooperative framework are indirectly reflected in the performance gains of the proposed scheme, compared to the non-cooperative and random allocation approaches, as shown in Figs. 4.3a-4.5. Such an increase in the data rate of the backhaul network, resulting from the cooperative framework, will provide additional revenues for the MNOs, either by offering services with higher QoS to the users, or by increasing the users served by each SBS. Here, we note that the revenue for each MNO explicitly depends on the cost of maintenance per SBS, leasing the spectrum, deployment of SBSs, providing power supply for SBSs, service plans by MNOs and other specific metrics that may differ from one geographical area to another. Therefore, there is no direct and general mechanism to map the physical layer metrics,



(a) Proposed cooperative mmW-MBN

(b) Non-cooperative mmW-MBN

Figure 4.8: A snapshot of multi-hop mmW backhaul network via the proposed cooperative scheme and the non-cooperative baseline approach.

such as rate into revenue. However, such a mapping is definitely being used by the economic departments of global operators to define their KPI performance metrics.

Consequently, in Fig. 4.7, we have shown the robustness of the proposed framework with regard to the pricing mechanisms. In fact, we have shown that the proposed resource management framework allows MNOs to choose whether to cooperate or not, depending on the system metrics, including the rate, their available budget, and the pricing policy by other MNOs.

4.5.4 Snapshot of the mmW-MBN

Fig. 4.8 shows a snapshot of the mmW multi-hop backhaul network (mmW-MBN) for both the proposed cooperative scheme and the non-cooperative baseline approach. In this figure, each circle shows an SBS and corresponding pair (m, n) means SBS m belongs to MNO n ($m \in \mathcal{M}_n$). Moreover, the MBS is shown by a triangle. For illustration purposes, we show the network for $N = 2$ MNOs and a total of $M = 10$ SBSs, and the quota for each A-BS is $Q_m = 5$.

From Fig. 4.8b, we first observe that SBS $9 \in \mathcal{M}_2$ is not connected to the non-cooperative mmW-MBN, since no other SBS belonging to MNO 2 is located within SBS 9's communication range. That is, MNO 2 must increase the density of its SBSs to provide ubiquitous backhaul connectivity. However, deploying additional SBSs will increase the costs for the MNO, including site rental costs, power consumption, and cell maintenance, among others. In contrast, the proposed cooperative scheme provides backhaul support for the SBS $9 \in \mathcal{M}_2$ via SBS $1 \in \mathcal{M}_1$ that belongs to the MNO 1, as shown in Fig. 4.8a. Second, Fig. 4.8 shows that SBS $8 \in \mathcal{M}_2$ is connected to A-BSs $7 \in \mathcal{M}_2$ and $5 \in \mathcal{M}_1$, respectively, in the non-cooperative and proposed cooperative

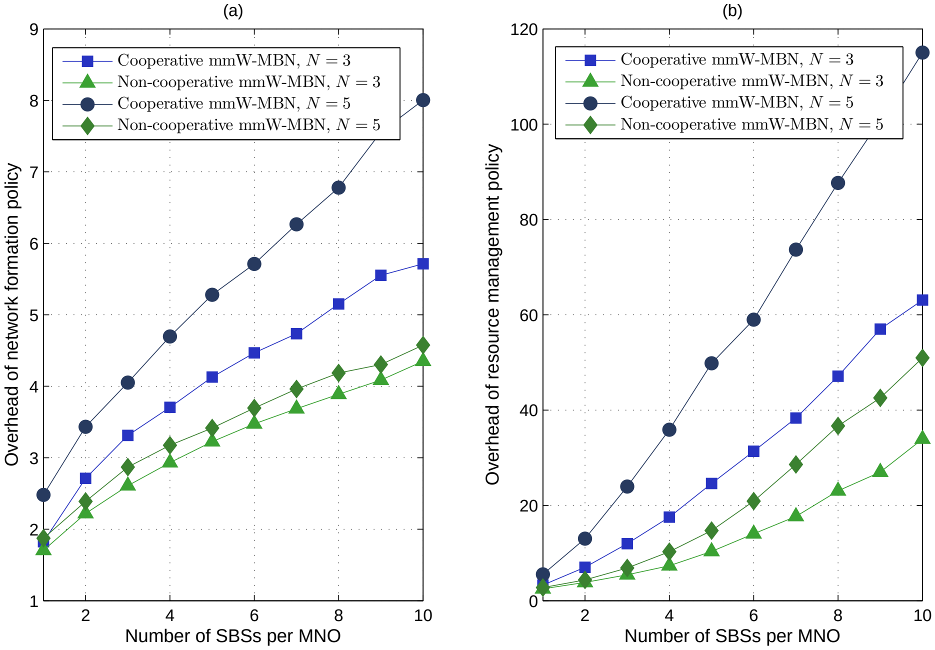


Figure 4.9: The average overhead of the network formation and resource allocation algorithms.

mmW-MBNs. We can easily observe that the proposed cooperative scheme provides a shorter path via a two-hop backhaul link for the SBS $m = 8$, compared to the non-cooperative approach.

4.5.5 Complexity analysis

In Fig. 4.9, the average signaling overhead of the proposed network formation and the resource allocation algorithms are analyzed, respectively, in (a) and (b). Here, the overhead captures the number of messages that must be exchanged between A-BSs and D-BSs. Fig. 4.9.a shows that the overhead of the proposed network formation policy increases with the number of SBS per MNOs, since the sets of A-BSs and D-BSs grow as more SBSs are deployed. However, we can see that the algorithm converges fast for all network sizes. Moreover, it can be observed that the proposed cooperative approach increases the overhead by 28% for $M = 18$ SBSs. This is because the proposed approach allows each SBS to communicate with more number of SBSs, compared to the non-cooperative scheme. Similarly, in Fig. 4.9.b, the overhead of the resource allocation algorithm increases as the number of SBSs increases. Regarding the complexity of the proposed approach, we note the followings:1) an optimal solution will require an exponential complexity which is not tractable for dense mmW network deployments, while the proposed approach yields a close-to-optimal performance, as shown in Fig. 4.3b, while requiring a manageable signaling overhead, 2) in this work, we have considered MBS as the only gateway, with a fiber backhaul, to the core network. Therefore, the scenario that is considered in our simulation is an extreme case, as in practice, there will be more than one gateway for up to $M = 65$ SBSs. Clearly, increasing the number of gateways will reduce the size of the problem, i.e., the number of SBSs to be managed, and 3) communication signals required by the proposed scheme will be incorporated within the common control signals of the system.

4.6 Summary

In this chapter, we have proposed a novel distributed backhaul management approach for analyzing the problem of resource management in multi-hop mmW backhaul networks. In particular, we have formulated the problem within a matching-theoretic framework composed of two, dependent matching games: a network formation game and a resource management game. For the network formation game, we have proposed a deferred acceptance-based algorithm that can yield a two-sided stable, Pareto optimal matching between the A-BSs and D-BSs. This matching represents the formation of the multi-hop backhaul links. Once the network formation game is determined, we have proposed a novel algorithm for resource management that allocates the sub-channels of each A-BSs to its associated D-BSs. We have shown that the proposed resource management algorithm is guaranteed to converge to a two-sided stable and Pareto optimal matching between the sub-channels and the D-BSs. Simulation results have shown that the proposed cooperative backhaul framework provides substantial performance gains for the network operators and incentivizes sharing of the backhaul links.

4.7 Appendix B

B.1 Proof of Proposition 2

We prove this using an example. Let $\mathcal{D}_j = \{m_1, m_2\}$, with preference profiles $k_1 P^D_{m_1} k_2 P^D_{m_1} k_3$ and $k_2 P^D_{m_2} k_3 P^D_{m_2} k_1$. Moreover, let $\mathcal{K} = \{k_1, k_2, k_3\}$ with preference profiles $m_1 P^K_{k_i} m_2$, for $i = 1, 2$, and $m_2 P^K_{k_3} m_1$. Considering achievable rates $r_{m_1}(k_1)$, $r_{m_2}(k_2)$ and $r_{m_2}(k_3)$ are greater than r_{th}, DA algorithm yields a matching μ where $\mu(k_1) = \{m_1\}$, $\mu(k_2) = \emptyset$, and $\mu(k_3) = \{m_2\}$. However, (m_2, k_2) form a blocking pair, which means μ is not stable.

B.2 Proof of Theorem 3

Algorithm 4 will always converge, since no sub-channel will apply for the same D-BS more than once, through steps 3-12 or 13-16. Next, we show that the proposed algorithm always converges to a two-sided stable matching. To this end, let D-BS m and sub-channel k form a blocking pair (m, k). That is, $m P^K_k \mu_j(k)$ and $k P^D_m k'$, where $k' \in \mu_j(m)$. We show that such a blocking pair does not exist. To this end, we note that there are two possible cases for sub-channel k: 1) $k \in \mathcal{K}'^r$, and 2) $k \notin \mathcal{K}'^r$, in step 12.

If $k \in \mathcal{K}'^r$ in step 12, that means k is unmatched and the minimum rate requirement is satisfied for all D-BSs, including m. Thus, (m,k) is a blocking pair only if $\mathbb{1}_{mm'} = 0$, otherwise, m will not accept more sub-channels. However, in step 14, k will be assigned to its most preferred D-BS $\mu_j(k)$ with $\mathbb{1}_{\mu_j(k)m'} = 0$, meaning that $\mu_j(k) P_k^K m$. Hence, (m,k) cannot be a blocking pair.

Next, if $k \notin \mathcal{K}'^r$, then k is matched to a D-BS $\mu_j(k)$ prior to step 12. Here, $m P_k^K \mu_j(k)$ implies that k has applied for m before $\mu_j(k)$ and is rejected. Thus, $k' P_m^D k$, for all $k' \in \mu_j(m)$. Therefore, (m,k) cannot be a blocking pair and matching μ_j is two-sided stable.

To prove Pareto optimality, we show that no sub-channel k can improve its utility by being assigned to another D-BS m, instead of $\mu_j(k)$. If $m P_k^K \mu_j(k)$, it means $k' P_m^D k$, for all $k' \in \mu_j(m)$, due to the two-sided stability of μ_j. Hence, a new matching μ_j' that allocates k to m instead of a sub-channel $k' \in \mu_j(m)$ will make (m,k') to be a blocking pair for μ_j'. Therefore, no other stable matching exists that improves the utility of a sub-channel.

Chapter 5

Millimeter Wave Communications for Enhanced Mobility Management

5.1 Background, Related Works, and Summary of Contributions

The proliferation of bandwidth-intensive wireless applications such as social networking, high definition video streaming, and mobile TV have drastically strained the capacity of wireless cellular networks. To cope with this traffic increase, several new technologies are anticipated for 5G cellular systems: 1) dense deployment of SBSs, 2) exploitation of the large amount of available bandwidth at mmW frequencies, and 3) enabling of *content caching* directly at the UEs to reduce delay and improve QoS. The dense deployment of SBSs with reduced cell-sizes will boost the capacity of wireless networks by decreasing UE-SBS distance, removing coverage holes, and improving spectral efficiency. Meanwhile, mmW communications will provide high data rates by leveraging directional antennas and transmitting over a large bandwidth that can reach up to 5 GHz. In addition, exploiting the high storage capacity of the modern smart handhelds to cache the data at the UE increases the flexibility and robustness of resource management, in particular, for *mobile UEs (MUEs)*. In fact, caching allows the network to store the data content in advance, while enabling MUEs to use the cached content when sufficient wireless resources are not available.

However, dense HetNets, composed of MBSs and SBSs with various cell sizes, will introduce three practical challenges for mobility management. First, MUEs will experience frequent HOs, while passing SBSs with relatively small cell sizes, which naturally increases the overhead and delay in HetNets. Such frequent HOs will also increase HOF, particularly for MUEs that are moving at high speeds [133]. In fact, due to the small and disparate cell sizes in HetNets, MUEs will not be able to successfully finish the HO process by the time they trigger HO and pass a target SBS. Second, the inter-frequency measurements that are needed to discover target SBSs can be excessively power consuming and detrimental for the battery life of MUEs, especially in dense

HetNets with frequent HOs. Third, μW frequencies are stringently congested, and thus, frequent HOs may introduce unacceptable overhead and limit the available frequency resources for the static users. In this regard, offloading MUEs from heavily utilized μW frequencies to mmW frequencies can substantially improve the spectral efficiency at the μW network.

5.1.1 Related works

In [17], the authors provide a comprehensive overview on mobility management in IP networks. The authors in [18] present different distributed mobility management protocols at the transport layer for future dense HetNets. In [19], an energy-efficient SBS discovery method is proposed for HetNets. The work in [20] investigates HO decision algorithms that focus on improving HO between femtocells and LTE-Advanced systems. The work presented in [21] overviews existing approaches for vertical handover decisions in HetNets. In [22], the authors study the impact of channel fading on mobility management in HetNets. In addition, the work in [22] shows that increasing the sampling period for HO decision decreases the fading impact, while increasing the ping-pong effect. In [23], the authors propose an HO scheme that takes into account the speed of MUEs to decrease frequent HOs in HetNets. The authors in [24] propose an HO scheme that supports soft HO by allowing MUEs to connect with both an MBS and SBSs. Furthermore, a distributed mobility management framework is proposed in [25] which uses multiple frequency bands to decouple the data and control planes.

Although interesting, the body of work in [17–25] does not consider mmW communications and caching capabilities for mobility management and solely focuses on HetNets operating over μW frequencies. In addition, it does not study the opportunities that caching techniques can provide for mobility management. In [134], an HO scheme for mmW networks is proposed in which the MBS acts as an anchor for mmW SBSs to manage control signals. However, [134] assumes that LoS mmW links are always available and provides no analytical results to capture the directional nature of mmW communications. In [135], the authors propose a resource allocation scheme for hybrid mmW-μW networks that enhances video streaming by buffering content over mmW links. However, [135] does not address any mobility management challenge, such as frequent HOs or HOF. Our early work in [136] provided some of the basic insights on mobility management in μW-mmW networks. However, in contrast to this work, [136] solely focuses on an average performance analysis, does not consider dynamic HO problem for multi-MUE scenarios, and does not propose any energy management mechanisms for handling inter-frequency measurements.

Proactive caching for enhancing mobility management has been motivated by the works in [44–47]. In [44], the authors discuss the potential of content caching at either evolved packet core network or radio access network to minimize the traffic overhead at the core network. Moreover, the authors in [45] propose a proactive caching framework in which an ongoing IP service can be cached in advance and continuously transferred among different data centers as MUEs move across different cells. In [46], the authors propose a caching framework that stores different parts of a content at different base stations, allowing MUEs to randomly move across different cells and download different cached parts of the original content whenever possible. In addition, in [47], a proactive caching solution is proposed for mobility management by exploiting MUEs' trajectory information. Although interesting, the body of work in [44–47] focuses on adopting protocols that are designed for higher network layers. Moreover, these solutions do not consider caching directly at the MUEs and focus on mobility management at the core network. However, we will show how leveraging high capacity mmW communication complements the notion of caching at MUEs. In addition, caching at MUEs will provide opportunities to perform *transparent* HOs in HetNets, without requiring any data session with a target SBS.

5.1.2 Summary of contributions

The main contribution of this chapter is a novel mobility management framework that addresses critical handover issues, including frequent HOs, HOF, and excessive energy consumption for seamless HO in emerging dense wireless cellular networks with mmW capabilities. In fact, we propose a model that allows MUEs to cache their requested content by exploiting high capacity mmW connectivity whenever available. As such, the MUEs will use the cached content and avoid performing any HO, while passing SBSs with relatively small cell sizes. First, we propose a geometric model to derive tractable, closed-form expressions for key performance metrics, including the probability of caching, cumulative distribution function of caching duration, and the average data rate for caching at an MUE over a mmW link. Moreover, we provide insight on the achievable gains for reducing the number of HOs and the average HOF, by leveraging caching in mmW-μW networks. Then, we formulate the proposed cache-enabled mobility management framework as a dynamic matching game, so as to provide a distributed solution for mobility management in HetNets, while taking the dynamics of the system into account. To solve the formulated dynamic matching problem, we first show that conventional algorithms such as the deferred acceptance algorithm adopted in [130] and [76], fail to guarantee a dynamically stable HO between MUEs and SBSs. Therefore, we propose a novel distributed algorithm that is guaranteed to converge to a dynamically stable HO policy in dense HetNets. Subsequently, the complexity of the proposed algorithm in terms of signaling overhead is analyzed. Under practical settings, we show that the proposed cache-enabled HO framework can decrease the average HOF rate by up to 45%, even for MUEs with high speeds. In addition, simulation results provide insights on the achievable gains by the proposed distributed algorithm, in terms of reducing energy consumption for cell search, as well as increasing traffic offloads from the μW frequencies.

The rest of this chapter is organized as follows. Section 5.2 presents the system model. Section 5.3 presents the analysis for caching in mobility management. Performance analysis of the cache-enabled mobility management is provided in Section 5.4. Section 5.5 formulates the mobility management as a dynamic matching and presents the proposed algorithm. Simulation results are presented in Section 5.6 and conclusions are drawn in Section 5.7.

Table 5.1: Variables and notations

Notation	Description	Notation	Description
K	Number of SBSs	$\mathcal{K}$	Set of SBSs
U	Number of MUEs	$\mathcal{U}$	Set of MUEs
θ_u	Moving angle of MUEs	v_u	Speed of MUEs
p_k	Transmit power of SBS k	B	Segment size of video (bits)
$\mathit{\Omega}_u$	Cache size of MUE u	$\mathit{\Omega}_u^{\text{max}}$	Maximum cache size
t_u^c	Caching duration of MUE u	Q	Video play rate
$\bar{R}^c(u,k)$	Average achievable caching rate	d^c	Traversed distance using cached content
$\mathit{\Delta T}$	Time-to-trigger (TTT)	r^c	Traversed distance in caching duration
T_s	Inter-frequency cell scanning interval	t_{MTS}	Minimum time-of-stay (ToS)
θ_k	Beamwidth for SBS k	E^s	Consumed energy per cell search
$t_{u,k}$	Time-of-stay for MUE u at SBS k	t_{MTS}	Minimum required time-of-stay

5.2 System Model

Consider a HetNet composed of an MBS and K SBSs within a set $\mathcal{K}$ distributed uniformly across an area. Each SBS $k \in \mathcal{K}$ can be viewed as a picocell or a femtocell, depending on its transmit power p_k. Picocells are typically deployed in outdoor venues while femtocells are relatively low-power and suitable for indoor deployments. The SBSs operate at μW frequencies that are different than those used by the MBS and, thus, there is no interference between SBSs and the MBS [19, 137]. The SBSs are also equipped with mmW front-ends to serve MUEs over either mmW or μW frequency bands [138]. The dual-mode capability allows to integrate mmW and μW RATs at the MAC layer of the air interface and reduce the delay and overhead for fast vertical handovers between both RATs [138]. Within this network, we consider a set $\mathcal{U}$ of U MUEs that are distributed randomly and that move across the considered geographical area during a time frame T. Each user $u \in \mathcal{U}$ moves in a random direction $\theta_u \in [0, 2\pi]$, with respect to the $\theta = 0$ horizontal angle, which is assumed fixed for each MUE over a considered time frame T. In addition, we consider that an MUE u moves with an average speed $v_u \in [v_{\text{min}}, v_{\text{max}}]$. The MUEs can receive their requested traffic over either the mmW or the μW band.

5.2.1 Channel model

The large-scale channel effect over mmW frequencies for a link between an SBS k and an MUE $u \in \mathcal{U}$, in dB, is given by[1]:

$$L(u,k) = 20\log_{10}\left(\frac{4\pi r_0}{\lambda}\right) + 10\alpha\log_{10}\left(\frac{r_{u,k}}{r_0}\right) + \chi, \tag{5.1}$$

[1] The free space path loss model in (5.1) has been adopted in many existing works, such as in [10], that carry out real-world measurements to characterize mmW large scale channel effects.

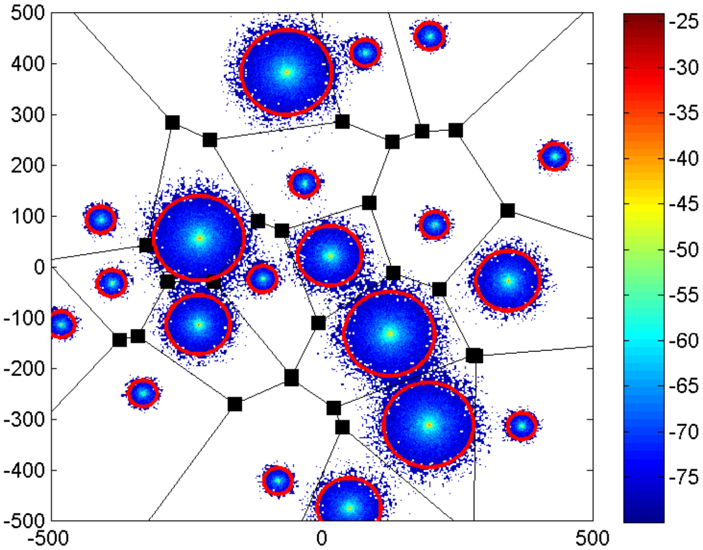


Figure 5.1: SBSs coverage with RSS threshold of -80 dB. Red circles show the simplified cell boundaries.

where (5.1) holds for $r_{u,k} \geq r_{\text{ref}}$, with r_{ref} and $r_{u,k}$ denoting, respectively, the reference distance and distance between the MUE u and SBS k. In addition, α is the path loss exponent, λ is the wavelength at carrier frequency $f_c = 73$ GHz over the E-band, due to the low oxygen absorption, and χ is a Gaussian random variable with zero mean and variance ξ^2. The path loss parameters α and ξ will have different values, depending on whether the mmW link is LoS or NLoS. Over the μW frequency band, the path loss model follows (5.1), however, with parameters that are specific to sub-6 GHz frequencies.

An illustration of the considered HetNet is shown in Fig. 5.1. The coverage for each SBS at the μW frequency is shown based on the maximum received signal strength (max-RSS) criteria with a threshold of -80 dB. White spaces in Fig. 5.1 delineate the areas that are covered solely by the MBS. Here, we observe that shadowing effect can adversely increase the ping-pong effect for MUEs. To cope with this issue, the 3GPP standard suggests L1/L3 filtering which basically applies averaging to RSS samples, as explained in [22].

5.2.2 Antenna model and configuration

To overcome the excessive path loss at the mmW frequency band, the MUEs will be equipped with electronically steerable antennas which allow them to achieve beamforming gains at a desired direction. The antenna gain pattern for MUEs follows the simple and widely-adopted sectorized pattern which is given by [14]:

$$G(\theta) = \begin{cases} G_{\text{max}}, & \text{if} \quad \theta < |\theta_m|, \\ G_{\text{min}}, & \text{otherwise}, \end{cases} \tag{5.2}$$

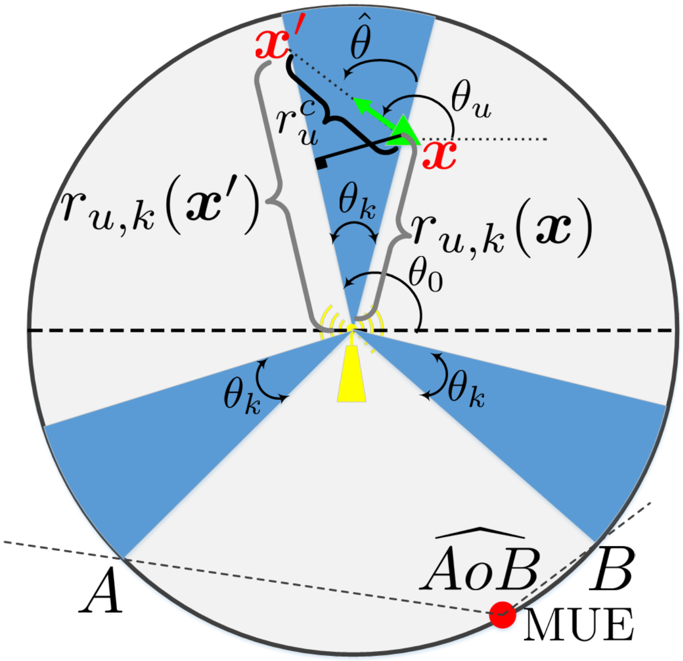


Figure 5.2: Antenna beam configuration of a dual-mode SBS with $N_k = 3$. Shaded areas show the mmW beams.

where θ and θ_m denote, respectively, the azimuth angle and the antennas' main lobe beamwidth. G_{max} and G_{min} denote, respectively, the antenna gain of the main lobe and side lobes. For SBSs, we use a model similar to the sectorized pattern in (5.2), however, we allow each SBS k to form N_k beams, either by using N_k antenna arrays or forming multi-beam beamforming. The beam patten configuration of an SBS k is shown in Fig. 5.2, where $N_k = 3$ equidistant beams in $\theta \in [0, 2\pi]$ are formed. To avoid the complexity and overhead of beam-tracking for mobile users, the direction of the SBSs' beams in azimuth is fixed. In fact, an MUE can connect to an SBS k over a mmW link, if the MUE traverses the area covered by the k's mmW beams. It is assumed that for a desired link between an SBS k and an MUE u, the overall transmit-receive gain is $\psi_{u,k} = G^2_{\text{max}}$.

5.2.3 Traffic model

Video streaming is one of the wireless services with most stringent QoS requirement. Meeting the QoS demands of such services is prone to the delay caused by frequent handovers in HetNets. In addition, HOFs can significantly degrade the performance by making frequent service interruptions. Therefore, our goal is to enhance mobility management for MUEs that request video or streaming traffic. Each video content is partitioned into small segments, each of size B bits. The network incorporates caching to transmit incoming video segments to an MUE, whenever a high capacity mmW connection is available. In fact, high capacity mmW connection, if available, allows to cache a large portion or even the entire video in a very short period of time. We define the cache size of $\Omega_u(k)$ for an arbitrary MUE u, associated with an SBS k, as the number of video segments that can be cached at MUE u as follows:

$$\Omega_u(k) = \min\left\{\left\lfloor \frac{\bar{R}^c(u,k)t_u^c}{B} \right\rfloor, \Omega_u^{\max}\right\}, \tag{5.3}$$

where $\lfloor . \rfloor$ and $\min\{.,.\}$ denote, respectively, the floor and minimum operands and $\Omega_u^{\max}$ is the maximum cache size. In addition, t_u^c is the *caching duration* which is equal to the time needed for an MUE u to traverse the mmW beam of its serving SBS. Considering the small green triangle in Fig. 5.2 as the current location of an MUE crossing a mmW beam, the caching duration will be $t_u^c = r_u^c/v_u$ where r_u^c is the distance traversed across the mmW beam. Moreover, $\bar{R}^c(u,k)$ is the *average achievable rate* for the MUE u during t_u^c. Given $\Omega_u(k)$ and the video play rate of Q, specified for each video content, the distance an MUE u can traverse with speed v_u, while playing the cached video content will be

$$d^c(u,k) = \frac{\Omega_u(k)}{Q}v_u. \tag{5.4}$$

In fact, the MUE can traverse a distance $d^c(u,k)$ by using the cached video content after leaving its serving cell k, without requiring an HO to any of the target cells. Meanwhile, the location information and control signals, such as paging, can be handled by the MBS during this time. As we discuss in details, such caching mechanism will help MUEs to avoid redundant cell search and HOs, resulting in an efficient mobility management in dense HetNets.

5.2.4 Handover procedure and performance metrics

The HO process in the 3GPP standard proceeds as follows: 1) Each MUE will do a cell search every T_s seconds, which can be configured by the network or directly by the MUEs, 2) If any target cell offers an RSS plus a hysteresis that is higher than the serving cell, even after L1/L3 filtering of input RSS samples, the MUE will wait for a time-to-trigger (TTT) of ΔT seconds to measure the average RSS from the target cell, 3) If the average RSS is higher than that of the serving SBS during TTT, the MUE triggers HO and sends the measurement report to its serving cell. The averaging over the TTT duration will reduce the ping-pong effect resulting from instantaneous CSI variations, and 4) HO will be executed after the serving SBS sends the HO information to the target SBS.

In our model, we modify the above HO procedure to leverage the caching capabilities of MUEs during mobility. Here, we let each MUE u dynamically determine T_s, depending on the cache size Ω_u, the video play rate Q, and the MUE's speed v_u. That is, an MUE u is capable of muting the cell search while Ω_u/Q is greater than ΔT, which enables it to have ΔT seconds to search for a target SBS before the cached content runs out.

Next, we consider the HOF as one of the key performance metrics for any HO procedure. One of the main reasons for the potential increase in HOF in HetNets is due to the relatively small cell sizes, compared to MBS coverage. In fact, HOF is typical if the time-of-stay (ToS) for an MUE is less than the minimum ToS (MTS) required for performing a successful HO. That is,

$$\gamma_{\text{HOF}}(u,k) = \begin{cases} 1, & \text{if } t_{u,k} < t_{\text{MTS}}, \\ 0, & \text{otherwise}, \end{cases} \tag{5.5}$$

where $t_{u,k}$ is the ToS for MUE u to pass across SBS k coverage. Although a short ToS may not be the only cause for HOFs, it becomes very critical within an ultra dense small cell network that encompasses MUEs moving at high speeds [139].

To search the μW carrier for synchronization signals and decode the broadcast channel (system information) of the detected SBSs, the MUEs have to spend an energy E^s per each cell search [19]. Hence, the total energy consumed by an MUE for cell search during time T will be

$$E^s_{\text{total}} = E^s \frac{T}{T_s}. \tag{5.6}$$

Note that increasing T_s reduces E^s_{total} which is desirable. However, less frequent scans will be equivalent to less HOs to SBSs. Therefore, there is a tradeoff between reducing the consumed power for cell search and maximizing traffic offloads from the MBS to SBSs. Content caching will allow increasing T_s, while maintaining traffic offloads from the MBS.

Next, we propose a geometric framework to analyze the caching opportunities, in terms of the caching duration t^c, and the average achievable rate $\bar{R}^c$, for MUEs moving at random directions in joint mmW-μW HetNets.

5.3 Analysis of Mobility Management with Caching Capabilities

In this section, we first investigate the probability of serving an arbitrary MUE over mmW frequencies by a dual-mode SBS.

5.3.1 Probability of mmW coverage

In Fig. 5.2, the small circle represents the intersection of an MUE u's trajectory with the coverage area of an SBS k. In this regard, $\mathbb{P}^c_k(N_k, \theta_k)$ represents the probability that MUE u with a random direction θ_u and speed v_u crosses the mmW coverage areas of SBS k. From Fig. 5.2, we observe that the MUE will pass through the area within mmW coverage only if the MUE's direction is inside the angle $\widehat{AoB}$. Hence, we can state the following.

Theorem 4. *If an SBS k has formed a mmW beam pattern with $N_k \geq 2$ main lobes, each with a beamwidth $\theta_k > 0$, the probability of content caching will be given by:*

$$\mathbb{P}^c_k(N_k, \theta_k) = \left[\frac{N_k\theta_k}{2\pi}\right] + \left[1 - \frac{N_k\theta_k}{2\pi}\right]\left[\frac{1}{2}\left(1 - \frac{1}{N_k}\right) + \frac{\theta_k}{4\pi}\right]. \tag{5.7}$$

Proof. See Appendix C.1. □

We can verify (5.7) by considering an example scenario with $N_k = 3$ and $\theta_k = \frac{2\pi}{3}$. For this example, (5.7) results in $\mathbb{P}^c_k(N_k, \theta_k) = 1$ which correctly captures the fact that the entire cell is covered by mmW beams.

5.3.2 Cumulative distribution function of the caching duration

To enable an MUE to use the cached content while not being associated to an SBS, it is critical to analyze the distribution of caching duration t^c for an arbitrary MUE with a random direction and speed. In this regard, consider the small green triangle in Fig. 5.2, which represents the location of an arbitrary MUE u, $\boldsymbol{x_u} = (x_u, y_u) \in \mathbb{R}^2$, crossing a mmW beam. First, we note that the geometry of the mmW beam of any given SBS can be defined by the location of the SBS, as well as the sides of the beam angle. Without loss of generality, we assume that the SBS of interest is located at the center, such that $\boldsymbol{x}_k = (0, 0)$. Therefore, the two sides of the beam angle will be given by

$$y = x\tan(\theta_0 - \theta_k), y = x\tan(\theta_0), \quad x > 0. \tag{5.8}$$

Assuming that the MUE u is currently located on the angle side $x = y\cos(\theta_0 - \theta_k)$, as shown by the small triangle in Fig. 5.2, then θ_0 in (5.8) will be $\theta_0 = \arccos\left(\frac{x_u}{r_{u,k}(\boldsymbol{x_u})}\right) + \theta_k$, where $r_{u,k}(\boldsymbol{x}) = \sqrt{x_u^2 + y_u^2}$. Hereinafter, we will use the parameter θ_0 to simplify our analysis. Let $F_{t^c}(.)$ be the cumulative distribution function (CDF) of the caching duration t^c. Thus,

$$F_{t^c_u}(t_0) = \mathbb{P}(t^c_u \leq t_0) = \mathbb{P}(r^c_u \leq v_u t_0), \tag{5.9}$$

where r^c_u is the distance that MUE u will traverse across the mmW beam, as shown in Fig. 5.2. Given the location of MUE $\boldsymbol{x}_u$, the minimum possible distance to traverse, r_u^{min}, is

$$r_u^{\text{min}} = \frac{\left|x_u \tan\theta_0 - y_u\right|}{\sqrt{1 + \tan^2\theta_0}}. \tag{5.10}$$

In fact, (5.10) gives the distance of the point $\boldsymbol{x}_u$ from the beam angle side $y = x\tan(\theta_0)$. If $r_u^{\text{min}} > v_u t_0$, then $F_{t^c_u}(t_0) = 0$. Therefore, for the remainder of this analysis we consider $r_u^{\text{min}} \leq v_u t_0$. Next, let $\boldsymbol{x}'_u$ denote the intersection of the MUE's path with line $y = x\tan(\theta_0)$. It is easy to see that $\boldsymbol{x}'_u = (x_u + r^c_u \cos\theta_u, y_u + r^c_u \sin\theta_u)$. Hence, $y_u + r^c_u \sin\theta_u = [x_u + r^c_u \cos\theta_u]\tan\theta_0$, and r^c_u, i.e., the distance that MUE u traverses during the caching duration t^c, is given by:

$$r^c_u = v_u t^c_u = \frac{y_u - x_u \tan\theta_0}{\tan\theta_0 \cos\theta_u - \sin\theta_u}. \tag{5.11}$$

Next, from (5.9) and (5.11), the CDF can be written as

$$F_{t^c_u}(t_0) = \mathbb{P}\left(\frac{y_u - x_u \tan\theta_0}{\tan\theta_0 \cos\theta_u - \sin\theta_u} \leq v_u t_0\right). \tag{5.12}$$

Using the geometry shown in Fig. 5.2, we find the CDF of the caching duration as follows:

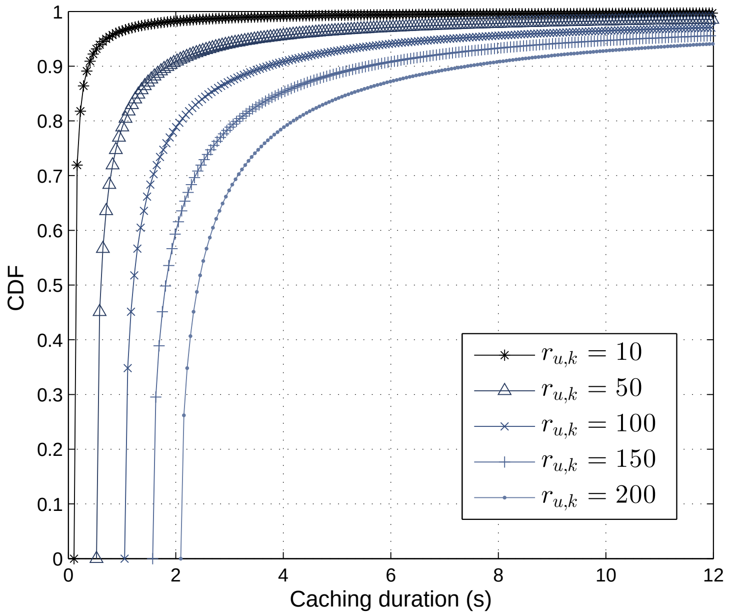


Figure 5.3: CDF of caching duration t^c.

Lemma 1. *The CDF of the caching duration, t^c, for an arbitrary MUE u with speed v_u is given by*

$$F_{t^c}(t_0) = \frac{1}{\pi - \theta_k}\left(\arccos\left(\frac{r_u^{min}}{v_u t_0}\right) + \min\left\{ \arccos\left(\frac{r_u^{min}}{r_{u,k}(\boldsymbol{x})}\right), \arccos\left(\frac{r_u^{min}}{v_u t_0}\right) \right\}\right). \quad (5.13)$$

Proof. See Appendix C.2. □

The CDF of t^c is shown in Fig. 5.3 for different MUE distances from the serving SBS. Fig. 5.3 shows that as the MUE is closer to the SBS, t^c takes smaller values with higher probability which is expected, since the MUE will traverse a shorter distance to cross the mmW beam.

5.4 Performance Analysis of the Proposed Cache-enabled Mobility Management Scheme

Next, we analyze the average achievable rate for content caching, for an MUE with speed v_u, direction θ_u, and initial distance $r_{u,k}(\boldsymbol{x})$ from the serving dual-mode SBS. In addition, we evaluate the impact of caching on mobility management. For this analysis, we ignore the shadowing effect and only consider distance path loss.

5.4.1 Average achievable rate for caching

The achievable rate of caching is given by:

$$R^c(u,k) = \frac{1}{v_u t_u^c} \int_{r_{u,k}(\boldsymbol{x})}^{r_{u,k}(\boldsymbol{x}')} w \log\left(1 + \frac{\beta P_t \psi r_{u,k}^{-\alpha}}{wN_0}\right) dr_{u,k}, \tag{5.14}$$

where $\beta = (\frac{\lambda}{4\pi r_0})^2 r_0^\alpha$. The integral in (5.14) is taken over the line with length r_u^c that connects the MUE location $\boldsymbol{x}$ to $\boldsymbol{x}'$, as shown in Fig. 5.2. With this in mind, we can find the average achievable rate of caching $\bar{R}^c$ as follows.

Theorem 5. *The average achievable rate for an MUE u served by an SBS k, $\bar{R}^c(u,k)$, is:*

$$\bar{R}^c(u,k) = \mathbb{P}_k^c(N_k,\theta_k) R^c(u,k), \tag{5.15}$$

$$= \delta_2 \int_{f(\theta_k)}^{f(0)} \frac{1}{f^2(\theta)} \log\left(1 + \delta_1 f^\alpha(\theta)\right) df(\theta), \tag{5.16}$$

$$\begin{aligned} &\overset{(a)}{=} \frac{\delta_2}{\ln(2)}\Bigg[2\sqrt{\delta_1}\arctan(\sqrt{\delta_1}f(\theta_k)) - \frac{\ln(\delta_1 f^2(\theta_k)+1)}{f(\theta_k)}, \\ &-2\sqrt{\delta_1}\arctan(\sqrt{\delta_1}f(0)) + \frac{\ln(\delta_1 f^2(0)+1)}{f(0)}\Bigg], \end{aligned} \tag{5.17}$$

where $\delta_1 = \frac{\beta P_t \psi}{wN_0}\Big[r_{u,k}(\boldsymbol{x})\sin\hat{\theta}\Big]^{-\alpha}$. Moreover, $\delta_2 = w r_{u,k}(\boldsymbol{x})\sin\hat{\theta}\mathbb{P}_k^c(N_k,\theta_k)/v_u t^c$, and $\hat{\theta} = \theta_u - \theta_0 + \theta_k$. For (a) to hold, we set $\alpha = 2$ which is a typical value for the path loss exponent of LoS mmW links [10].

Proof. See Appendix C.3. □

5.4.2 Achievable gains of caching for mobility management

From (5.3), (5.4), and (5.17), we can find $d^c(u,k)$ which is the distance that MUE u can traverse, while using the cached video content. On the other hand, by having the average inter-cell distances in a HetNet, we can approximate the number of SBSs that an MUE can pass over distance $d^c(u,k)$. Hence, the average number of SBSs that MUE is able to traverse without performing cell search for HO is

$$\eta \approx \left\lfloor \frac{\mathbb{E}\left[d^c(u,k)\right]}{l} \right\rfloor, \tag{5.18}$$

where the expected value is used, since $d^c(u,k)$ is a random variable that depends on θ_u. Moreover, l denotes the average inter-cell distance. Here, we note that

$$\mathbb{E}\left[d^c(u,k)\right] = \int_0^\infty \left(1 - F_{t_u^c}(v_u t)\right) dt, \tag{5.19}$$

where $F_{t^c}(.)$ is derived in Lemma 1. We note that (5.19) is the direct result of writing an expected value in terms of CDF. Based on the definition of η in (5.18) and considering that the inter-frequency energy consumption linearly scales with the number of scans, we can make the following observation.

Remark 1. *The proposed caching scheme will reduce the average energy consumption E^s for inter-frequency cell search by a factor of* $1/\eta$ *with* η *being defined in* (5.18).

Furthermore, from the definition of γ_{HOF} in (5.5), we can define the probability of HOF as $\mathbb{P}(D_{u,k} < v_u t_{\text{MTS}})$ [140], where $D_{u,k} = t_{u,k}/v_u$, and $t_{u,k}$ is the ToS. To compute the HOF probability, we use the probability density function (PDF) of a random chord length within a circle with radius a, as follows:

$$f_D(D) = \frac{2}{\pi\sqrt{4a^2 - D^2}}, \tag{5.20}$$

where (5.20) relies on the assumption that one side of the chord is fixed and the other side is determined by choosing a random $\theta \in [0, \pi]$. This assumption is in line with our analysis as shown in Fig. 5.2. Using (5.20), we can find the probability of HOF as follows:

$$\mathbb{P}(D_{u,k} < v_u t_{\text{MTS}}) = \int_0^{v_u t_{\text{MTS}}} \frac{2}{\pi\sqrt{4a_k^2 - D^2}} dD = \frac{2}{\pi} \arcsin\left(\frac{v_u t_{\text{MTS}}}{2a_k}\right). \tag{5.21}$$

In fact, γ_{HOF} is a binomial random variable whose probability of success depends on the MUE's speed, cell radius, and t_{MTS}. Hence, by reducing the number of HOs by a factor of $1/\eta$, the proposed scheme will reduce the expected value of the sum $\sum \gamma_{\text{HOF}}$, taken over all SBSs that an MUE visits during the considered time T.

Thus far, the provided analysis are focused on studying the caching opportunities for the mobility management in single-MUE scenarios. However, in practice, the SBSs can only serve a limited number of MUEs simultaneously. Therefore, an HO decision for an MUE is affected by the decision of the other MUEs. In this regard, we propose a cache-enabled mobility management framework to capture the inter-dependency of HO decisions in dynamic multi-MUE scenarios.

5.5 Dynamic Matching for Cache-enabled Mobility Management

Within the proposed mobility management scenarios, the MUEs have a flexibility to perform either a vertical or horizontal HO, while moving to their chosen target cell. Additionally, as elaborated in Section 5.4, caching enables MUEs to skip a certain HO, depending on the cache size Ω. In fact, there are three HO actions possible for an arbitrary MUE that is being served by an SBS: 1) Execute an HO for a new assignment with a target SBS, 2) Use the cached content and mute HO,

3) Perform an HO to the MBS. Similarly, an MUE assigned to the MBS can decide whether to handover to an SBS, use cached content, or stay connected to the MBS.

Our next goal is to maximize possible handovers to the SBSs in order to increase the traffic offload from the MBS, subject to constraints on the HOF, SBSs' quota, and limited cache sizes. With this in mind, our goal is to find an HO policy $\boldsymbol{\zeta}$ for MUEs and target BSs[2] that satisfies:

$$\underset{\boldsymbol{\zeta}}{\operatorname{argmin}} \sum_{u\in\mathcal{U}} \zeta(u,k_0), \tag{5.22a}$$

$$\text{s.t.}\quad \mathbb{P}\left(\sum_{k\in\mathcal{K}} \zeta(u,k) D_{u,k} < v_u t_{\text{MTS}}\right) \leq P_u^{\text{th}}, \tag{5.22b}$$

$$\left[1 - \sum_{k\in\mathcal{K}'} \zeta(u,k)\right] T_s \leq \frac{\Omega_u}{Q}, \tag{5.22c}$$

$$\sum_{k\in\mathcal{K}'} \zeta(u,k) \leq 1, \tag{5.22d}$$

$$\sum_{u\in\mathcal{U}} \zeta(u,k) \leq U_k^{\text{th}}, \qquad \forall k \in \mathcal{K}, \tag{5.22e}$$

$$\zeta(u,k) \in \{0,1\}, \tag{5.22f}$$

where $\mathcal{K}' = \mathcal{K} \cup \{k_0\}$ and $\boldsymbol{\zeta}$ is a vector of binary elements $\zeta(u,k) \in \{0,1\}$. In fact, a variable $\zeta(u,k) = 1$, if MUE u is chosen to execute an HO to the target cell k, otherwise, $\zeta(u,k) = 0$. Constraints (5.22b)-(5.22f) must hold for all $u \in \mathcal{U}$. In fact, the objective in (5.22a) is to minimize the number of MUEs associated with MBS k_0. (5.22b) ensures that once an MUE u is assigned to an SBS, i.e. $\sum_{k\in\mathcal{K}} \zeta(u,k) = 1$, the probability of HOF must be less than a threshold P_u^{th}, determined based on the QoS requirement of the MUE u's service. Constraint (5.22c) ensures that if an MUE u is not assigned to any SBS nor the MBS, there will be enough cached video segments for the next T_s time duration. Moreover, constraints (5.22d) and (5.22e) indicate, respectively, that each MUE can be assigned to at most one BS and each SBS can serve maximum U_k^{th} MUEs simultaneously.

We note that using (5.21), we can rewrite (5.22b) as $\sum_{k\in\mathcal{K}} \frac{2}{\pi} \arcsin\left(\frac{v_u t_{\text{MTS}}}{2a_k}\right) \zeta(u,k) \leq P_u^{\text{th}}$, which is a linear constraint. Hence, the posed problem in (5.22a)-(5.22f) is an integer linear programming (ILP), and thus, it is NP-hard. Although an approximation algorithm can be employed to solve (5.22a)-(5.22f), centralized algorithms are not scalable and typically introduce latency which is not desired for real-time applications such as streaming for mobile users. Moreover, these solutions will typically rely on the current network instances, such as the location, speed and cache size of the MUEs, and, hence, they fail to capture the dynamics of the system. To show this, we consider two critical scenarios, shown in Fig. 5.4, as follows:

Illustrative Scenario 1: Consider a feasible solution for (5.22a)-(5.22f), where an MUE u is not assigned to the target SBS k and will use the cached content for the next T_s time duration, as

[2] For brevity, if not specified, we refer to a BS as either an SBS $k \in \mathcal{K}$ or the MBS k_0.

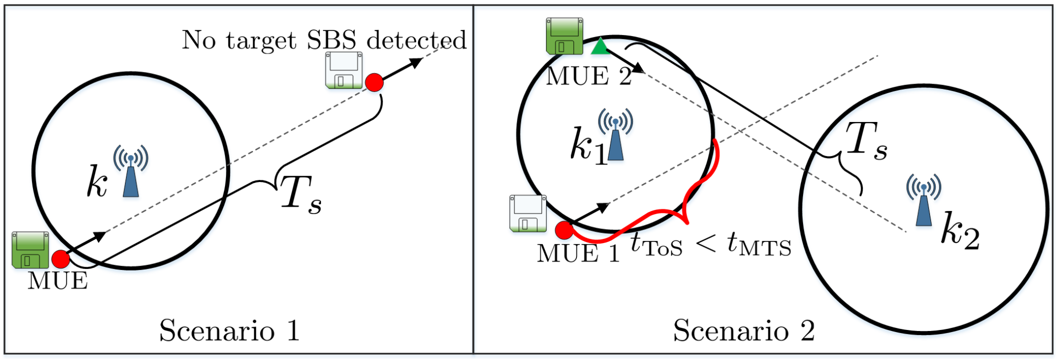


Figure 5.4: Two dynamic HO scenarios for cache-enabled mobile users.

shown in scenario 1 of Fig. 5.4. However, the MUE has to be assigned to the MBS after T_s, since eventually no target SBS is detected. Alternatively, the MUE could be assigned to k initially and fill up the cache, while later, it could use the saved cached content to reach the next target cell without requiring to be assigned to the MBS.

Illustrative Scenario 2: Consider a feasible solution for (5.22a)-(5.22f) which assigns an arbitrary MUE 1, in Fig. 5.4, to a target SBS k_1. If there are not enough cached contents for the MUE 1 to move to the next SBS and at the same time HO fails, the MUE has to be assigned to the MBS as shown in scenario 2 of Fig. 5.4. Alternatively, we could assign MUE 2 with a large cache size to the SBS k_1 such that in case of an HOF, the MUE 2 can reach the next target SBS k_2 by using its available cached contents.

These examples show that taking into account the future network information, such the estimated distance from the next target SBS, is imperative to effectively maximize the traffic offloads from the MBS. Therefore, an efficient HO policy must take into account post-handover scenarios that may occur due to the HOFs. In this regard, we propose a framework based on *dynamic matching theory* [141] which allows effective mobility management, as presented in (5.22a)-(5.22f), while capturing the future network instances, such as the cache size, MUEs' trajectory, and the topology of the network. Next, we explain how the proposed problem can be formulated as a dynamic matching problem.

5.5.1 Handover as a matching game

As proved in previous chapters, matching theory is useful to find polynomial time solutions for combinatorial assignment problems such as (5.22a)-(5.22f) [130]. In a static form, a matching game is defined as a two-sided assignment problem between two disjoint sets of players in which the players of each set are interested to be matched to the players of the other set, according to their preference profiles. A *preference profile* for player i, denoted by $\succ_i$, is defined as a complete, reflexive, and transitive binary relation between the elements of a given set. Within the context of our proposed cache-enabled HO problem, we define the matching problem as follow:

Definition 8. Given the two disjoint sets of MUEs and BSs, respectively, in $\mathcal{U}$ and $\mathcal{K}' = \mathcal{K} \cup \{k_0\}$, a single-period *HO matching* is defined as a many-to-one mapping $\mu : \mathcal{U} \cup \mathcal{K}' \to \mathcal{U} \cup \mathcal{K}'$ that satisfies:

1) $\forall u \in \mathcal{U}$, $\mu(u) \in \mathcal{K}' \cup \{u\}$. In fact, $\mu(u) = k$ means u is assigned to k, and $\mu(u) = u$ indicates that the MUE u is not matched to any BS, and thus, will use the cached content.
2) $\forall k \in \mathcal{K}'$, $\mu(k) \subseteq \mathcal{U} \cup \{k\}$, and $\forall k \in \mathcal{K}$, $|\mu(k)| \leq U_k^{\text{th}}$. In fact, $\mu(k) = k$ implies that no MUE is assigned to the BS k.
3) $\mu(u) = k$, if and only if $u \in \mu(k)$.

Note that, by definition, the matching game satisfies constraints (5.22d)-(5.22f). More interestingly, the matching framework allows defining relevant utility functions per MUE and SBSs, which can capture the preferences of MUEs and SBSs. In this regard, the utility that an arbitrary MUE $u \in \mathcal{U}$ assigns to an SBS $k \in \mathcal{K}$ will be:

$$\varPhi(u,k) = P_u^{\text{th}} - \mathbb{P}\left(\sum_{k\in\mathcal{K}} \zeta(u,k) D_{u,k} < v_u t_{\text{MTS}}\right) = P_u^{\text{th}} - \frac{2}{\pi}\arcsin\left(\frac{v_u t_{\text{MTS}}}{2a_k}\right). \tag{5.23}$$

Here, we observe that the utility in (5.23) is larger for SBSs having a larger cell radius a_k. In addition, as the speed of the MUEs increases, the utility generated from those MUEs being assigned to an SBS decreases. Meanwhile, the utility that an SBS k assigns to an MUE u is given by

$$\varGamma(u,k) = T_s - \frac{\varOmega_u}{Q}. \tag{5.24}$$

In fact, an SBS assigns higher utility to MUEs that are not capable of using caching for the next time duration T_s. Based on the defined utility functions, the preference profile of an arbitrary MUE u, $\succ_u$, will be:

$$k \succ_u k' \;\Leftrightarrow\; \varPhi(u,k) > \varPhi(u,k'), \tag{5.25a}$$

$$u \succ_u k \;\Leftrightarrow\; \varPhi(u,k) < 0, \tag{5.25b}$$

where $k \succ_u k'$ implies that SBS k is strictly more preferred than SBS k' by MUE u. Moreover, $u \succ_u k$ means that an SBS k is not acceptable to an MUE u, if and only if the assigned utility is negative. In fact, (5.25b) known as *an individual rationality constraint* and is in line with satisfying the feasibility condition in (5.22b). Similarly, we can define the preference profile of an SBS k, $\succ_k$, as follows

$$u \succ_k u' \;\Leftrightarrow\; \varGamma(u,k) > \varGamma(u',k), \tag{5.26a}$$

$$k \succ_k u \;\Leftrightarrow\; \varGamma(u,k) < 0, \tag{5.26b}$$

where (5.26b) is the individual rationality requirement for SBSs which is equivalent to satisfying the feasibility constraint in (5.22c). With this in mind, the proposed matching game is formally defined as a tuple $\varPi \triangleq (\mathcal{U} \cup \mathcal{K}, \succ_{\boldsymbol{u}}, \succ_{\boldsymbol{k}})$, where $\succ_{\boldsymbol{u}} = \{\succ_u\}_{u\in\mathcal{U}}$ and $\succ_{\boldsymbol{k}} = \{\succ_k\}_{k\in\mathcal{K}}$.

To solve this game, one desirable solution concept is to find a *two-sided stable matching* between the MUEs and SBSs, μ^*, which is defined as follow [90]:

Definition 9. An MUE-SBS pair $(u,k) \notin \mu$ is said to be a *blocking pair* of the matching μ, if and only if $k \succ_u \{\mu(u), u\}$ and $u \succ_k \{\mu(k), k\}$. Matching μ is *stable*, $\mu \equiv \mu^*$, if there is no blocking pair.

Algorithm 5 DA Algorithm for Single-period Association Between MUEs and SBSs

Inputs: $\mathit{\Pi} \triangleq (\mathcal{U} \cup \mathcal{K}, \succ_{\boldsymbol{u}}, \succ_{\boldsymbol{k}})$.

Outputs: Stable matching μ^*.

```
If not already accepted by an SBS, each unmatched MUE $u \in \mathcal{U}$ applies for its most preferred SBS $k \succ_u u$. Remove $k$ from $u$'s preference profile $\succ_u$.
Each SBS $k \in \mathcal{K}$ receives the proposals from the applicants in Step 1, tentatively accepts $U_k^{\text{th}}$ of most preferred MUEs from new applicants and the MUEs that are so far accepted in $\mu(k)$, and rejects the rest.
repeat Steps 1 to 2
until Each MUE $u$ is accepted by an SBS, or $u$ is applied for all SBSs $k \succ_u u$.
if $\exists u \in \mathcal{U}, \mu(u) \notin \mathcal{K}$ and $\mathit{\Omega}_u/Q < T_s$, then
    $\mu(u) = u$,
else
    Assign $u$ to the MBS.
end if
```

A two-sided stable association between MUEs and SBSs ensures fairness for the MUEs. That is, if an MUE u envies the association of another MUE u', then u' must be preferred by the SBS $\mu^*(u')$ to u, i.e., the envy of MUE u is not justified.

Remark 2. *For a given single-period HO matching game $\mathit{\Pi}$, the* deferred acceptance (DA) *algorithm [130], presented in Algorithm* 5, *is guaranteed to find a two-sided stable association μ^* between MUEs and SBSs.*

Unfortunately, the DA algorithm is not suitable to capture the dynamics of the system which arise from the mobility of the MUEs. In fact, the preference profiles of the MUEs and SBSs only depend on the current state of the system, such as the location of the MUEs, and the cache sizes. In addition, the DA algorithm cannot guarantee stability, if the preference of the MUEs change after HOFs. Thus, to be able to achieve stability for dynamic settings, such as in Scenarios 1 and 2, we need to incorporate the post-HO scenarios into the matching game, such that no MUE can block the stability even after experiencing an HOF. To this end, we extend the notion of one-stage stability in Algorithm 5 into a *dynamic stability* concept that is suitable for the problem at hand.

5.5.2 Dynamic matching for mobility management in heterogeneous networks

To account for possible scenarios that may occur after HO, we consider a two-stage dynamic matching game that incorporates within the preference profiles, some of the possible scenarios that may face the MUEs and base stations after handover execution. Such a dynamic matching will allow the MUEs to build preference profiles over different *association plans* rather than SBSs. An association plan is defined as a sequence of two matchings for a given MUE or SBS. For example, kk' is an association plan that indicates an MUE will be assigned to the SBS k followed by another HO to SBS k'. In this regard, $k_1k_2 \succ_u k_1'k_2'$ means that MUE u prefers plan k_1k_2 to $k_1'k_2'$. With this in mind, we can modify the one-period matching in Definition 8 to a relation

$\mu^\dagger : \mathcal{U} \cup \mathcal{K}' \to (\mathcal{U} \cup \mathcal{K}')^2$, such that $\mu^\dagger(u) = (\mu_1(u), \mu_2(u))$, where μ_1 and μ_2 are one-period matchings. For example, $\mu^\dagger(u) = (k, u)$ indicates that MUE u will first perform an HO to SBS k, $\mu_1(u) = k$, followed by using the content of the cache after exiting the coverage of SBS k, $\mu_2(u) = u$. Next, we use the following definitions to formally define the stability in dynamic matchings [141]:

Definition 10. An MUE-BS pair (u, k) can *period-1 block* the matching, if any of the following conditions is satisfied: 1) $kk \succ_u \mu^\dagger(u)$ and $uu \succ_k \mu^\dagger(k)$; 2) $ku \succ_u \mu^\dagger(u)$ and $uk \succ_k \mu^\dagger(k)$; 3) $uk \succ_u \mu^\dagger(u)$ and $ku \succ_k \mu^\dagger(k)$; or 4) $uu \succ_u \mu^\dagger(u)$ and $kk \succ_k \mu^\dagger(k)$. A matching is *ex ante stable*, if it cannot be period-1 blocked by any MUE/BS or MUE-BS pair.

In a dynamic matching problem, either the MUEs or the BSs may block the matching, after knowing the outcome of the first matching μ_1. In this regard, we define the notion of period-2 blocking and dynamic stability as follows:

Definition 11. An MUE u can *period-2 block* a matching $\mu^\dagger$ if $(\mu_1(u), u) \succ_u \mu^\dagger(u)$. Similarly, an MUE-BS pair (u, k) can *period-2 block* if any of the following conditions is satisfied: 1) $(\mu_1(u), k) \succ_u \mu^\dagger(u)$ and $(\mu_1(k), u) \succ_k \mu^\dagger(k)$, or 2) $(\mu_1(u), u) \succ_u \mu^\dagger(u)$ and $(\mu_1(k), k) \succ_k \mu^\dagger(k)$. A matching is said to be *dynamically stable*, if it cannot be period-1 or period-2 blocked by any MUE or MUE-BS pair[3].

From Definitions 3 and 4, we can see that, any dynamically stable matching is also an ex ante stable matching. However, ex ante stability does not guarantee dynamic stability. For example, if $\mu^\dagger(u) = (k, u)$ for an MUE u, ex ante stability does not guarantee that the MUE commits to use the cache, if the first handover to SBS k fails. In other words, the MUE may block an ex ante stable matching after the actual outcome of the first matching is known. To help better understand the stability for dynamic matchings, we consider the following simple example.

Example 1. Consider a dynamic matching game $\Pi^\dagger$, composed of MUEs $\mathcal{U} = \{u_1, u_2\}$, MBS k_0, and SBSs $\mathcal{K} = \{k_1, k_2\}$, with $U_k^{\text{th}} = 1$ for $k = k_1, k_2$, as shown in scenario 2 of Fig. 5.4. The preference plans of MUEs, MBS k_0, and SBSs are as follows:

$\succ_{u_1}$: $k_1k_0, \underline{k_1u_1}, u_1k_0, u_1u_1$; $\succ_{u_2}$: $k_1u_2, \underline{u_2k_2}, u_2u_2$;

$\succ_{k_1}$: $\underline{u_1k_1}, u_2k_1, k_1k_1$; $\succ_{k_2}$: $\underline{k_2u_2}, k_2k_2$; $\succ_{k_0}$: $k_0u_1, \underline{k_0k_0}$;

where the preference profiles are sorted in descending order and association plans that are not included do not meet the individual rationality constraint. Here, the underlined matching is one of the possible ex ante stable matchings. However, this matching is not dynamically stable. That is because conditioned to $\mu_1(u) = k_1$, the MUE-MBS pair (u_1, k_0) will period-2 block the matching, since $k_1k_0 \succ_{u_1} k_1u_1$ and $k_0u_1 \succ_{k_0} k_0k_0$. In practice, such a blocking occurs if the MUE experiences an HOF with its first matching to k_1.

Next, we propose an algorithm that finds a dynamically stable solution for the proposed mobility management problem.

[3] In general, a matching is dynamically stable for any time t, if it cannot be period-t blocked by any MUE or MUE-BS pair. Extending the dynamic matching to more than two periods depends on how much information is available for MUEs about the network. In this work, we focus on a two-period matching problem, since it is more tractable and practical.

5.5.3 Dynamically stable matching algorithm for mobility management

To find the dynamically stable solution, we note that the solution must first admit the ex ante stability. Therefore, we propose an algorithm, inspired from [141] that yields an ex ante stable association in the first stage, followed by a simple modification to resolve any possible period-2 blocking cases. For each MUE u, let $\mathcal{P}_u = \cup_{k\in\mathcal{K}}\{kk, uk, ku\}$ be the set of all plans considered by u. The algorithm proceeds as follows:
Stage-1 (Finding an ex ante stable matching):

1. For each MUE $u \in \mathcal{U}$, if $uu \succ_u \kappa$, for all $\kappa \in \mathcal{P}_u$, then u does not send any plan proposal to the BSs. Otherwise, MUE u sends a plan proposal to a BS, according to the most preferred plan κ_u^* as follows. If $\kappa_u^* = kk$, MUE u sends a request for a two-period association to the BS k. If $\kappa_u^* = ku$, the MUE sends an association request to BS k, only for period-1. Similarly, if $\kappa_u^* = uk$, the MUE sends an association request to k only for period-2. The MUE removes κ_u^* from its preference profile for the rest of the procedure.
2. Each SBS $k \in \mathcal{K}$ receives the plan proposals and tentatively accepts the most preferred plans, such that the quota U_k^{th} is not violated at each period. Clearly, any accepted plan κ by SBS k satisfies $\kappa \succ_k kk$.
3. MUEs with rejected plans apply in the next round, based on their next most preferred plan. The first stage of the algorithm converges, once no plan is rejected.

Proposition 3. Stage-1 of the proposed algorithm in Algorithm 6 converges to an ex ante stable association between MUEs and BSs.

Proof. See Appendix C.4. □

To avoid period-2 blockage, we introduce a certain structure to the preference profile of the SBSs as follows. For any SBS for whom the maximum quota of U_k^{th} MUEs are assigned, i.e. $|\mu_2^\dagger(k)| = U_k^{\text{th}}$,

$$\mu^\dagger \succ_k \left(\mu_1^\dagger(k), \tilde{\mu_2}^\dagger(k) \cup \{u\}\right), \tag{5.27}$$

where $\tilde{\mu_2}^\dagger(k)$ is $\mu_2{}^\dagger(k)$ with one associated MUE removed to accommodate a new matching with MUE u. In fact, (5.27) implies that an MUE cannot period-2 block the matching with any SBS k that is associated to U_k^{th} MUEs. In addition,

$$(\mu_1(k_0), u) \succ_{k_0} \mu^\dagger \iff P^{\text{th}}_{\mu_1^\dagger(u)} - \frac{2}{\pi}\arcsin\left(\frac{v_u t_{\text{MTS}}}{2a_{\mu_1^\dagger(u)}}\right) < \epsilon, \tag{5.28}$$

where ϵ is a non-negative scalar. In fact, (5.28) allow MUEs that are assigned to SBSs in period 1, with not small enough HOF probability, to be assigned to the MBS in period 2. Another alternative was to set P^{th} a small value from the start. However, this policy will discourage MUEs to be assigned to SBSs and could increase the load on the MBS. With this in mind, we construct the second stage of the algorithm as follows:

Algorithm 6 Proposed Algorithm for Dynamic Matching Between MUEs and BSs

```
Inputs: Preference plans $\kappa$ for all MUEs, MBS, and SBSs.
Outputs: Dynamically stable matching $\mu^*$.
   Phase 1:
1: For each MUE $u \in \mathcal{U}$, if $uu \succ_u \kappa$, for all $\kappa \in \mathcal{P}_u$, then $u$ does not send any plan proposal to the BSs.
   Otherwise, MUE $u$ sends a plan proposal to a BS, according to the most preferred plan $\kappa_u^*$.
2: Each SBS $k \in \mathcal{K}$ receives the plan proposals and tentatively accepts most preferred plans (also compared
   to plans that are previously accepted), such that the quota $U_k^{\text{th}}$ is not violated at each period. Clearly,
   any accepted plan $\kappa$ by SBS $k$ satisfies $\kappa \succ_k kk$.
3: repeat Steps 1 to 2
4: until No plan is rejected. The yielded ex ante stable matching is denoted by $\mu^\dagger = (\mu_1^\dagger, \mu_2^\dagger)$.
   Phase 2:
5: if $\exists u \in \mathcal{U}, \mu_2^\dagger(u) = u$, then apply DA algorithm in Algorithm 5 to the subset of MUEs with $\mu_2^\dagger(u) = u$
   and the subset of BSs with $|\mu_2^\dagger(k)| < U_k^{\text{th}}$, considering the constraints in (5.27) and (5.28). Return
   yielded matching.
6: else
7:     return $\mu^\dagger$.
8: end if
```

Stage-2 (Remove period-2 blocking pairs): Apply the deferred acceptance algorithm shown in Algorithm 5 to a subset of MUEs with $\mu_2^\dagger(u) = u$, and subset of BSs with $|\mu_2^\dagger(k)| < U_k^{\text{th}}$, while considering the constraints in (5.27) and (5.28). The proposed two-stage algorithm is summarized in Algorithm 6. Reconsidering Example 1, it is easy to follow that Algorithm 6 yields the following solution which is dynamically stable[4]:

$\succ_{u_1}: \underline{k_1k_0}, k_1u_1, u_1k_0, u_1u_1;$ $\succ_{u_2}: k_1u_2, \underline{u_2k_2}, u_2u_2;$

$\succ_{k_1}: \underline{u_1k_1}, u_2k_1, k_1k_1;$ $\succ_{k_2}: \underline{k_2u_2}, k_2k_2;$ $\succ_{k_0}: \underline{k_0u_1}, k_0k_0.$

For the proposed algorithm, we can state the following results:

Theorem 6. *The proposed two-stage algorithm in Algorithm* 6 *is guaranteed to converge to a dynamically stable association between MUEs and BSs.*

Proof. See Appendix C.5. □

To analyze the signaling overhead of the proposed algorithm, we consider the total number of HO requests sent to a target SBS by the MUEs. Additional control signals from the SBSs to MUEs can be managed by using a broadcast channel and do not significantly contribute to the overhead of the proposed scheme. In this regard, consider the worst-case scenario in which the initial cache size is $\varOmega_u = 0$ for all $u \in \mathcal{U}$. Therefore, all MUEs seek to perform an HO to the target SBS k by sending a request for plan $\kappa = ku$ during Stage-1 of the proposed algorithm. The SBS k accepts up to U_k^{th} association plans and rejects the rest. Clearly, if there is one target SBS for the MUEs, the signaling overhead will be $\mathcal{O}(U)$. Otherwise, rejected MUEs will send an HO request to the

[4] Here, we assume that (5.28) holds for u_1. Otherwise, the ex ante stable solution in Example 1 is also dynamically stable, since k_0 will not make a period-2 block pair with u_1.

Table 5.2: Simulation parameters

Notation	Parameter	Value
f_c	Carrier frequency	73 GHz
$P_{t,k}$	Total transmit power of SBSs	$[20, 27, 30]$ dBm
K	Total number of SBSs	50
w	Available Bandwidth	5 GHz
$(\alpha_{\text{LoS}},\alpha_{\text{NLoS}})$	Path loss exponent	(2, 3.5) [10]
d_0	Path loss reference distance	1 m [10]
G_{max}	Antenna main lobe gain	18 dB [14]
G_{min}	Antenna side lobe gain	-2 dB [14]
N_k	Number of mmW beams	3
θ_m, θ_k	beam width	10° [14]
N_0	Noise power spectral density	-174 dBm/Hz
t_{MTS}	Minimum time-of-stay	1s [139]
Q	Play rate	1k segments per second
B	Size of video segments	1 Mbits
$(v_{\text{min}}, v_{\text{max}})$	Minimum and maximum MUE speeds	$(1, 16)$ m/s
E^s	Energy per inter-frequency scan	3 mJ [137]

next target SBS, based on their preference profiles. The maximum signaling overhead occurs for a case when all MUEs have the same preference profile as it introduces the highest competition among MUEs. In this case, the signaling overhead of the proposed algorithm will be $\mathcal{O}(UK)$. In addition, in Section 5.6, we will discuss how caching capabilities will reduce the overhead of the proposed algorithm.

5.6 Simulation Results

For simulations, we consider a HetNet composed of $K = 50$ SBSs distributed uniformly across a circular area with radius 500 meters with a single MBS located at the center and a minimum inter-cell distance of 30 meters. Moreover, the transmit power of SBSs are chosen randomly from the set of powers in $[20, 27, 30]$ dBm. The main parameters are summarized in Table 5.2. In our simulations, we consider the overall transmit-receive antenna gain from an interference link to be random. All statistical results are averaged over a large number of independent runs. Next, we first investigate the gains achievable by the proposed cache-enabled scheme for a single user scenario. Then, we will evaluate the performance of the proposed dynamic matching approach by extending the results for scenarios with multiple MUEs in which SBSs can only serve a limited number of MUEs.

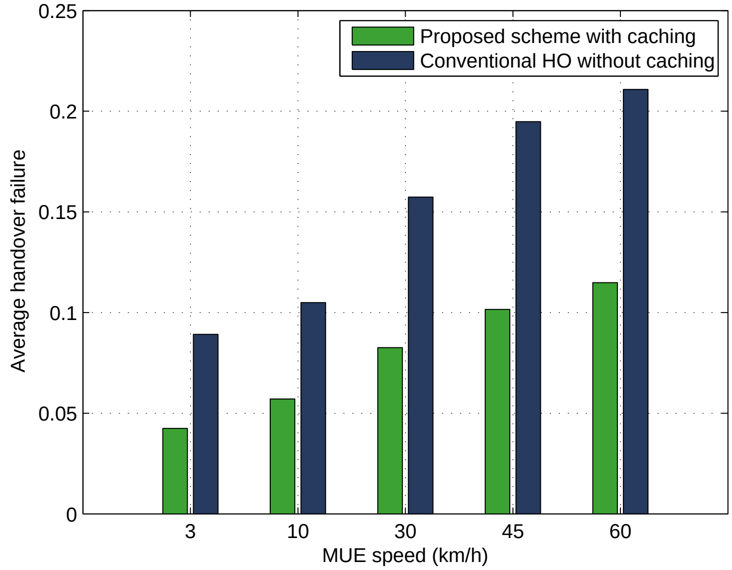


Figure 5.5: HOF vs different MUE speeds.

5.6.1 Analysis of the proposed cache-enabled mobility management for single user scenarios

Fig. 5.5 compares the average HOF of the proposed scheme with a conventional HO mechanism without caching. The results clearly demonstrate that caching capabilities, as proposed here, will significantly improve the HO process for dense HetNets. In fact, the results in Fig. 5.5 show that caching over mmW frequencies will reduce HOF for all speeds, reaching up to 45% for MUEs with $v_u = 60$ km/h.

Fig. 5.6 shows the achievable rate of caching for an MUE with $v_u = 60$ km/h, as a function of different initial distances $r_{u,k}(\boldsymbol{x})$ for various θ_u. The results in Fig. 5.6 show that even for MUEs with high speeds, the achievable rate of caching is significant, exceeding 10 Gbps, for all θ_u values and inital distance of 20 meters from the SBS. However, we can observe that the blockage can noticeably degrade the performance. In fact, for NLoS scenarios, the maximum achievable rate at a distance of 20 meters decreases to 2 Gbps.

5.6.2 Performance of the proposed dynamically stable mobility management algorithm

Here, we consider the set of MUEs entering a target cell coverage region with random directions and speeds. Moreover, the cache sizes of the MUEs are initially $\Omega_u = 10^4$ segments for all MUEs. In addition, each SBS can serve up to $U_k^{\text{th}} = 10$ MUEs. Depending on the speed of the MUE, its direction, and the location of the next target SBS, MUEs form their preferences over different plans as elaborated in Section 5.5.

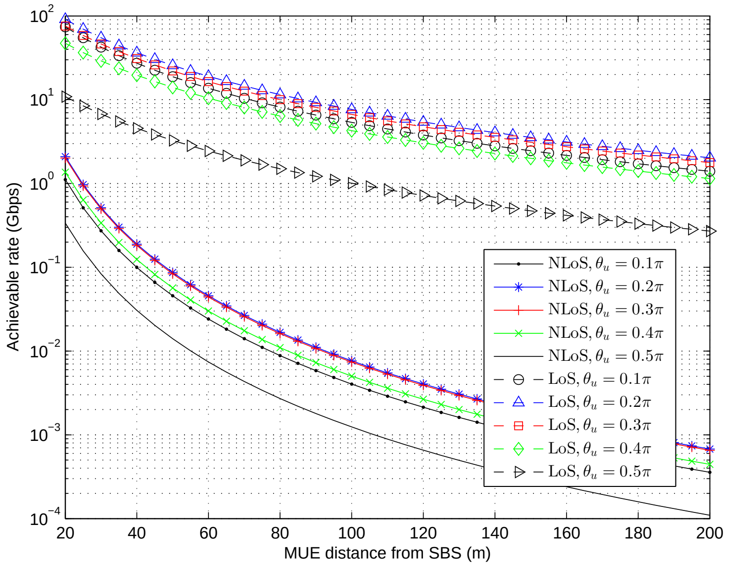


Figure 5.6: Achievable rate of caching vs $r_{u,k}(\boldsymbol{x})$ for different θ_u.

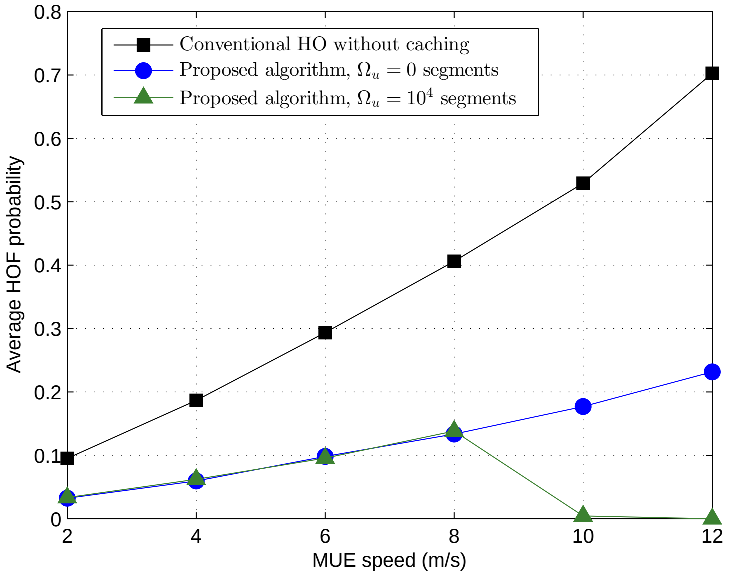


Figure 5.7: Average HOF probability versus MUEs' speeds.

In Fig. 5.7, the average HOF probability of the proposed algorithm is compared with a conventional scheme that does not incorporate caching, versus the speed of the MUEs. The HOF probability is defined as the ratio of the MUEs with HOF to the total number of MUEs, for $U = 20$ and $U_k^{\text{th}} = 10$. The results in Fig. 5.7 show that the HOF probability increases with the speed of the MUEs, since the ToS will decrease for higher MUE speeds. In addition, we observe that the proposed algorithm can significantly reduce the HOF probability by leveraging the information on the MUE's trajectory and the network's topology. Fig. 5.7 also shows that for a non-zero initial

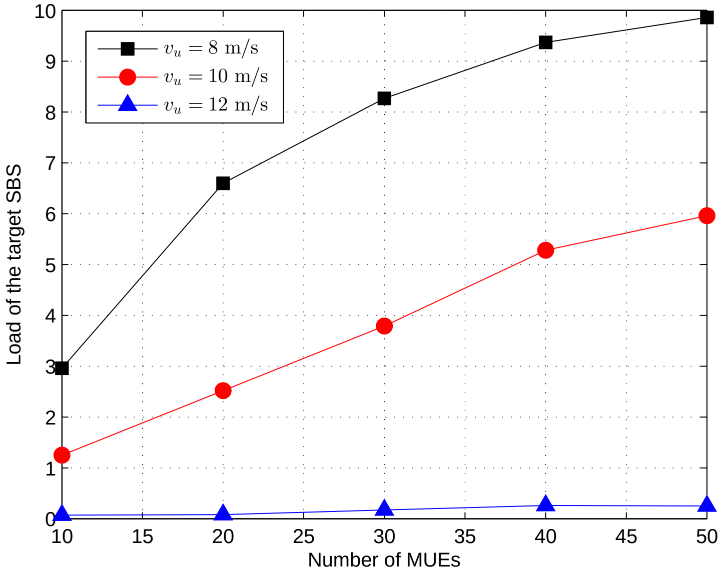


Figure 5.8: Load of the target SBS vs the number of MUEs.

cache sizes of $\Omega_u = 10^4$ segments, the algorithm is considerably robust against HOF. In fact, the HOF probability declines for speeds beyond $v_u = 8$ m/s, since higher speed allows the MUE to traverse larger distance before the cached video segments run out. Therefore, more MUEs will be able to skip an HO to the target cell and use the cached content to move to the next available SBS.

Fig. 5.8 shows the load of the target cell versus the number of MUEs for different MUE speeds $v_u = 8, 10$, and 12 m/s, SBS quota $U_k^{\text{th}} = 10$, and initial cache size $\Omega_u = 10^4$ segments. Here, we observe that the proposed algorithm associates less MUEs to the target cell as the speed increases. That is due to two reasons: 1) higher speeds decrease the ToS and increase the chances of HOFs, and 2) with higher speeds, MUEs can traverse longer distances by using Ω_u cached segments and it is more likely that they can reach to the next target SBS. Fig. 5.8 shows that the load of the target cell reduces up to 45% when v_u increases from 8 to 10 m/s for $U = 40$ MUEs.

In Fig. 5.9, the inter-frequency measurement energy savings yielded by the proposed algorithm are shown as a function of the number of MUEs. Fig. 5.9 shows the total saved energy for MUEs that will use the cached content and do not perform any inter-frequency measurements for handover to an SBS for an initial cache size of $\Omega_u = 10^4$ segments and different MUE speeds. For $U = 50$, MUEs that perform conventional handover without caching will require $UE^s = 150$ mJ total energy for performing inter-frequency measurements. However, the results in Fig. 5.9 show that the proposed scheme achieves up to 80%, 52%, and 29% gains in saving energy, respectively, for MUE speeds $v_u = 8, 10$, and 12 m/s by leveraging cached segments and muting unnecessary cell search. Given that the required energy for measurements linearly scales with the number of MUEs, the results in Fig. 5.9 can also be interpreted as the offloading gains of the proposed approach, compared with conventional HO with no caching. Moreover, these results are consistent with those shown in Fig. 5.8. In fact, as the speed of MUEs increases, the HOF probability increases, and thus, MUEs tend to be assigned to the MBS or use their cached content. In addition,

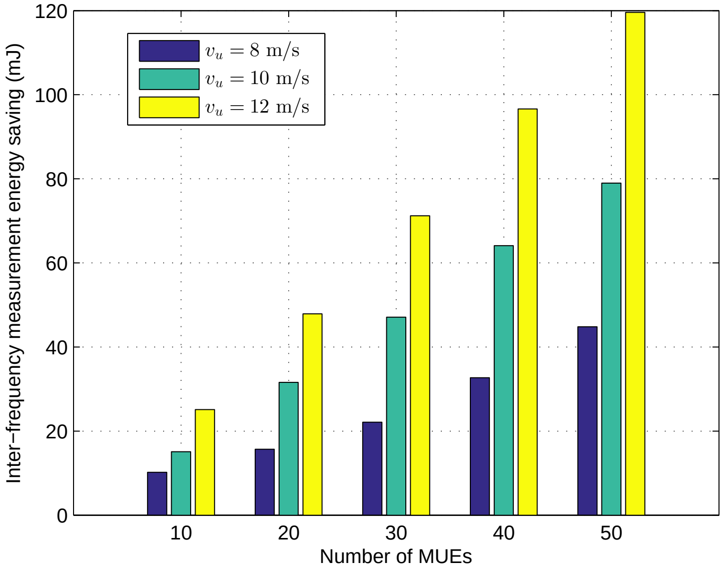


Figure 5.9: Energy savings for inter-frequency measurements vs number of MUEs.

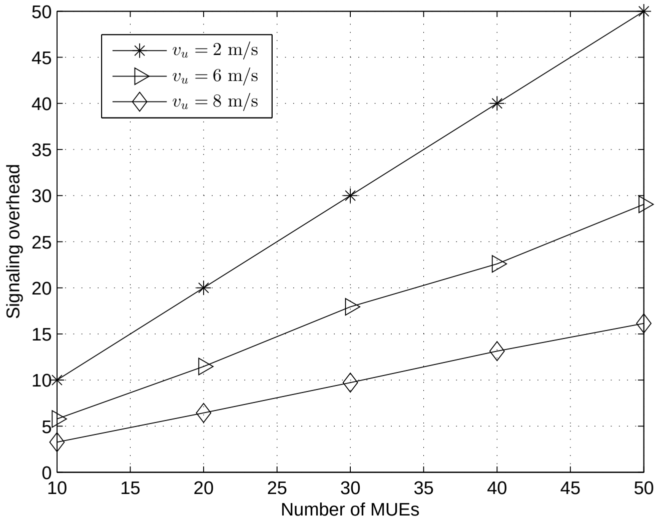


Figure 5.10: Signaling overhead vs number of MUEs.

fast moving MUEs are more likely to reach the next target cell before the cached content runs out.

In Fig. 5.10, we show the signaling overhead resulting from the proposed algorithm versus the number of MUEs, for $\Omega_u = 10^4$ initial cache size and different MUE speeds. Here, we refer to the signaling overhead as the number of HO requests sent to the target SBS by the MUEs. Fig. 5.10 shows that for low speeds $v_u = 2$ m/s, almost all MUEs will attempt to hand over to the target SBS, since the time needed for traversing the SBS coverage is longer than the time available by using the cached content. Nonetheless, the results in Fig. 5.10 clearly demonstrate that the proposed algorithm has a manageable overhead, not exceeding 17 requesting signals for a network size of

$U = 50$ with $v_u = 8$ m/s. In fact, it is interesting to note that although mobility management is, in general, more challenging for high speed MUEs, the overhead of the proposed algorithm decreases for high speed scenarios. This is due to the fact that high speed MUEs use the cached content more effectively than slow-moving MUEs.

5.7 Summary

In this chapter, we have proposed a comprehensive framework for mobility management in integrated microwave-millimeter wave cellular networks. In particular, we have shown that by smartly caching video contents while exploiting the dual-mode nature of the network's base stations, one can provide seamless mobility to the users. We have derived various fundamental results on the probability and the achievable rate for caching video contents by leveraging millimeter wave high capacity transmissions. In addition, to capture the dynamics of the mobility management, we have formulated the multi-user handover problem as a dynamic matching game between the mobile users and small base stations. To solve this game, we have proposed a novel algorithm that is guaranteed to converge to a dynamically stable handover mechanism. Moreover, we have shown that the proposed cache-enabled mobility management framework provides significant gains in reducing the number of handovers, energy consumption for inter-frequency scanning, as well as mitigating the handover failure. Numerical results have corroborated our analytical results and showed that the significant rates for caching can be achieved over the mmW frequencies, even for fast mobile users. In addition, the results have shown that the proposed approach substantially decreases the handover failures and provides significant energy savings in heterogeneous networks.

5.8 Appendix C

C.1 Proof of Theorem 4

Due to the equidistant beams, we have

$$\widehat{AoB} = \frac{1}{2}\widehat{AB} = \frac{1}{2}\left[2\pi - \widehat{AoB}\right] = \frac{1}{2}\left[2\pi - \left(\frac{2\pi}{N_k} - \theta_k\right)\right] = \left(1 - \frac{1}{N_k}\right)\pi + \frac{\theta_k}{2}. \tag{5.29}$$

Given that an arbitrary MUE can enter the circle in Fig. 5.2 at any direction, this MUE will be instantly covered by mmW with probability $\mathbb{P}(\boldsymbol{x}_u \in \mathcal{A}) = \frac{N_k\theta_k}{2\pi}$, where $\mathcal{A} \subset \mathbb{R}^2$ denotes the part of circle's perimeter that overlaps with mmW beams. Therefore,

$$\mathbb{P}_k^c(N_k, \theta_k) = \mathbb{P}(\boldsymbol{x}_u \in \mathcal{A}) + \left[1 - \mathbb{P}(\boldsymbol{x}_u \in \mathcal{A})\right]\frac{1}{2\pi}\widehat{AoB}, \tag{5.30}$$

where (5.30) results from the fact that $\theta_u \sim U\,[0, 2\pi]$. Therefore, from (5.29) and (5.30), the probability of crossing a mmW beam follows (5.7).

C.2 Proof of Lemma 1

From (5.9), $F_{t^c}(t_0) = \mathbb{P}(r_u^c \leq v_u t_0)$. To find this probability, we note that $r_u^c \leq v_u t_0$ if MUE moves between two line segments of length $v_u t_0$ that connect MUE to line $y = x\cos\theta_0$. Depending on $r_{u,k}(\boldsymbol{x})$, the intersection of line segment with $y = x\cos\theta_0$ may have one or two solutions. In case of two intersection points, the two line segments will make two equal angles with the perpendicular line from $\boldsymbol{x}_u$, to $y = x\cos\theta_0$, which each is obviously equal to $\pi-(\pi/2-\theta_k)-\hat{\theta} = \pi/2+\theta_k-\hat{\theta} = \arccos\left(\frac{r_u^{\min}}{v_u t_0}\right)$. Therefore,

$$F_{t^c}(t_0) = \frac{2}{\pi-\theta_k}\arccos\left(\frac{r_u^{\min}}{v_u t_0}\right). \tag{5.31}$$

In fact, θ_u must be within a range of $\pi-\theta_k$ for $r_u^c \leq v_u t_0$ to be valid. Now, if this angle is greater than $\pi/2-\theta_k$, only one intersection point exists. Equivalently,

$$F_{t^c}(t_0)=\frac{1}{\pi-\theta_k}\left(\arccos\left(\frac{r_u^{\min}}{v_u t_0}\right)+\arccos\left(\frac{r_u^{\min}}{r_{u,k}(\boldsymbol{x})}\right)\right). \tag{5.32}$$

Integrating (5.31) and (5.32), the CDF for caching duration can be written as (5.13).

C.3 Proof of Theorem 5

Theorem 4 implies that with probability $1-\mathbb{P}_k^c(N_k,\theta_k)$, only μW coverage is available for an MUE. Therefore, the average achievable rate for caching over the mmW frequencies is given by (5.15). To simplify (5.15), we have

$$r_{u,k}\cos\theta = r_{u,k}(\boldsymbol{x}) + r_u\cos\hat{\theta},\;\; r_{u,k}\sin\theta = r_u\sin\hat{\theta}, \tag{5.33}$$

where $\hat{\theta} = \theta_u-\theta_0+\theta_k$ and θ is an angle between the line connecting MUE to SBS, ranging from 0 to θ_k. Moreover, r_u is the current traversed distance, with $r_u = r_u^c$ once the MUE reaches $\boldsymbol{x}'$ by the end of caching duration, as shown in Fig. 5.2. From (5.33), we find $r_{u,k} = r_{u,k}(\boldsymbol{x})\sin\hat{\theta}/\sin(\hat{\theta}-\theta)$. By changing the integral variable r_u to θ, we can write (5.15) as

$$\bar{R}^c(u,k) = \delta_2\int_0^{\theta_k}\log\left(1+\delta_1\sin^{\alpha}(\hat{\theta}-\theta)\right)\frac{\cos(\hat{\theta}-\theta)}{\sin^2(\hat{\theta}-\theta)}d\theta, \tag{5.34}$$

where $\delta_1 = \beta P_t\psi(r_{u,k}(\boldsymbol{x})\sin\hat{\theta})^{-\alpha}/wN_0$ and $\delta_2 = wr_{u,k}(\boldsymbol{x})\sin\hat{\theta}\mathbb{P}_k^c(N_k,\theta_k)/v_u t^c$. Next, we can directly conclude (5.16) from (5.34) by substituting $f(\theta) = \sin(\hat{\theta}-\theta)$ in (5.34). For $\alpha = 2$, which is a typical value for the path loss exponent for LoS mmW links, (5.16) can be simplified into (5.17) by taking the integration by parts in (5.16).

C.4 Proof of Proposition 3

Assume an MUE-BS pair (u, k) period-1 blocks the matching $\mu^{\dagger}$. In consequence, there is a plan $\kappa \in \{ku, uk, kk\}$ for u and a corresponding plan for k that both prefer to their current matching in $\mu^{\dagger}$. If $\kappa \succ_u \mu^{\dagger}(u)$, then the MUE u must have sent a proposal for κ to k prior to its associated plan in $\mu^{\dagger}$. Since κ is not eventually accepted, that means at some point, the SBS k has rejected κ in favor of another plan. Since the matching for SBSs improves at each round, we conclude that κ is less preferred by k compared to $\mu^{\dagger}(k)$. This contradicts the first assumption, thus, such a period-1 blocking pair does not exist and $\mu^{\dagger}$ is ex ante stable.

C.5 Proof of Theorem 6

From Proposition 3, the solution is guaranteed to be ex ante stable. Therefore, MUEs and BSs will not period-1 block the matching. The rest of the proof easily follows the fact that the BSs will not make a period-2 blocking pair with any MUE, due to the constraints in (5.27) and (5.28). In fact, if there is any period-2 blocking pair (u, k), there are four possible cases to consider: 1) $kk \succ_u \mu^*(u)$ and $uu \succ_k \mu^*(k)$, 2) $uk \succ_u \mu^*(u)$ and $ku \succ_k \mu^*(k)$, 3) $uk \succ_u \mu^*(u)$ and $u'u \succ_k \mu^*(k)$, where $u' \neq u$, or 4) $k'k \succ_u \mu^*(u)$ and $u'u \succ_k \mu^*(k)$, where $k' \neq k$ and $u' \neq u$. The first two cases are not possible, since they indicate that (u, k) can period-1 block μ^* which contradicts ex ante stability. Considering the last two cases, since MUE u is not associated with SBS k in period-2, that implies that k has already been assigned to U_k^{th} MUEs. Otherwise, u would be assigned to k during the second stage of Algorithm 6. Hence, due to the constraint (5.27), k will not make a period-2 blocking pair with u. Similarly, the MBS will not make a period-2 blocking pair with any MUE u that is not assigned to the MBS during the second stage. Thus, no period-2 blocking pair exists and the solution μ^* satisfies dynamic stability.

Chapter 6

Cell Association and Load Balancing for Joint Millimeter Wave-Microwave Cellular Networks

6.1 Background, Related Works, and Summary of Contributions

The integration of cellular networks with mmW communication links is a promising solution to meet the high data traffic requirements of tomorrow's wireless services [7, 10, 27, 75, 142, 143]. However, mmW communication is known to be inherently intermittent, due to the susceptibility of its links to signal blockage, due to shadowing by human, buildings, and other obstacles. To this end, mmW base stations (mmW-BSs) must coexist with the conventional microwave base stations (μW-BSs) to provide μW connectivity for users, when a reliable mmW communication is not feasible [142, 143].

Such integrated mmW-μW networks introduce new challenges for cellular resource management. In particular, the association of UEs to the BSs must now account for the presence of two RATs with significantly different propagation environments. In fact, conventional approaches such as maximum signal-to-interference-plus-noise-ratio (max-SINR) and maximum signal strength indicator (max-RSSI) may result in significantly unbalanced load distributions and may not be directly applicable to the multi-RAT setting. That is due to three key reasons: a) mmW links are highly intermittent and have a higher path loss than μW, b) mmW communication is mostly limited by noise rather than interference, and c) more bandwidth is available at mmW band compared to the μW frequency band.

6.1.1 Related works

The problem of cell association with load balancing has been extensively studied in heterogeneous cellular networks [142–147]. The work in [144] studies the performance of the max-SINR cell association for heterogeneous networks (HetNets) with load balancing via cell range expansion (CRE). The authors in [145] propose a cell association approach based on convex optimization to find a load-aware distributed cell association algorithm for HetNets. Moreover, in [146], a game-theoretic approach is adopted for network selection in HetNets, using an evolutionary game approach. For mmW networks, the work in [147] presents a distributed algorithm that yields a fair cell association. A stochastic geometry framework is used in [142] for the decoupled uplink-downlink cell association for traditional macrocells and mmW small cell networks. In addition, the authors in [143] study resource allocation for mmW-μW networks where cell association is decoupled in the uplink for mmW users.

The existing works in [144–147] have focused on μW or mmW networks, separately and in isolation, and thus, they cannot be applied to integrated mmW-μW cellular networks. In addition, the authors in [142] and [143] consider max-RSSI cell association. However, max-RSSI is not a proper association metric for integrated mmW-μW networks, since it does not properly reflect the achievable rate of the users. Indeed, this rate depends on the allocated bandwidth and the interference, which are completely different between mmW and μW.

6.1.2 Summary of contributions

The main contribution of this chapter is to introduce a novel cell association framework with load balancing for integrated mmW-μW cellular networks. First, we show that conventional max-SINR and max-RSSI cell associations can result in significant unbalanced load in mmW and μW networks. Then, we formulate the proposed cell association problem as a *matching game with minimum quota constraints*, in which the BSs can adjust their minimum quota, in terms of the number of UEs they serve, to balance the network's load. For this game, we show that classical matching solutions such as in [90] and [76] cannot be applied. In contrast, to solve our problem, we propose a novel distributed algorithm that allows UEs to submit association requests to either the mmW-BS or μW-BS that maximizes its average achievable rate. To achieve a balanced load, BSs approve UEs' requests such that the quota constraints are met. We show that the proposed algorithm yields a PO and stable solution for the UEs. Simulation results show the effectiveness of our approach in integrated mmW-μW networks.

The rest of this chapter is organized as follows. Section 6.2 presents the problem formulation. Section 6.3 formulates the problem as a matching game. Section 6.4 presents the proposed algorithm. Simulation results are analyzed in Section 6.5. Section 6.6 concludes the chapter.

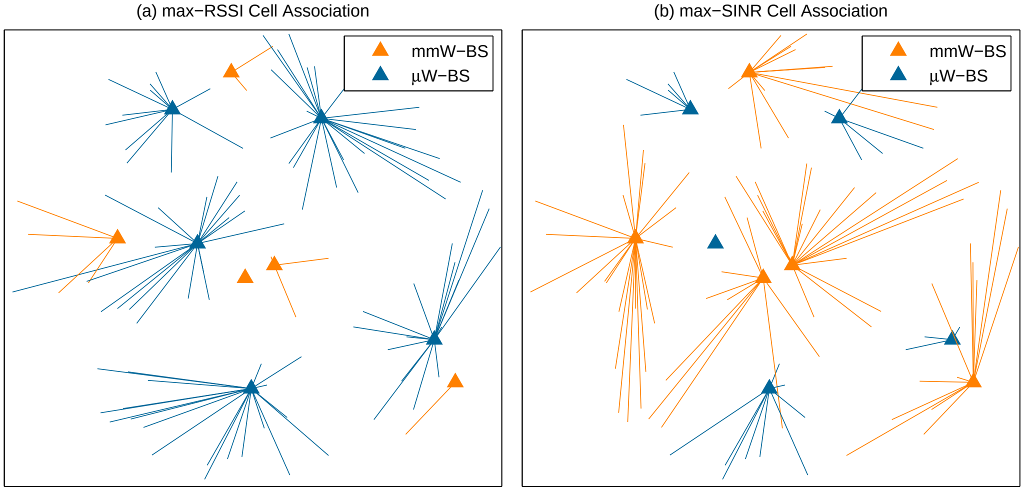


Figure 6.1: Cell association for an integrated μW-mmW network using (a) max-RSSI, and (b) max-SINR approaches. The triangles show the BSs and the orange and blue colors represent, respectively, the mmW and μW links.

6.2 System Model

Consider the downlink of a cellular network, composed of a set $\mathcal{N}_1$ of N_1 mmW-BSs and a set $\mathcal{N}_2$ of N_2 μW-BSs. In this network, a set $\mathcal{M}$ of M UEs are deployed and must be assigned to one mmW-BS or μW-BS. UEs and BSs are distributed uniformly and randomly within a planar area with radius r centered at $(0,0) \in \mathbb{R}^2$. UEs are equipped with both mmW and μW RF interfaces allowing them to manage their traffic at both frequency bands.

6.2.1 Propagation model at mmW and μW frequency bands

Each mmW link between mmW-BS $n \in \mathcal{N}_1$ and UE $m \in \mathcal{M}$, located at $\boldsymbol{y}_n \in \mathbb{R}^2$ and $\boldsymbol{y}_m \in \mathbb{R}^2$, respectively, is characterized by the transmit power p_n, channel gain $g(\boldsymbol{y}_m, \boldsymbol{y}_n)$ and the antenna gain $\psi(\boldsymbol{y}_m, \boldsymbol{y}_n)$. Assuming that the total power p_n is distributed uniformly over the mmW bandwidth, the achievable rate per unit of bandwidth for a UE m assigned to mmW-BS n is given by:

$$c_{m,n}^{\text{mmW}} = \log_2\left(1 + \frac{p_n\psi(\boldsymbol{y}_m, \boldsymbol{y}_n)g(\boldsymbol{y}_m, \boldsymbol{y}_n)}{w_1 N_0}\right), \tag{6.1}$$

where w_1 is the mmW bandwidth, $g(\boldsymbol{y}_m, \boldsymbol{y}_n)$ is the link channel gain, and N_0 is the noise power spectral density. Hereinafter, we represent $c_{m,n}^{\text{mmW}}$ by $c_{m,n}^{\text{LoS}}$ and $c_{m,n}^{\text{NLoS}}$, respectively, if the link is LoS and NLoS. Here, $g(\boldsymbol{y}_m, \boldsymbol{y}_n) = L(\boldsymbol{y}_m, \boldsymbol{y}_n)^{-1}$, where the path loss $L(\boldsymbol{y}_m, \boldsymbol{y}_n)$ in dB follows the model of [10]:

$$L(\boldsymbol{y}_m, \boldsymbol{y}_n) = b_1 + a_1 10 \log_{10}(\|\boldsymbol{y}_m - \boldsymbol{y}_n\|) + \chi, \tag{6.2}$$

where a_1 represents the slope of the best linear fit to the propagation measurement in mmW frequency band and b_1 is the path loss (in dB) for 1 meter of distance. In addition, χ models the deviation in fitting (in dB) which is a Gaussian random variable with zero mean and variance ξ_1^2.

For each UE-BS pair (m, n), let $\zeta_{m,n}$ be a Bernoulli random variable with success probability $\rho_{m,n}$ that indicates whether the mmW link is LoS, $\zeta_{m,n} = 1$, or NLoS, $\zeta_{m,n} = 0$. Different path loss parameters in (6.2) are considered for the LoS and NLoS links, as listed in Table. 6.1.

At μW band, the achievable rate per unit of bandwidth for a UE $m \in \mathcal{M}$ associated with μW-BS $n \in \mathcal{N}_2$ is given by:

$$c_{m,n}^{\mu\text{W}} = \log_2\left(1 + \frac{p_n g(\boldsymbol{y}_m, \boldsymbol{y}_n)}{\sum_{n' \neq n} p_{n'} g(\boldsymbol{y}_m, \boldsymbol{y}_{n'}) + w_2 N_0}\right), \tag{6.3}$$

where the total power p_n is distributed uniformly over the μW bandwidth, w_2, and the channel gain is characterized by parameters, a_2, b_2 and ξ_2, similar to (6.2).

6.2.2 Problem formulation

The cell association problem can be defined as a decision policy π which, for any UE-BS pair (m, n), it outputs a binary variable $x_{m,n} \in \{0, 1\}$, where $x_{m,n} = 1$ indicates that UE m is assigned to BS n, otherwise, $x_{m,n} = 0$. Further, we define the BS n's load, κ_n as

$$\kappa_n = \sum_{m=1}^{M} x_{m,n}. \tag{6.4}$$

Using (6.4), the *maximum load difference* can be defined as the difference of the load for the BSs with the maximum and minimum number of associated UEs, as follow:

$$\varDelta_\kappa(\pi) = \max(\kappa_n) - \min(\kappa_n). \tag{6.5}$$

In (6.5), a smaller $\varDelta_\kappa(\pi)$ implies better load balancing. In general, it is desirable to achieve uniform loads for all BSs, i.e., $\varDelta_\kappa(\pi) = 0$. However, by using conventional cell association approaches, such as max-SINR and max-RSSI schemes [144, 145], as shown in Fig. 6.1, the network will exhibit a severely unbalanced load. In fact, the max-RSSI scheme assigns most of the UEs to μW-BS, due to the smaller path loss over the μW frequency band. On the other hand, the max-SINR scheme assigns most of the UEs to the mmW-BSs, due to the directional transmissions and less interference. In addition, in Figs. 6.4 and 6.5, we show by simulations that the CRE techniques used in small cell networks [144] may not effectively improve load balancing in mmW-μW networks, due to the large gap in the RSSI and SINR values for mmW and μW links. Our joint mmW-μW cell association problem is thus given by:

$$
\begin{aligned}
&\underset{\boldsymbol{x}}{\text{maximize}} \sum_{n=1}^{N}\sum_{m=1}^{M} x_{m,n}U_{m,n}(\boldsymbol{x}), && \text{(6.5a)}\\
\text{s.t.}\quad &\sum_{n\in\mathcal{N}} x_{m,n} \leq 1, \qquad \forall m \in \mathcal{M}, && \text{(6.5b)}\\
&\kappa_n \leq q_n^{\text{max}}, \qquad \forall n \in \mathcal{N}, && \text{(6.5c)}\\
&\kappa_n \geq q_n^{\text{min}}, \qquad \forall n \in \mathcal{N}, && \text{(6.5d)}\\
&x_{m,n} \in \{0,1\}, && \text{(6.5e)}
\end{aligned}
$$

where $\mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_2$ is the set of all $N = N_1 + N_2$ BSs and $U_{m,n}$ denotes the utility of the UE m associated to the BS n. Moreover, q_n^{max} and q_n^{min} denote, respectively, *the maximum and the minimum quotas* for BS n which represent the maximum and the minimum number of UEs that it can serve. We let $0 \leq q_n^{\text{min}} \leq q_n^{\text{max}}$ and $\sum_{n\in\mathcal{N}} q_n^{\text{min}} \leq M \leq \sum_{n\in\mathcal{N}} q_n^{\text{max}}$ to ensure that a feasible solution exists. As we elaborate later in Section IV, constraints (6.5c)-(6.5d) are introduced to balance the network's load. Next, we make the following observation:

Remark 3. *With $q_n^{min} = 0$ and $q_n^{max} = M$ for $\forall n \in \mathcal{N}$, the optimization problem* (6.5a)-(6.5e) *does not incorporate load balancing.*

The cell association for an arbitrary UE depends on the associations of the other UEs, due to the quota constraints (6.5c)-(6.5d). In addition, the utility of a UE may depend on whether the associated BS is a mmW-BS or a μW-BS.

6.3 Cell Association as a Matching Game with Minimum Quotas

The downlink association problem in (6.5a)-(6.5e) is a 0-1 integer programming problem for assigning UEs to BSs which does not admit a closed-form solution and has exponential complexity [89]. In fact, for such a cell association problem, an exhaustive search requires a comparison of $O(N^M)$ assignments, which cannot adapt to the dynamics of dense cellular networks, particularly, when using the mmW frequency band.

In this regard, centralized cell association schemes require the BSs to send the network information to the radio network controller (RNC). Such implementations carried out by the RNC are updated at relatively long timescales. This can be detrimental for the mmW UEs that frequently experience NLoS transmissions. To this end, we propose a distributed solution for the mmW-μW cell association problem.

6.3.1 Cell association as a matching game

To solve the problem in (6.5a)-(6.5e), we propose a novel decentralized solution with tractable complexity based on matching theory concept [76, 90]. In our model, over each cell association time frame, the set of BSs, $\mathcal{N}$, and the set of UEs, $\mathcal{M}$, are the two sets of players of the matching game. A preference relation $\succ$ is defined as a complete, reflexive, and transitive binary relation between the elements of a given set. Here, we let $\succ_m$ be the preference relation of UE m and denote $n \succ_m n'$, if UE m prefers BS n more than n'. Similarly, we use $\succ_n$ to denote the preference relation of BS $n \in \mathcal{N}$.

To define the preference relations, we can introduce individual utility functions for each UE and BS, using which they can rank one another. In the proposed cell association problem, the preference relations of UEs will depend only on the local average achievable rate information, while the BSs will use network-wide information to distribute the loads and maximize the sum utility. In fact, matching-based cell association provides a suitable framework to balance the load by properly adjusting the maximum and minimum BS quotas.

Each cell association policy π determines the allocation of a subset of UEs to each BS. Thus, the problem can be defined as a *one-to-many matching game*:

Definition 12. Given two disjoint finite sets of players $\mathcal{M}$ and $\mathcal{N}$, the cell association policy, π, can be defined as a a one-to-many *matching relation*, $\pi : \mathcal{N} \to \mathcal{M}$ that satisfies 1) $\forall n \in \mathcal{N}, \pi(n) \subseteq \mathcal{M}$, 2) $\forall m \in \mathcal{M}, \pi(m) \in \mathcal{N}$, and 3) $\pi(m) = n$, if and only if $m \in \pi(n)$.

In fact, $\pi(m) = n$ implies that $x_{m,n} = 1$, otherwise $x_{m,n} = 0$. One can easily see from Definition 12 that the proposed matching game inherently satisfies the constraints in (6.5b) and (6.5e). In addition, π is a *feasible matching*, if it satisfies the quota constraints, i.e., $|\pi(m)| \in \{0, 1\}$ and $q_n^{\text{min}} \leq \kappa_n = |\pi(n)| \leq q_n^{\text{max}}$, where $|.|$ denotes the set cardinality. Next, we define suitable utility functions.

6.3.2 Utility and preference relations of the UEs and BSs

For mmW links, a UE may experience multiple LoS/NLoS transmissions with different rates during the time that cell association is not updated. Thus, the utilities of UEs to BSs must be a function of the average rate. Here, we define the utility function of UE m for BS n as:

$$\begin{aligned} U_m(n) = \log \left[f(\boldsymbol{k}_{m,n}) c_{m,n}^{\text{LoS}} + (1 - f(\boldsymbol{k}_{m,n}))\, c_{m,n}^{\text{NLoS}} \right] \mathbb{1}_{n \in \mathcal{N}_1} \\ + \log \left[c_{m,n}^{\mu\text{W}} \right] \mathbb{1}_{n \in \mathcal{N}_2}, \end{aligned} \tag{6.6}$$

where,

$$\mathbb{1}_{n \in \mathcal{N}_i} = \begin{cases} 1 & \text{if } n \in \mathcal{N}_i, \\ 0, & \text{if } n \in \mathcal{N}_{j \neq i}, \end{cases} \tag{6.7}$$

Algorithm 7 Proposed Cell Association Algorithm

```
Inputs: $\succ_{\text{ML}}, \succ_m, \forall m \in \mathcal{M}, q_n^{\max}, q_n^{\min}, \forall n \in \mathcal{N}$.
Outputs: $\pi, \boldsymbol{x}$.
1: Initialize: $\pi(m) = \emptyset, \forall m \in \mathcal{M}, \mathcal{M}' = \mathcal{M}$.
2: Choose the UE $m^* \in \mathcal{M}'$ that has the highest rank in ML profile, i.e., $m^* \succ_{\text{ML}} m, \forall m \in \mathcal{M}'$.
3: Let $\pi(m^*) = n$, where $n$ is the most preferred BS based on $\succ_{m^*}$ with $\kappa_n < q_n^{\max}$. Moreover, add $m^*$ to $\pi(n)$ and
   remove it from $\mathcal{M}'$.
4: repeat Steps 3 to 4
5: until $\sum_{n\in\mathcal{N}} \lfloor q_n^{\min} - \kappa_n \rfloor^+ = |\mathcal{M}'|$.
6: while $\mathcal{M}' \neq \emptyset$ do
7:     Choose the UE $m^* \in \mathcal{M}'$ that has the highest rank in ML profile, i.e., $m^* \succ_{\text{ML}} m, \forall m \in \mathcal{M}'$.
8:     Let $\pi(m^*) = n$, where $n$ is the most preferred BS based on $\succ_{m^*}$ with $\kappa_n < q_n^{\min}$. Add $m^*$ to $\pi(n)$ and remove
   it from $\mathcal{M}'$.
9: end while
```

and $\boldsymbol{k}_{m,n}$ is a vector composed of elements, $k_{m,n}(t')$ where $t' = t-1, t-2, \cdots, 0$, is the number of successful LoS transmissions from mmW-BS n to UE m and $f(k_{m,n}(t))$ is a metric that each UE uses to estimate the LoS probability $\rho_{m,n}$ for cell association at time t. In practice, the UEs can update a moving average of the number of LoS transmissions from each mmW-BS by using:

$$f(\boldsymbol{k}_{m,n}(t)) = \lambda \frac{k_{m,n}(t) x_{m,n}}{k} + (1-\lambda) f(\boldsymbol{k}_{m,n}(t-1)), \tag{6.8}$$

where λ is a constant smoothing factor between 0 and 1 and k is the number of transmission slots within the time window in which the association policy π is not updated. Using the utilities in (6.6), the preference relations of UEs are:

$$n \succ_m n' \Leftrightarrow U_m(n) \geq U_m(n'), \tag{6.9}$$

for $\forall m \in \mathcal{M}$, and $\forall n, n' \in \mathcal{N}$.

We note that assigning UEs to their most preferred BS may not admit a feasible matching in general. In other words, in order to satisfy the minimum quotas of the BSs, some UEs may have to be assigned to a lower ranked BS. Therefore, a suitable mechanism is required at the level of the BSs to determine which UEs must be assigned to the BSs with unsatisfied minimum quotas. To this end, all BSs must use the same preference profile, known as *master list (ML)*, $\succ_{\text{ML}}$ with $\succ_n \equiv \succ_{\text{ML}}, \forall n \in \mathcal{N}$, as follow:

$$m \succ_{\text{ML}} m' \Leftrightarrow U_{\text{ML}}(m) \geq U_{\text{ML}}(m'), \tag{6.10}$$

where,

$$U_{\text{ML}}(m) = \{U_m(n') | U_m(n') \geq U_m(n), \forall n \in \mathcal{N}\}. \tag{6.11}$$

In fact, (6.11) implies that BSs give higher priority to a UE that can achieve higher utility by being assigned to its preferred BS. This allows maximizing the sum utility in (6.5a). To form the ML in practice, BSs only require to exchange the ordering of their nearby UEs to neighboring BSs.

Table 6.1: Simulation parameters

Notation	**Parameter**	**Value**
p_n	Transmit power	30 dBm
(ω_1, ω_2)	Bandwidth	(1 GHz, 10 MHz)
$(\xi_{1,\text{LoS}}, \xi_{1,\text{NLoS}}, \xi_2)$	Standard deviation of path loss	$(5.2, 7.6, 10)$ [142]
$(a_{1,\text{LoS}}, a_{1,\text{NLoS}}, a_2)$	Path loss exponent	(2,4,3) [142]
(b_1, b_2)	Path loss at 1 m	$(70, 38)$ dB
ψ	Antenna gain	18 dBi [142]
N_0	Noise power spectral density	-174 dBm/Hz
M	Number of UEs	From 10 to 100
q_n^{max}	Maximum quota	M

6.4 Proposed Cell Association and Load Balancing Algorithm

To solve the proposed cell association matching problem, we consider two important concepts of *Pareto optimality* and *two-sided stability*. A PO matching is defined as follow [148]:

Definition 13. A cell association policy, π, is *Pareto optimal*, if there is no other feasible matching policy π' such that π' is preferred by all UEs over π, $\pi' \succeq_m \pi$, for all $m \in \mathcal{M}$, and strictly preferred over π, $\pi' \succ_m \pi$, for some UEs $m \in \mathcal{M}$.

In fact, PO is a widely adopted notion of efficiency for distributed mechanisms where each entity, here each UE, aims to maximize its own utility. Furthermore, the concept of two-sided *stable matching* between UEs and BSs is defined as follows [90]:

Definition 14. A UE-BS pair $(m, n) \notin \pi$ is said to be a *blocking pair* of the matching π, if and only if $m \succ_n m'$ for some $m' \in \pi(n)$ and $n \succ_m \pi(m)$. Matching π is *stable*, if there is no blocking pair.

A stable cell association policy ensures fairness for the UEs. That is, if a UE m envies the assignment of another UE m', then m' must be preferred by the BS $\pi(m')$ to m, i.e., the envy of UE m is not justified. When $\succ_n \equiv \succ_{\text{ML}}, \forall n \in \mathcal{N}$, as in our problem, the two-sided stable π is also known as *ML-fair* matching.

For matching problems with no minimum quota, i.e., $q_n^{\text{min}} = 0$, the well-known *deferred acceptance (DA)* algorithm is used to find a stable matching such as in [90], [76], and [91]. However, with minimum quotas, DA is no longer guaranteed to find a feasible solution.

Proposition 4. For cell association problems with minimum quota constraints, the standard DA algorithm may not admit a feasible solution.

Proof. See Appendix D.1. □

Therefore, a new algorithm must be developed to solve the problem. To this end, we propose the matching with minimum quota (MMQ) algorithm shown in Algorithm 7, which is designed based on [148]. The proposed algorithm proceeds as follows. After initialization, in step 2, UE m^* with the highest rank in the ML profile requests a connection with its most preferred BS n. If κ_n is less than its maximum quota, UE m^* will be accepted by BS n. This procedure continues in Steps 3 and 4 for the remaining UEs until the number of UEs is equal to the required number of UEs for meeting the minimum quota constraints, i.e., $\sum_{n\in\mathcal{N}}\lfloor q_n^{\text{min}} - \kappa_n\rfloor^+ = |\mathcal{M}'|$, where $\lfloor x\rfloor^+ = max(x,0)$. Next, in Step 7, the most preferred UE based on the ML profile must be assigned only to its most preferred BS from the subset of $\mathcal{N}$ with $\kappa_n < q_n^{\text{min}}$. In fact, our algorithm allows each UE to be assigned to its most preferred BS, as long as the minimum and maximum quota constraints are not violated. The algorithm terminates once all the UEs are assigned to a BS. The proposed, distributed matching algorithm exhibits the following properties:

Theorem 7. *Algorithm* 7 *is guaranteed to yield a feasible PO and stable matching between UEs and BSs.*

Proof. See Appendix D.2. □

We must note that Pareto optimality and stability cannot be inherently achieved if the BSs do not follow the ML preference profile. In fact, for $\succ_n\neq\succ_{\text{ML}}$, there is no algorithm in general that can guarantee a feasible PO and stable solution [148].

6.5 Simulation Results

For simulations, we consider a network with $N_1 = 10$ mmW-BSs, $N_2 = 10$ μW-BSs, and up to $M = 100$ UEs located uniformly and randomly over an area with diameter $r = 1$ km. The main parameters are summarized in Table 6.1. The average probability of LoS for each mmW BS-UE pair is sampled from a uniform distribution, $\rho_{m,n} \in [0,1]$. All statistical results are averaged over a large number of independent runs.

We compare the performance of the proposed MMQ algorithm with both conventional max-SINR and max-RSSI approaches. We also consider a CRE with bias factor γ_{RSSI} and γ_{SINR}, respectively, for the max-RSSI and max-SINR schemes for further comparisons. To calculate the rates, the total bandwidth at each BS is allocated equally to the associated UEs. That is, $r_{m,n}^{\text{mmW}} = \frac{w_1}{\kappa_n}c_{m,n}^{\text{mmW}}$, where $r_{m,n}^{\text{mmW}}$ denotes the achievable rate for UE m associated with mmW-BS n. Moreover, $r_{m,n}^{\mu\text{W}} = \frac{w_2}{\kappa_n}c_{m,n}^{\mu\text{W}}$, where $r_{m,n}^{\mu\text{W}}$ denotes the achievable rate for UE m assigned to μW-BS n. In [145], it is shown that for logarithmic utilities, as in (6.6), uniform resource allocation maximizes the sum utility.

Fig. 6.2 shows the average sum-rate for the proposed MMQ approach, compared to max-RSSI and max-SINR approaches versus the number of UEs. The bias factors are chosen such that near uniform loads are achieved for all the BSs. The minimum quotas for μW-BSs are chosen randomly

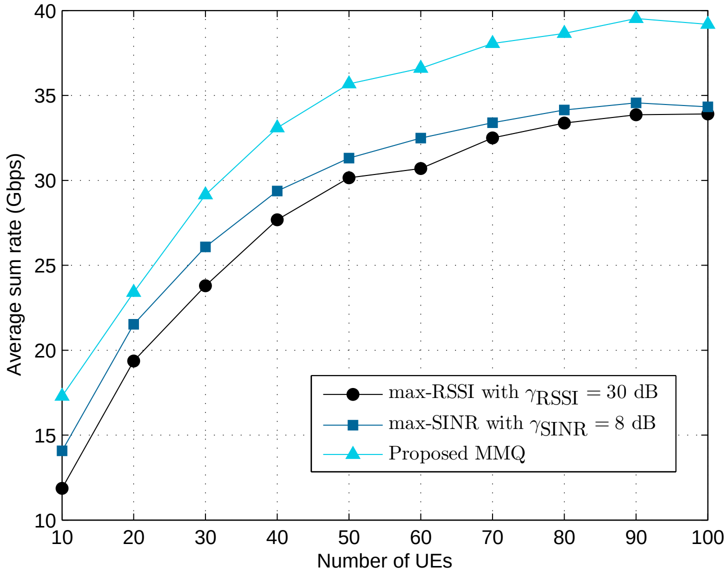


Figure 6.2: The average sum-rate (Gbps) versus the number of UEs M.

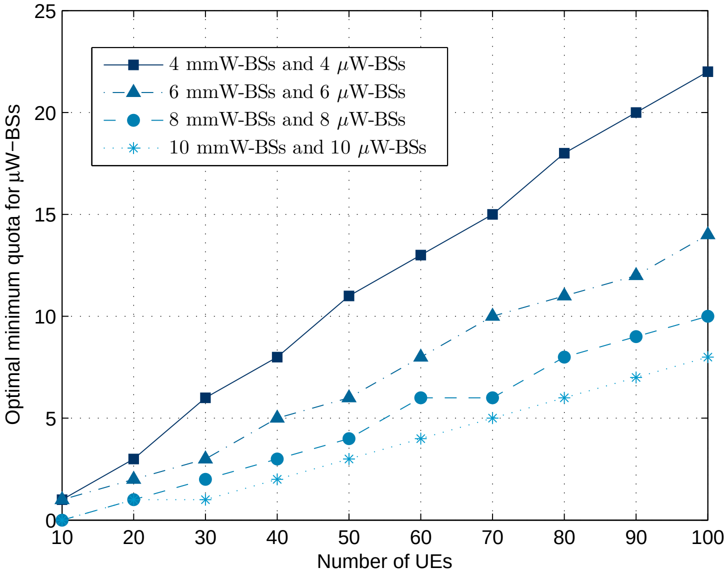


Figure 6.3: The optimal quota values for μW-BSs versus the number of UEs.

from 0 to $\lfloor M/N_2 \rfloor$, with $\lfloor . \rfloor$ denoting the floor operand. The results show that the proposed approach achieves up to 14% and 18% improvements compared to, respectively, the max-SINR and the max-RSSI schemes, for $M = 50$. This is due to the fact that the achievable rate is a nonlinear function of the SINR or RSSI metrics. Hence, average SINR or RSSI , with respect to $\zeta_{m,n}$, cannot be used to find the average achievable rate. However, the proposed approach directly relies on the average achievable rate, as shown in (6.6).

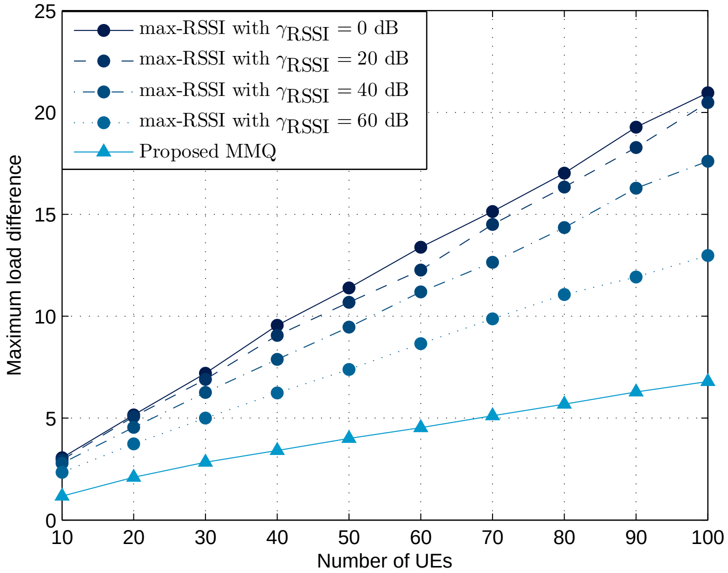


Figure 6.4: The maximum load difference, Δ_κ, for the proposed MMQ approach, compared to the max-RSSI with CRE.

Fig. 6.3 shows the optimal minimum quota for μW-BSs that yields the maximum average sum-rate, as the number of UEs varies, for different values of $N_1 = N_2$. The minimum quota for mmW-BSs is zero, since the load of the mmW-BSs are higher than μW-BSs. The results show that the optimal minimum quota increases, as M increases, since more UEs must be associated with the μW-BSs. Moreover, the optimal q_n^{min} decreases as N_1 and N_2 increase, since more BSs are available. For $M = 100$ UEs, we observe that the optimal minimum quotas are $q_n^{\text{min}} = 8$, for all $n \in \mathcal{N}_2$, which implies that 80% of the UEs must be assigned to the μW-BSs. Hence, if sum rate is considered as the optimality criterion, the result does not yield a balanced network.

In Fig. 6.4, the maximum load difference Δ_κ, is evaluated for the proposed algorithm compared to max-RSSI approach with CRE under biasing values ranging from 0 to 60 dB. The results show that, as biasing increases, the load balancing decreases and then increases. For all biasing values, Δ_κ for the max-RSSI approach is significantly larger than the proposed MMQ algorithm. In fact, we observe that the proposed MMQ algorithm substantially improves the load balancing, reaching up to 48% compared to the max-RSSI with $\gamma_{\text{RSSI}} = 40$ dB for $M = 70$. This improvement is due to the fact that the CRE with biasing cannot precisely control the number of UEs re-associated from μW-BSs to the mmW-BSs. However, in the proposed approach, the BSs can directly control the number of associated UEs by adjusting their minimum quotas.

Fig. 6.5 compares the maximum load difference resulting from the proposed MMQ algorithm, compared to the max-SINR approach with CRE. We observe that, as γ_{SINR} is increased from 0 to 8 dB, the maximum load difference decreases. However, for $\gamma_{\text{SINR}} > 8$ dB, the load difference increases, since a larger number of UEs is being assigned to the μW-BSs. Moreover, Fig. 6.5 shows that for all network sizes, the proposed approach substantially outperforms the max-SINR

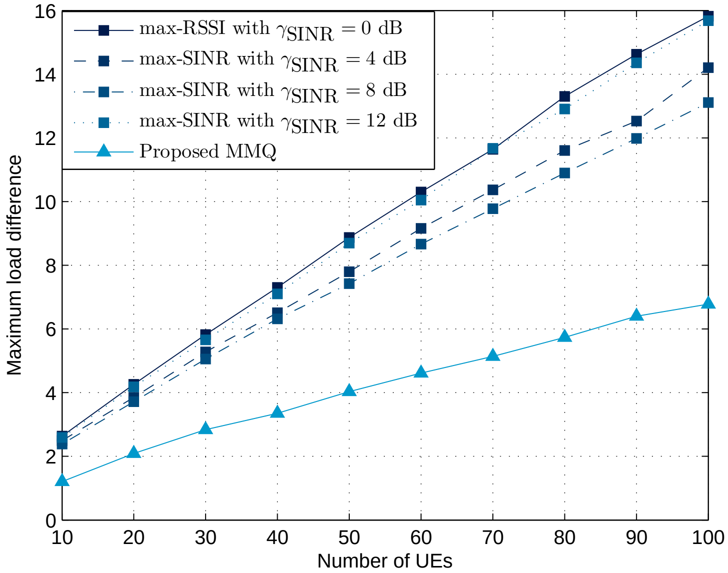


Figure 6.5: The maximum load difference, $\varDelta_\kappa$, for the proposed MMQ approach, compared to the max-SINR with CRE.

approach with CRE. In fact, the proposed approach decreases $\varDelta_k$ by 47%, compared to max-SINR with $\gamma_{\text{SINR}} = 8$ dB for $M = 70$. Here, we can once again see that the minimum quota constraints allow BSs to control the load more precisely, compared to max-SINR with CRE.

In Fig. 6.6, the statistics of the average rate per UE are shown over the μW band, compared to max-RSSI and max-SINR. Here, we focus on the average rate for μW links, since the mmW links achieve higher rates, due to the available bandwidth. The results show that, an inherent byproduct of any load balancing technique is the fact that some of the UEs will eventually be associated with an unpreferred μW-BS to satisfy the minimum quota constraints. Such UEs will then trade off rate for load balancing. To this end, parameter c_{th} is defined as a utility threshold for UEs. That is, the UE m is assigned to an unpreferred μW-BS n, if $U_m(n) \geq c_{\text{th}}$. c_{th} allows controlling the tradeoff between a highly balanced load and a low average rate for the cell edge UEs. Fig. 6.6 shows that for $c_{\text{th}} = 0.5$, the proposed MMQ algorithm outperforms the max-RSSI and the max-SINR approaches with CRE.

6.6 Summary

In this chapter, we have proposed a novel cell association and load balancing framework for small base stations operating at mmW and μW frequency bands. We have formulated the problem as a one-to-many matching game with minimum quotas. To solve this game, we have proposed a distributed algorithm that considers the average LoS probability in addition to the achievable rate, while assigning UEs to the BSs. We have shown that the proposed algorithm yields a Pareto

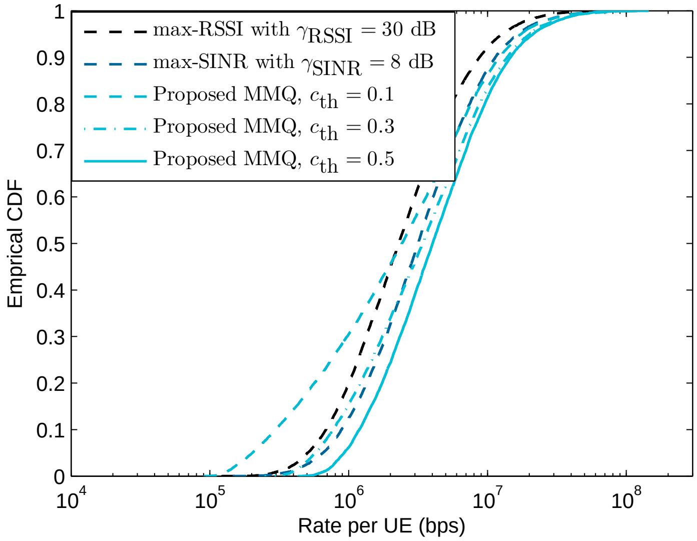


Figure 6.6: The empirical CDF of the rate per UE over μW frequency band for $M = 100$ UEs.

optimal and stable association policy. Simulation results have shown that the proposed MMQ algorithm outperforms the conventional max-RSSI and max-SINR schemes in terms of both performance and load balancing.

6.7 Appendix D

D.1 Proof of Proposition 4

We prove this using an example. Let $\mathcal{M} = \{m_1, m_2, m_3\}$ and $\mathcal{N} = \{n_1, n_2, n_3\}$, with ML profile $m_1 \succ_{\text{ML}} m_2 \succ_{\text{ML}} m_3$. In addition, assume $q_n^{\text{min}} = 1$, $q_n^{\text{max}} = 2$ for all BSs, and $n_1 \succ_{m_i} n_2 \succ_{m_i} n_3$, for all $m_i \in \mathcal{M}$. The DA algorithm for the UE-proposed solution yields $\pi(n_1) = \{m_1, m_2\}$, $\pi(n_2) = \{m_3\}$, and $\pi(n_3) = \emptyset$, which does not satisfy the minimum quota constraint for n_3.

D.2 Proof of Theorem 7

If the cell association π, given by Algorithm 7 is not PO, a UE m must exist that can benefit by being assigned to another BS n, i.e., $n \succ_m \pi(m)$. There are two possible cases to consider. First, $n \succ_m \pi(m)$ and $m \notin \pi(n)$ imply that UE m has applied to BS n prior to $\pi(m)$ and is rejected, due to $\kappa_n = q_n^{\text{max}}$ and $m' \succ_{\text{ML}} m$, for all $m' \in \pi(n)$. Therefore, adding m to $\pi(n)$ does not yield a feasible solution. Second, UE m is assigned to $\pi(m)$ to satisfy minimum quota constrain for $\pi(m)$. This means re-allocating m to BS n will violate the minimum quota criterion for $\pi(m)$ and is not feasible. Therefore, the given solution is feasible Pareto optimal.

To prove the stability, we note that if UE m prefers to be assigned to BS $\pi(m')$, that implies $m' \succ_{\text{ML}} m$, otherwise, $\pi(m) = \pi(m')$. Hence, no blocking pair exists and the solution is stable.

Chapter 7

Performance Analysis of Integrated Sub-6 GHz-Millimeter Wave Wireless Local Area Networks

7.1 Background, Related Works, and Summary of Contributions

Advanced wireless stations (STAs) are capable of supporting multiple WLAN standards, including legacy IEEE 802.11 over the *sub-6 GHz (microwave) unlicensed bands*, as well as IEEE 802.11ad over the 60 *GHz mmW band* [149]. These modern STAs, also known as tri-band WiGig devices, can potentially benefit from high capacity mmW communications along with flexible, simple, and more reliable networking at the sub-6 GHz bands. Reaping the benefits of such a multi-band WLAN capability is contingent upon adopting new MAC protocols that can support flexible and dynamic traffic scheduling over the aggregated mmW-μW unlicensed frequency bands[1]. Such promising integrated mmW-μW protocols also provide substantial motivation to revisit the existing MAC solutions for traditional, yet important challenges of WLANs. One such problem is the excessive delay at the contention-based medium access of the IEEE 802.11 standards that prevents WLANs to meet the stringent QoS requirements of emerging technologies, such as smart home applications [150, 151].

7.1.1 Related works

The performance of IEEE 802.11 MAC protocols has been thoroughly studied in the literature [152–156]. The seminal work of Bianchi in [152] presents a comprehensive analysis for the per-

[1] Hereinafter, μW unlicensed band refers to either 2.4 GHz, 5 GHz, or both.

formance of distributed coordination function (DCF) of the IEEE 802.11. The authors in [153] study the modeling and performance analysis of IEEE 802.11 DCF in unsaturated scenarios with heterogeneous traffic arrival rates for STAs. In [154], the authors propose a cooperative MAC protocol that leverages spatial diversity across the network to increase system throughput. The authors in [155] study the performance of enhanced-DCF (EDCF) for IEEE 802.11e standard. Moreover, the work in [156] and references therein propose different MAC protocols to improve QoS in IEEE 802.11. Although interesting, the body of work in [152–156] solely focuses on the WLAN standards at the μW unlicensed bands.

However, mmW communications over the 60 GHz unlicensed band is one of the key enablers to support emerging bandwidth-intensive technologies, such as virtual reality, in WLANs [157–160]. In fact, the large available bandwidth at 60 GHz mmW band allows STAs to potentially achieve higher data rates, compared with the data rates at the sub-6 GHz unlicensed μW bands. However, mmW links are inherently intermittent, due to extreme susceptibility of mmW signals to blockage. In addition, the challenges of bidirectional transmissions at the 60 GHz band, such as deafness, increase the complexity of MAC protocols.

In 2012, the IEEE 802.11ad standard [157] was introduced as an amendment to IEEE 802.11 in order to enable bidirectional transmissions over the unlicensed 60 GHz mmW frequency band and support a variety of services with different QoS requirements. In addition, this standard supports fast session transfer (FST) that enables STAs to dynamically migrate from one frequency band to another. This capability will enable advanced multi-band STAs to jointly manage their traffic over either 2.4, 5, or 60 GHz unlicensed frequency bands. The performance of IEEE 802.11ad is studied in [158–160]. The authors in [158] analyze the performance of IEEE 802.11ad MAC protocol using a three-dimensional Markov chain model. In [159], a directional cooperative scheme is proposed for 60 GHz mmW communications which is shown to improve the system performance, compared with the standard IEEE 802.11ad. In [160], the throughput analysis of IEEE 802.11ad under different modulation schemes is presented. The works in [158–160] focus solely on performance analysis of the IEEE 802.11ad as a stand-alone system, although this standard has been designed to coexist with legacy IEEE 802.11.

7.1.2 Summary of contributions

The main contribution of this chapter is to propose an integrated mmW-μW MAC protocol that enables STAs to dynamically leverage the bandwidth available at the 60 GHz mmW band and alleviate the excessive delay caused by the contention-based medium access over the μW frequencies. In addition, we present a comprehensive performance analysis for the proposed protocol by adopting a Markov chain model for backoff time that accommodates FST between mmW and μW frequency bands. Furthermore, simulation results are provided and shown to perfectly corroborate the derived analytical results. Both analytical and simulation results show that the proposed MAC protocol significantly increases the saturation throughput and reduces the delay, compared with the legacy IEEE 802.11 DCF. Moreover, the impact of different network parameters, such as mmW

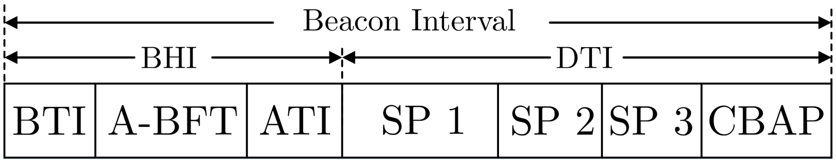


Figure 7.1: Beacon Interval structure [157].

link state, initial backoff window size, and maximum backoff stage on the performance are studied.

The rest of this chapter is organized as follows. Section 7.2 presents the proposed MAC protocol. Section 7.3 presents the analytical results. Simulation results are provided in Section 7.4. Section 7.5 concludes the chapter.

7.2 MAC Protocol Integration for Multi-Band Sub-6 GHz and MMW WLANs

The contention-based medium access in the IEEE 802.11 DCF suffers from increased backoff time and excessive delays in congested scenarios [150, 151]. To alleviate this problem, our goal is to leverage the multi-band operability of modern STAs to avoid excessive backoff times for collided frames and thus, decrease the associated contention delay for services in WLANs. *Prior to presenting the proposed scheme, we briefly overview some of the key definitions in the IEEE 802.11ad MAC protocol and 802.11 DCF that will be used in our analysis within the subsequent sections.*

7.2.1 IEEE 802.11ad MAC protocol overview

IEEE 802.11 standards, including IEEE 802.11ad, organize the medium access using periodic recurring beacon intervals (BIs). To accommodate bidirectional transmissions over the 60 GHz mmW band, some adjustments are introduced in the IEEE 802.11ad BI structure, as shown in Fig. 7.1. These modifications include: 1) sending directional beacon frames via an antenna sweeping mechanism, implemented within the beacon time interval (BTI). This sweeping process allows to extend the communication range and resolve the issue of STA discovery with unknown directions, 2) association beamforming training (A-BFT) used by stations to train their antenna sector for communication with the personal basic service set (PBSS) control point (PCP)/access point (AP), and 3) the PCP/AP exchanges management information, including scheduling, with beam-trained STAs prior to the data transmission interval (DTI).

During DTI, three different medium access schemes are supported, namely, 1) contention-based access, 2) scheduled channel time allocation, and 3) dynamic channel time allocation. The first scheme which is conventional in IEEE 802.11 protocols allows STAs to access channel during contention-based access periods (CBAPs). Two latter approaches are based on TDMA that dedicate a session period (SP) to each pair of scheduled STAs. The dynamic channel time allocation method includes a polling phase (PP) that enables STAs to request a channel time from the PCP/AP. The PCP/AP allocates the available channel time according to these requests. This polling-based scheduling mechanism is implemented within the beacon header interval (BHI).

7.2.2 IEEE 802.11 DCF overview

In this widely adopted protocol, STAs follow the contention-based carrier-sense multiple access with collision avoidance (CSMA/CA) scheme to reduce collisions [161]. That is, an STA senses the channel prior to sending its packet. If channel is sensed busy, the STA defers the transmission until the channel is sensed idle for a DCF Interframe Space (DIFS) time. Afterwards, the STA chooses a random backoff counter (BC). Then, time is divided into slots and the BC will be decremented after each idle slot time. Moreover, the BC countdown is stopped, whenever the channel is sensed busy during a slot time. The BC count down is reactivated once the channel is sensed idle again for a DIFS. The STA sends its packet immediately after BC reaches zero.

The BC is randomly selected from integers within an interval [0, CW-1], where CW is called contention window. CW depends on the number of transmissions failed for the packet. Initially, CW is set equal to a value W, called minimum contention window. After each unsuccessful transmission, W is doubled, up to a maximum value of $2^m W$. At this point, if transmission fails again, the packet is either dropped or a new BC is chosen randomly from $[0, 2^m W - 1]$.

7.2.3 Proposed integrated MmW-microwave MAC protocol

In this work, we focus on the IEEE 802.11 DCF and IEEE 802.11ad dynamic channel time allocation, respectively, at the μW and mmW unlicensed bands. In order to reduce the excessive delay caused by the collisions at the IEEE 802.11 DCF, in this section, we propose a novel protocol that enables STAs with multi-band capability to transfer their traffic to the contention-free 60 GHz mmW band, whenever available, and avoid intolerable large backoff times. The proposed protocol is shown in Fig. 7.2. In this example scenario, STAs 1 and 2 are, respectively, the transmitting and receiving stations. The communications between STAs 1 and 2 can be explained in three following phases:

Phase 1: STA 1 aims to transmit its packet to STA 2, over the μW band using a CSMA/CA scheme, as explained in Sec. 7.2.2. Due to its omnidirectional MAC protocol, the DCF of IEEE 802.11 requires minimum coordination among STAs, which provides a fast and flexible medium access. However, as the number of STAs increases, larger backoff times are required, resulting in more delay for packet transmissions. According to this protocol, STA 1 increases its backoff stage after each unsuccessful transmission. After reaching the maximum backoff stage m, STA 1 initiates Phase 2 with probability β and remains in Phase 1 with probability $1 - \beta$. The merit of using this control parameter will be elaborated in the next section.

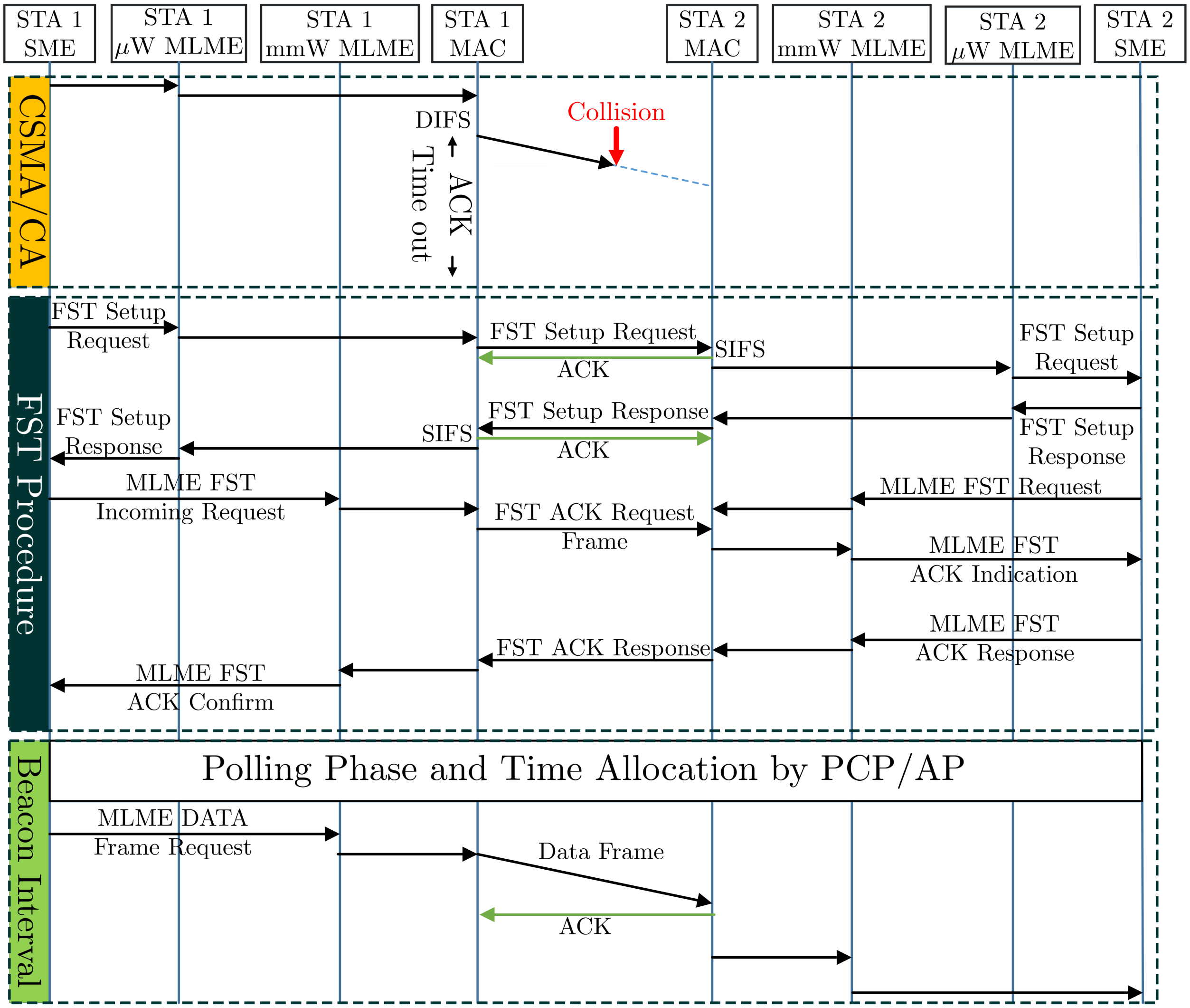


Figure 7.2: Proposed Multi-Band MAC Protocol.

Phase 2: In this phase, STA 1 initiates an FST with STA 2. FST capability is introduced in the IEEE 802.11ad Extended version [157] that enables STAs to swiftly move their traffic from one transmission band/channel to another. Since the FST is managed at a separate control channel, it will not be prone to collisions at the data channel. As shown in Fig. 7.2, to invoke FST, the station management entity (SME) unit in STA 1 sends an *FST Setup Request* to the μW MAC layer management entity (MLME), followed by informing the STA 1's MAC to forward the *FST Setup Request* frame to STA 2. Then, a handshaking procedure is done between STAs 1 and 2 in which STA 2 confirms that it is ready to move the communication to the 60 GHz band. Up to this stage, the control messages between STAs 1 and 2 are exchanged at the μW band. Next, an *FST ACK Request* is initiated by the STA 1's mmW MLME to request an FST ACK frame from STA 2. This message is transferred over the 60 GHz band and FST is done once STA 1 receives the *FST ACK Response* frame from STA 2.

Here, we note that the FST procedure is revoked if STA 1 does not receive ACK frames in any stage during Phase 2. This can happen if the link between STAs 1 and 2 is blocked by an obstacle, or A-BFT is failed. In that case, STA 1 continues following the CSMA/CA in Phase 1.

Phase 3: This phase starts with the next BI of the IEEE 802.11ad, in which STA 1 participates in the polling within PP of BI and requests a contention-free time for communication with STA 2, as elaborated in Sec. 7.2.1. Next, STA 1 will transmit its packet to STA 2 during the allocated SP in DTI. Afterwards, STA 1 will reset its CW to the minimum value W and it will initiate Phase 1.

The proposed multi-band MAC benefits from the flexible and simple CSMA/CA protocol at the μW unlicensed bands, while preventing excessive delays caused by the contention-based medium access. Next, we present analytical results to evaluate the performance of the proposed MAC protocol.

7.3 Modeling and Analysis of the Proposed Multi-Band MAC Protocol

In this section, we present analytical results to evaluate the performance of the proposed multi-band MAC protocol. First, we study the operation of an arbitrary STA that follows the proposed MAC protocol. In particular, we determine the probability of packet transmissions over either mmW or μW frequencies at a randomly chosen time slot. Then, we use these transmission probabilities to find a suitable expression for the saturation throughput.

7.3.1 Probability of packet transmission over mmW and μW frequencies

In our analysis, we assume non-empty queues for all STAs, i.e., the network operates at a saturation condition. As such, a new packet will be ready for transmission immediately after each successful transmission. These consecutive transmissions will require each STA transmitting over the μW frequency band to wait for a random backoff time prior to sending the next packet. In this regard, let $b(t)$ be the stochastic process for the BC of an arbitrary STA. A discrete and integer time scale is adopted in which t and $t+1$ present the beginning of two consecutive slot times, and the BC of each STA is decremented at the beginning of each slot time. According to the works in [152] and [153], the DCF of IEEE 802.11 can be modeled as a two-dimensional discrete-time Markov chain $(s(t), b(t))$, where $s(t) \in \{0, 1, \cdots m\}$ represents the backoff stage of an STA at time t, with m being the maximum backoff stage. For an arbitrary backoff stage $s(t) = i$, the CW will be $W_i = 2^i W$. In these Markov chain models, it is collectively assumed that, regardless of state, each packet collides with a constant and independent probability p as concretely discussed in [152] and [153].

To study the performance of the proposed protocol, we adopt a Markov chain model, as shown

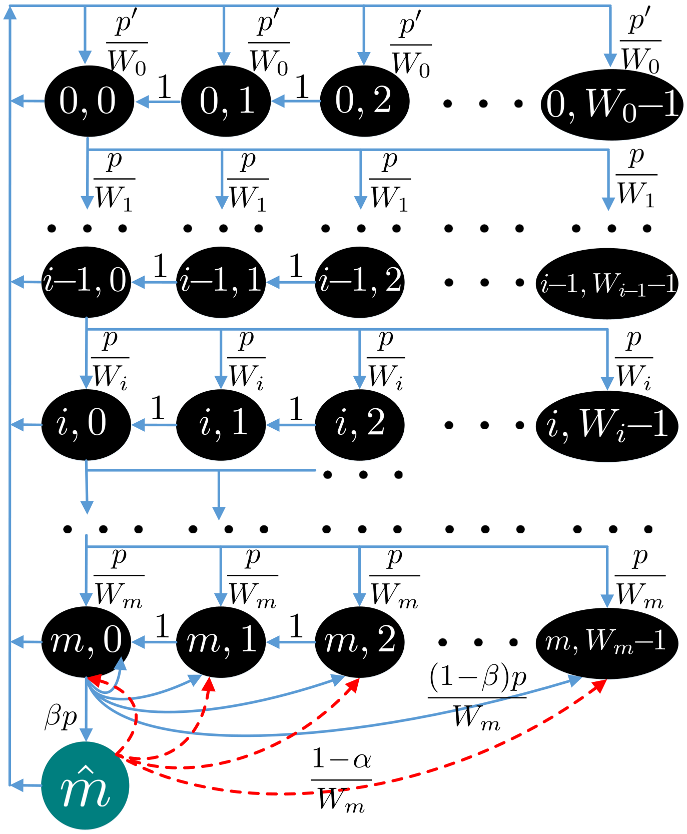


Figure 7.3: Markov Chain model for the backoff window size

in Fig. 7.3, where each state (i,k) indicates that $s(t) = i$ and $b(t) = k$, i.e., the BC of an STA is at the k-th step of stage i. In addition, by introducing a new state $\hat{m}$, this model captures the capability of multi-band STAs to operate at the mmW frequency band. In fact, while being at state $(m,0)$, the STA can choose to either stay at the μW band and follow the DCF protocol or perform an FST to transmit over the mmW band. We note that performing FST by an arbitrary STA j does not alter the collision probability p for other packets, since the next backlogged packet of STA j will be ready to be sent over the μW frequency band. In this model, $\beta \in [0,1]$ is a control parameter that allows an STA to manage unnecessary FSTs to reduce signaling overhead or avoid the mmW frequency band whenever the transmission of a number of previous packets has failed, due to unsuccessful A-BFTs. Moreover, β provides backward compatibility for legacy STAs with no mmW communications capability[2]. The state of each mmW link is determined by a Bernoulli random variable η with success probability α. That is, with probability α_j, a transmitting STA j and its desired receiving STA can successfully perform the A-BFT and execute the transmission.

[2] By choosing $\beta = 0$, the proposed model will converge to the corresponding Markov chain for the conventional DCF in IEEE 802.11 standards.

Here, the single-step nonzero transition probabilities are

$$\mathbb{P}\left\{i,k|i,k+1\right\}=1, \qquad i\in[0,m], k\in[0,W_i-2], \tag{7.1a}$$

$$\mathbb{P}\left\{0,k|i,0\right\}=p'/W_0, i\in[0,m]\cup\left\{\hat{m}\right\}, k\in[0,W_0-1], \tag{7.1b}$$

$$\mathbb{P}\left\{m,k|\hat{m}\right\}=(1-\alpha)/W_m, \qquad k\in[0,W_m-1], \tag{7.1c}$$

$$\mathbb{P}\left\{i,k|i-1,0\right\}=p/W_i, \qquad i\in[1,m], k\in[0,W_i-1], \tag{7.1d}$$

$$\mathbb{P}\left\{m,k|m,0\right\}=p(1-\beta)/W_m, \qquad k\in[0,W_m-1], \tag{7.1e}$$

$$\mathbb{P}\left\{\hat{m}|m,0\right\}=\beta p, \tag{7.1f}$$

where (7.1a) shows the backoff time count down at each time slot. Moreover, (7.1b) indicates that, after a successful packet transmission, an STA will randomly choose a BC from stage 0, i.e., k is uniformly chosen from $[0, W-1]$. In (7.1b), p' is equal to $1-p$ and α, respectively, if $i\in[0,m]$ and $i=\hat{m}$. In addition, (7.1c) captures an unsuccessful mmW transmission, after which the STA will remain at the μW frequency band and will choose a random backoff time at stage m. (7.1d) shows that backoff stage will incremented after an unsuccessful μW transmission. Furthermore, (7.1e) and (7.1f) indicate, respectively, that an STA will remain at stage m with probability $1-\beta$ after a collision, or will perform an FST with probability β.

For this Markov chain model, we next determine the stationary probability for each state (i,k). Let $h_{i,k}=\lim\limits_{t\to\infty}\mathbb{P}\{s(t)=i,b(t)=k\}$, $i\in[0,m]\cup\{\hat{m}\}$, $k\in[0,W_i-1]$. From the Markov chain model in Fig. 7.3, it is easy to see that

$$h_{i,0}=ph_{i-1,0}=p^i h_{0,0}, \quad i\in(0,m). \tag{7.2}$$

Furthermore, for $i=m$ and $\hat{m}$, we note that

$$ph_{m-1,0}+p(1-\beta)h_{m,0}+(1-\alpha)h_{\hat{m}}=h_{m,0}, \tag{7.3a}$$

$$p\beta h_{m,0}=h_{\hat{m}}. \tag{7.3b}$$

Using (7.2), we solve (7.3a) and (7.3b) with respect to $h_{m,0}$ and $h_{\hat{m}}$, which yields

$$h_{m,0}=\frac{p^m}{1-p+\alpha\beta p}h_{0,0}, \;\; h_{\hat{m}}=\frac{\beta p^{m+1}}{1-p+\alpha\beta p}h_{0,0}. \tag{7.4}$$

Next, by following the chain regularities, we can represent the remaining stationary state probabilities as:

$$h_{i,k}=W_i'\begin{cases}\frac{1-p}{W_0}\sum_{j=0}^{m}h_{j,0}+\frac{\alpha}{W_0}h_{\hat{m}}, & i=0,\\ \frac{p}{W_m}h_{m-1,0}+\frac{p(1-\beta)}{W_m}h_{m,0}+\frac{1-\alpha}{W_m}h_{\hat{m}}, & i=m,\\ \frac{p}{W_i}h_{i-1,0}, & i\in(0,m),\end{cases} \tag{7.5}$$

where $W_i'=W_i-k$, and $k\in(0,W_i-1]$. In addition, we note that

$$h_{0,0}=(1-p)\sum_{j=0}^{m}h_{j,0}+\alpha h_{\hat{m}}. \tag{7.6}$$

Thus, by using (7.2), (7.3a), and (7.6), $h_{i,k}$ in (7.5) simplifies to:

$$h_{i,k} = \frac{W_i - k}{W_i} h_{i,0}, \qquad i \in [0, m], k \in (0, W_i - 1]. \tag{7.7}$$

Finally, we find $h_{0,0}$ by noting that the sum of all state probabilities is 1. That is,

$$\begin{aligned} 1 &= \sum_{i=0}^{m} \sum_{k=0}^{W_i - 1} h_{i,k} + h_{\hat{m}} \\ &\overset{\text{(a)}}{=} \sum_{i=0}^{m} h_{i,0} \sum_{k=0}^{W_i-1} \frac{W_i - k}{W_i} + h_{\hat{m}}, \\ &\overset{\text{(b)}}{=} \sum_{i=0}^{m-1} h_{i,0} \frac{W_i + 1}{2} + \frac{W_m + 1}{2} b_{m,0} + h_{\hat{m}}, \\ &\overset{\text{(c)}}{=} \left[\sum_{i=0}^{m-1} (W_i + 1)p^i + \frac{(W_m + 1)p^m}{1 - p + \alpha\beta p} + \frac{2\beta p^{m+1}}{1 - p + \alpha\beta p} \right] \frac{h_{0,0}}{2}. \end{aligned} \tag{7.8}$$

In (7.8), (a) and (b) result from (7.7) and noting that $\sum_{k=0}^{W_i-1}(W_i - k)/W_i = (W_i + 1)/2$, respectively. In addition, (c) results from (7.2) and (7.4). From (7.8), we can find $h_{0,0}$ as follows:

$$\begin{aligned} h_{0,0} = 2 \Bigg[& W \left(\frac{1 - (2p)^m}{1 - 2p} \right) \\ & + \frac{1 - p^m}{1 - p} + \frac{(2^m W + 1 + 2\beta p)p^m}{1 - p + \alpha\beta p} \Bigg]^{-1}. \end{aligned} \tag{7.9}$$

Next, we can compute the transmission probability over the μW band, $\varTheta^{\mu\text{W}}$, for an STA in a random time slot. To this end, we note that μW transmission occurs only if the backoff time countdown for an STA reaches zero. That is, an STA transmits a packet if it is at any states $(i, 0), i \in [0, m]$. Thus,

$$\varTheta^{\mu\text{W}} = \sum_{i=0}^{m} h_{i,0} = \frac{1}{1 - p} \left[1 - \frac{\alpha\beta p^{m+1}}{1 - p + \alpha\beta p} \right] h_{0,0}. \tag{7.10}$$

Remark 4. *Without mmW communications ($\beta = 0$), we can easily verify that $\varTheta^{\mu W}$ in* (7.10) *simplifies to*

$$\varTheta^{\mu W} = \frac{2(1 - 2p)}{(1 - 2p)(W + 1) + pW(1 - (2p)^m)}, \tag{7.11}$$

which is shown to be the transmission probability in DCF protocol of the IEEE 802.11 [152].

Over the mmW frequency band, STAs that are in state $\hat{m}$ will be scheduled to transmit within the next available DTI. Given that the mmW transmissions follow a TDMA scheme during each

SP, as proposed in IEEE 802.11ad, no collision will happen. However, as mentioned in section 7.2.1, a mmW transmission is contingent upon a successful A-BFT phase. Hence, the probability of transmission over the mmW frequency band is

$$\Theta^{\text{mmW}} = \mathbb{P}\{\eta = 1\}h_{\hat{m}} = \frac{\alpha\beta p^{m+1}}{1-p+\alpha\beta p}h_{0,0}. \tag{7.12}$$

After deriving the transmission probability at both mmW and μW frequencies for an arbitrary STA, our next step is to compute the saturation throughput as a key performance metric.

7.3.2 Throughput analysis of the proposed multi-band mmW-μW MAC protocol

Here, we focus on analyzing the system throughput R at the saturation conditions. This throughput is collectively defined as the average payload that is successfully transmitted across the network during a randomly chosen time slot, divided by the average time slot duration $\mathbb{E}[T]$. In multi-band WLANs, parallel streams of data can be sent simultaneously over different frequency bands. Thus, our analysis will focus on finding throughput across the aggregated mmW-μW frequencies.

Consider a WLAN, composed of J STAs within a set $\mathcal{J}$. Over the μW frequency band, the protocol follows the standard CSMA/CA. In other words, only one STA can successfully transmit at a given time, otherwise, collision happens. In this regard, $P_t^{\mu\text{W}}$ is defined as the probability that at least one STA is transmitting over the μW frequency band. Since each STA $j \in \mathcal{J}$ transmits with probability $\Theta_j^{\mu\text{W}}$, $P_t^{\mu\text{W}}$ is given by:

$$P_t^{\mu\text{W}} = 1 - \prod_{j\in\mathcal{J}}(1-\Theta_j^{\mu\text{W}}). \tag{7.13}$$

In addition, transmission of an arbitrary STA j is successful, if no other STA transmits at the same time. Hence, the probability of successful transmission can be written as:

$$P_s^{\mu\text{W}} = \frac{\sum_{j\in\mathcal{J}}\Theta_j^{\mu\text{W}}\prod_{j'\in\mathcal{J}\setminus j}(1-\Theta_{j'}^{\mu\text{W}})}{P_t^{\mu\text{W}}}. \tag{7.14}$$

To compute $\mathbb{E}[T]$, we note that there are three possible cases for the transmission scenarios over the μW band: 1) having an empty slot which occurs with probability $1 - P_t^{\mu\text{W}}$, since no STA is transmitting. 2) Successful transmission of a packet during a time slot which happens with probability $P_t^{\mu\text{W}}P_s^{\mu\text{W}}$, and 3) collision scenario that occurs with probability $P_t^{\mu\text{W}}(1-P_s^{\mu\text{W}})$. Hence, the average slot time is

$$\mathbb{E}[T] = (1-P_t^{\mu\text{W}})\sigma + P_t^{\mu\text{W}}P_s^{\mu\text{W}}T_s + P_t^{\mu\text{W}}(1-P_s^{\mu\text{W}})T_c, \tag{7.15}$$

where T_s, T_c, and σ denote the slot time duration, respectively, in successful, collision, and no transmission scenarios.

Table 7.1: Simulation parameters

Notation	Parameter	Value
H_{MAC}	MAC header	272 bits
H_{PHY}	PHY header	128 bits
$B^{\mu\text{W}}$	μW packet payload	8184 bits
B^{mmW}	mmW packet payload	81840 bits
ACK	ACK	112 bits + PHY header
δ	Propagation delay	1 μs
σ	Slot time	50 μs
SIFS	Short interframe space	28 μs
DIFS	Distributed interframe space	128 μs
$r^{\mu\text{W}}$	μW channel bit rate	1 Mbps
r^{mmW}	mmW channel bit rate	1 Gbps
SETUP_REQ	FST setup request	240 bits
SETUP_RES	FST setup response	240 bits

For mmW transmissions, we must note that only FST is performed over the μW band, while other phases during BHI as well as payload transmissions in DTI will be done simultaneously with the μW band transmissions. To properly capture the mmW band contribution in the system throughput, we consider the time overhead associated with performing FST and we find the average number of STAs that can be scheduled at the mmW frequency band within a coarse of $\mathbb{E}[T]$ time. In this regard, let $\hat{J} \leq J$ be the maximum number of STAs that can be scheduled over the mmW band during $\mathbb{E}[T]$, each transmitting a payload of size B^{mmW} bits. Considering r^{mmW} as the mmW channel bit rate, $\hat{J} = \lfloor \mathbb{E}[T] r^{\text{mmW}} / B^{\text{mmW}} \rfloor$, where $\lfloor . \rfloor$ is the floor operand. Consequently, the average number of STAs transmitting at the mmW frequency band, $\mathbb{E}[J^{\text{mmW}}]$, is

$$\mathbb{E}[J^{\text{mmW}}] = \sum_{u=1}^{\hat{J}} \sum_{s=1}^{\binom{J}{u}} \prod_{j=1}^{|\mathcal{J}'|=u} \Theta_j^{\text{mmW}}, \tag{7.16}$$

where the inner sum is taken over all possible subsets $\mathcal{J}' \subseteq \mathcal{J}$ with $|\mathcal{J}'| = u$ number of STAs. Clearly, there are $\binom{J}{u}$ distinct subsets with size u. Moreover, the product is for all STAs in the chosen subset $\mathcal{J}'$. In addition, since the protocol employs TDMA scheme for mmW communications, no collision will occur between multiple mmW transmissions during a DTI and the probability of successful transmission is $P_s^{\text{mmW}} = 1$.

Therefore, the system throughput R is calculated by finding the aggregated transmitted payload over both μW and mmW frequency bands, divided by the average time slot duration $\mathbb{E}[T]$ plus the time overhead associated with FST process. That is,

$$R = \frac{P_s P_t B^{\mu\text{W}} + \mathbb{E}[J^{\text{mmW}}] B^{\text{mmW}}}{\mathbb{E}[T] + \mathbb{E}[J^{\text{mmW}}] T_{\text{FST}}}, \tag{7.17}$$

where $B^{\mu\text{W}}$ is the payload size over the μW frequency band. Given the high available bandwidth

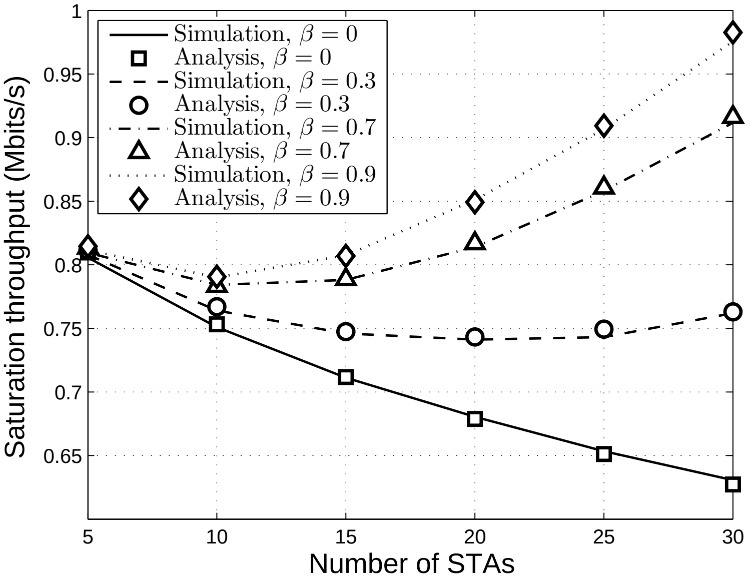


Figure 7.4: Saturation throughput vs the number of STAs.

at the mmW band, B^{mmW} is considered larger than $B^{\mu\text{W}}$. Moreover, T_{FST} is the required time for performing an FST.

7.4 Simulation Results

Here, we validate our analytical results by simulating the proposed protocol in a multi-band WLAN. The number of STAs varies from $J = 5$ to 50. The considered network is simulated in MATLAB and the total simulation time extends to 500 seconds. We consider $\alpha_j = \alpha, j \in \mathcal{J}$ to simplify the performance analysis. In this case, (7.13)-(7.16) can be written as:

$$P_t^{\mu\text{W}} = 1 - (1 - \Theta^{\mu\text{W}})^J, \tag{7.18a}$$

$$P_s^{\mu\text{W}} = J\Theta^{\mu\text{W}}(1 - \Theta^{\mu\text{W}})^{J-1}/P_t^{\mu\text{W}}, \tag{7.18b}$$

$$\mathbb{E}[J^{\text{mmW}}] = \sum_{u=1}^{\hat{J}} \binom{J}{u} \left(\Theta^{\text{mmW}}\right)^u. \tag{7.18c}$$

The effect of α and β on the network performance will be evaluated subsequently. For μW communications, we consider the basic access scheme[3] in which the receiving STA will send an acknowledgment (ACK) signal after successfully decoding the sent packet. Hence, T_s, T_c and T_{FST}

[3] Other access schemes such as request-to-send/clear-to-send (RTS/CTS) mechanisms can be applied similarly.

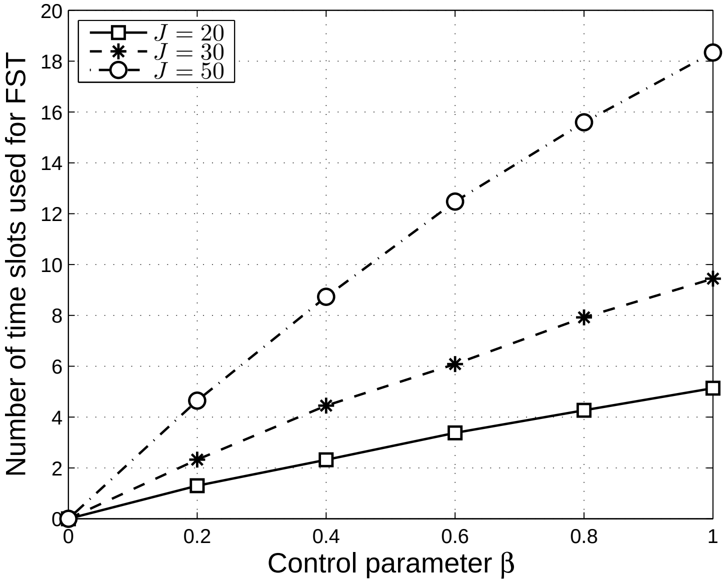


Figure 7.5: Number of time slots used in FST procedure vs the control parameter β, for different network size J.

are calculated as follows:

$$\begin{aligned} T_s &= \mathit{\Gamma} + \text{SIFS} + \text{ACK} + \text{DIFS} + 2\delta, \\ T_c &= \mathit{\Gamma} + \text{DIFS} + \delta, \\ T_{\text{FST}} &= \text{SETUP\_REQ} + \text{SETUP\_RES} + 2\text{ACK} + 4\delta, \end{aligned} \tag{7.19}$$

where $\mathit{\Gamma}$ is the required time for transmitting PHY header H_{PHY}, MAC header H_{MAC}, and payload $B^{\mu\text{W}}$ of a μW packet. Moreover, δ models the propagation delay. T_{FST} is calculated based on the FST procedure, as shown in Fig. 7.2, composed of sending FST Setup Request/Response frames, each followed by an ACK signal. Here, we note that FST ACK Request/Response frames are sent over the mmW frequency band, thus, they are not included in the time overhead. All network parameters are summarized in Table 7.1.

Fig. 7.4 shows the effect of control parameter β on the performance, for different number of STAs, with $m = 3$, $W = 32$, and $\alpha = 0.6$. From Fig. 7.4, we can see that the throughput increases as β becomes large. Interestingly, this performance gain is more evident for large network sizes, since the collision probability is higher and thus, the proposed protocol sends more packets over the mmW band. In addition, for any fixed non-zero β, we observe that throughput initially decreases and then increases, as the number of STAs grows. That is because for a larger network size J, collision initially increases which results in a lower throughput. However, with further network size growth, more STAs reach the maximum backoff stage m and initiate FST to the mmW band.

In Fig. 7.5, the overhead of the proposed protocol is evaluated in terms of the number of time slots used in the FST procedure. From Figs. 7.4 and 7.5, we observe an interesting tradeoff between the saturation throughput and the overhead of switching between mmW and μW frequency bands.

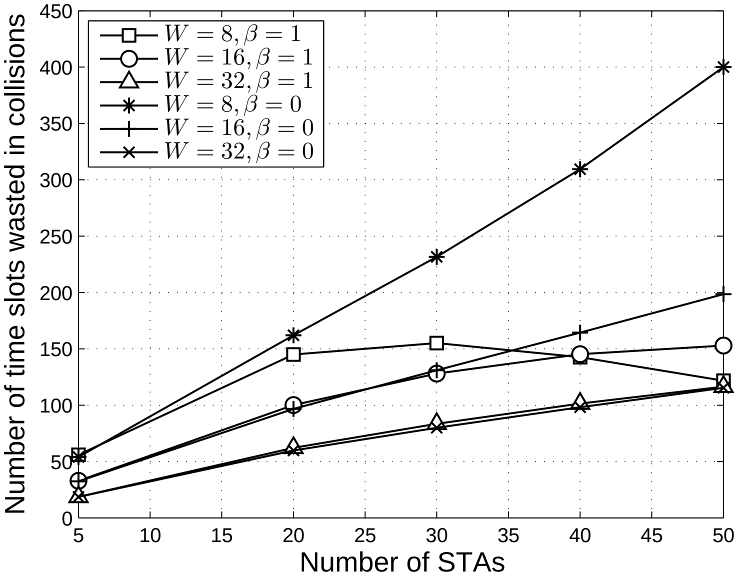


Figure 7.6: Number of time slots wasted in collisions vs the number of STAs, for different W and β values.

For example, from Fig. 7.4, we can see that the throughput is improved by 28% for $J = 30$, when β is increased from $\beta = 0.3$ to $\beta = 0.9$. Moreover, Fig. 7.5 shows that the overhead increases from 3 slots to 9 slots in order to achieve this performance gain. Next, we show that this overhead is negligible compared with the time wasted in collisions.

Fig. 7.6 shows another key merit of mmW-μW MAC layer integration which is reducing the packet transmissions delay caused by the collisions. This figure compares the number of time slots that are wasted in collisions by the proposed protocol ($\beta = 1$) and legacy IEEE 802.11 ($\beta = 0$), for different initial contention window and network sizes. From Fig. 7.6, it is clear that the proposed scheme significantly reduces the delay, e.g., up to three times for $J = 50$ STAs and $W = 8$. Moreover, we observe that the performance gap between the two schemes is larger for smaller W values. That is because more collisions occur when initial backoff window size is small, which increases the probability for STAs to transmit their packets over the mmW frequency band.

Fig. 7.7 shows the saturation throughput as a function of α for different number of STAs, with $m = 3$, $W = 32$, and $\beta = 1$. We can observe that, as mmW communication is more feasible, the throughput will increase with all network sizes. For example, the throughput increase by 37% for $J = 20$ and $\alpha = 0.9$, compared with the stand-alone IEEE 802.11 system ($\alpha = 0$). Similar to Fig. 7.4, the throughput varies as a convex function with respect to the number of STAs.

In Fig. 7.8, the impact of initial backoff window size, W, on the throughput is studied for $m = 3$, $\alpha = \beta = 0.5$, and three network sizes $J = 5, 10, 20$. Fig. 7.8 also shows the optimal W for maximizing the throughput. We can observe that the optimal W grows as the number of STAs J increases.

Furthermore, the effect of maximum backoff stage, m, on throughput is shown in Fig. 7.9 with

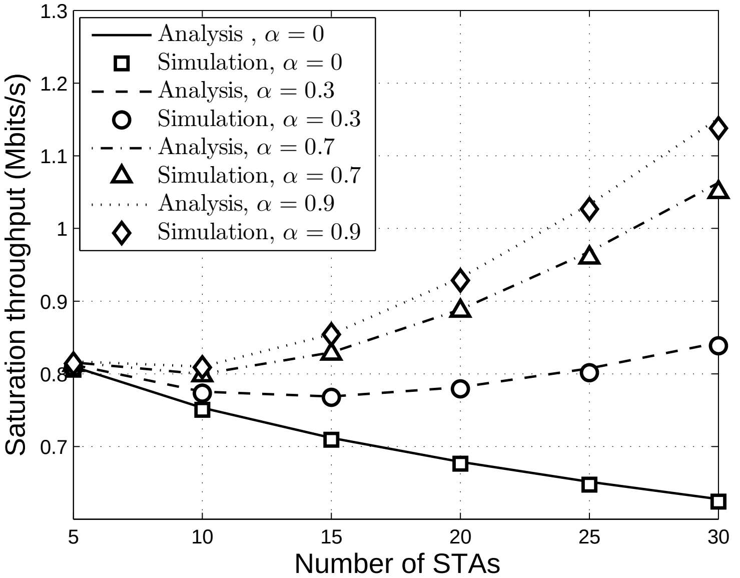


Figure 7.7: Saturation throughput vs the number of STAs.

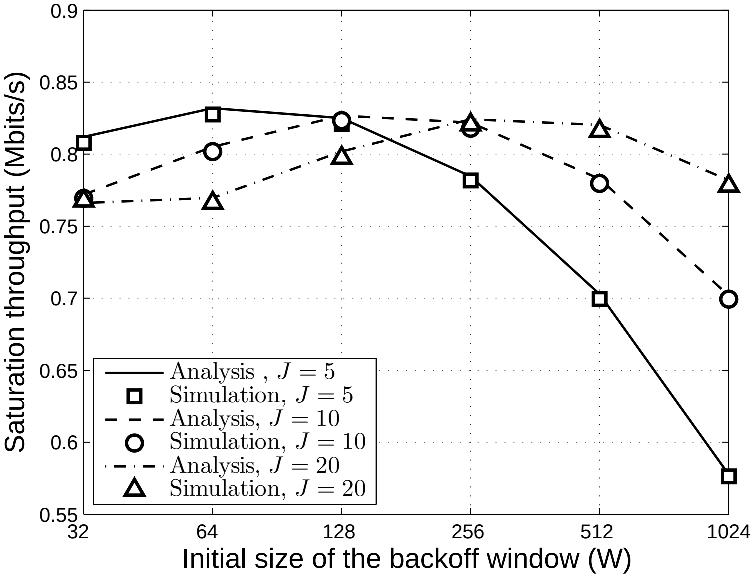


Figure 7.8: Saturation throughput vs W for different network size J.

$\beta = 0.5$, $W = 16$, and $J = 50$. It is interesting to note that for $\alpha = 0$, i.e., with no mmW communications, throughput increases as m grows. That is because less collisions happen with larger maximum backoff. However, this trend is opposite for nonzero α values. In fact, even for $\alpha = 0.2$ and small m, we observe a significant performance gain which results from STAs' frequent switching to the mmW frequency band, due to the high collision at the μW frequency band.

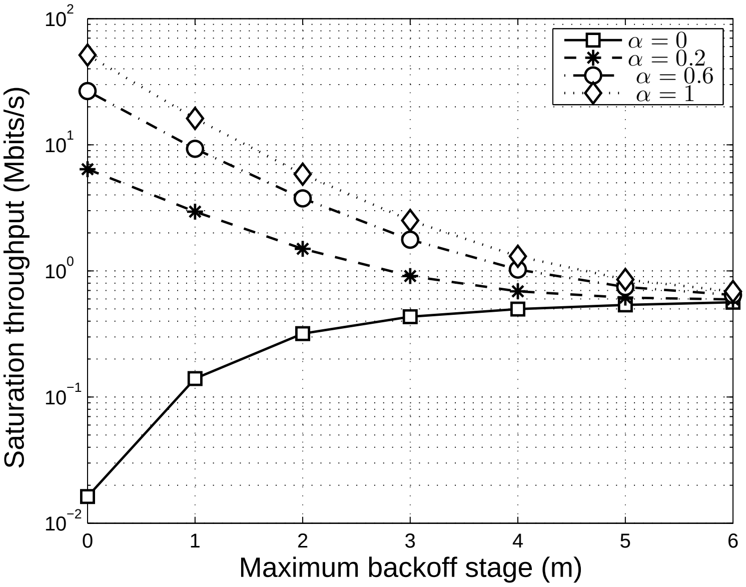


Figure 7.9: Saturation throughput vs m for different α values.

7.5 Summary

In this chapter, we have proposed a novel MAC protocol that leverages the capability of advanced wireless stations to decrease the contention-based delay and increase throughput in WLANs. In fact, the proposed protocol allows stations to perform fast session transfer to the 60 GHz mmW band, and avoid excessive delay caused by collisions at the μW unlicensed bands. To analyze the performance of the proposed scheme, we have adopted a Markov chain model that captures the fast session transfer across mmW-μW bands. We have shown the accuracy of the model by providing comprehensive simulation results. Both simulations and analytical results have shown that the proposed protocol yields significant gains in terms of maximizing the saturation throughput and minimizing the delay caused by collisions.

Chapter 8

Social-Aware Resource Allocation in Small Cell Networks with Underlaid Device-to-Device Communications

8.1 Background, Related Works, and Summary of Contributions

The introduction of smartphones and tablets has led to the proliferation of bandwidth-intensive wireless services, such as multimedia streaming and social networking, that have strained the capacity of present-day wireless communication networks [100]. This increasing trend led to the emergence of wireless SCNs as a promising solution to meet the QoS requirements of such emerging wireless services [3–6]. In SCNs, the main idea is to massively deploy small cell base stations (SCBSs) with relatively low transmit power, overlaid on existing cellular infrastructure. Small cells allow to increase the capacity and coverage of a wireless network by bringing the UEs closer to their serving base stations. Nonetheless, the deployment of small cells introduces new challenges in terms of interference management, resource allocation, and network modeling. These challenges stem from many key features of SCNs such as the unplanned SCBS distribution, limited coverage, dense SCBS deployment, and limited backhaul capacities, among others [3–6, 162–165].

8.1.1 Related works

The authors in [162] proposed a control-based scheduler for traffic management at small cells. In [163], an optimization problem is solved at each cell to perform resource allocation while taking into account cell range expansion and offloading metrics. Most of these existing approaches mainly adopt centralized methods for resource allocation [162, 163]. Although interesting, such centralized approaches have several drawbacks since they assume the presence of a centralized controller for the small cells, depend on SCBSs cooperation, and require MBS coordination. However, resource allocation in SCNs needs to be decentralized, self-organizing, and computationally efficient; specifically when the number of small cells increases. In this regard, game theory has emerged as a popular tool to realize distributed approaches for wireless networks [164–169]. In [164], the authors proposed a distributed resource allocation in the uplink of a two tier network, by posing the problem as a matching game. They solved the game using the Hungarian algorithm. In [165], the resource allocation in SCNs is formulated as an evolutionary game. In [166], the theory of one-to-one and many-to-one matching markets is extended for the resource allocation in wireless networks. In [167], the authors used matching theory to perform distributed scheduling at the downlink of a MIMO-OFDMA system. Other works that apply matching in some limited wireless settings are found in [168] and [169]. In fact, prior works do not handle the challenges of SCNs that are underlaid with D2D connections and in which there is a need not only to manage interference, but to also account for redundant transmissions by exploiting the ability of D2D to provide popular content chaching. The body of work in [164–169] focuses on resource allocation while only accounting for classical physical layer metrics such as the SINR and is restricted to networks without D2D. In addition, it is based on the classical deferred acceptance algorithm which cannot be applied for scenarios with peer effects such as in our case. Context-aware resource allocation, as done in this chapter, is a new design paradigm that can help to boost the performance of small cell networks and to exploit D2D for popular content distribution.

Along with the use of SCNs for improving network performance, D2D communications has recently emerged as an interesting approach to provide proximity services to users of an SCN, thus assisting in further offload of the cellular system's traffic [37–41]. Indeed, due to the evolution of numerous data centric applications, it is very likely that devices in proximity of one another tend to interact directly over the wireless spectrum. Communication of such neighboring devices via the infrastructure of the SCN (i.e., via SCBSs) is neither spectrum nor power efficient [37]. In addition, SCNs are envisioned to have a capacity-limited backhaul [5] and, thus, the use of underlaid D2D can help offload traffic from the SCNs' backhaul. In this respect, D2D communication over the cellular spectrum is viewed as an attractive candidate to handle these scenarios [40, 41, 170]. D2D over cellular networks is significantly different from D2D over unlicensed bands such as WiFi or traditional short-range D2D via Zigbee and Bluetooth. Indeed, D2D over cellular allows longer ranges and higher QoS, while also requiring to properly manage interference with cellular transmissions [38, 39], and [170]. Some of the main challenges associated with deploying the D2D technology include introducing proximity services that leverage D2D, managing the wireless resources in D2D deployments, and protecting such low power and vulnerable communication links from interference [37–40].

In addition to the conventional physical layer metrics to optimize the SCN's performance, modern UEs can offer a versatile range of information from higher network layers that could help to reap the prospective gains of D2D deployments in SCNs. Such additional information, referred to as *context information*, may include the data extracted from online social networks [51–54], the history of a user's throughput [48], prediction of a user's location [50], or the delay-throughput tolerance of the applications [49]. The work in [51] develops an analytical model for the epidemic information spreading among mobile users of an ad hoc network. In [52], the resource allocation in wireless LAN is defined as an optimization problem, taking the notion of social distance into account. The authors in [53] and [54] extend the work in [52] by introducing new utility functions which again account for the social distance of users, extracted from the social graph. Existing context-aware works [48–54] are mostly tailored to macrocell networks, are based on centralized approaches, and do not address the SCN or D2D over SCN challenges. In addition, although the use of social networks has been demonstrated to be useful to improve wireless systems, most existing works such as [51–54] are based solely on the physical aspects of the social network, e.g., centrality measures. Such notion of social context is insufficient to capture common interests. For example, the large number of friends of a user in a social network does not necessarily mean that such a user is influential enough to require more bandwidth. In contrast, there is a need to adopt a more holistic view for the social context by basing it on other social dimensions such as the actual interactions between users. Here, we note that, although another body of works such as [171–174] has further explored the use of social metrics in networking applications, such works are not adequate for deployment in wireless cellular systems such as SCNs with D2D as they do not deal with issues such as interference and network offload.

8.1.2 Summary of contributions

The main contribution of this chapter is to propose a novel, self-organizing, context-aware framework for optimizing resource allocation in D2D-enabled SCNs. We formulate the problem as a two-sided one-to-one matching game in which each UE is assigned to one RB. In this game, the UEs and SCBS-controlled RBs rank one another based on utilities that capture the social context of the users as well as the wireless physical layer metrics. The social context includes the information inferred from the social network profiles of the wireless users. This information is mainly based on the similarities between users' interests, activities, and their interactions such as tagging or wall posting. The proposed scheme allows to exploit the fact that users who are strongly connected in a social network are likely to request similar type of data over the physical wireless network. We show that the proposed game is a matching game with *peer effects* in which the strategy of each player is affected by the decisions of its peers. This is in contrast to most existing works on matching theory for wireless networks [164–169] that deal with conventional matching games in which there is no peer effect. To solve this context-aware resource management game, we propose a novel, self-organizing algorithm that allows to find a stable matching between users and RBs. We show that our proposed algorithm allows the SCBSs and UEs to interact and converge to a stable matching with manageable complexity. Simulation results using real traces are used to analyze the performance of the proposed approach.

The rest of this chapter is organized as follows. Section 8.2 describes the system model. Section 8.3 introduces the modeling of social context in wireless D2D-enabled SCNs. Section 8.4 defines the problem as the matching game and Section 8.5 presents the proposed algorithm. Simulation results are analyzed in Section 8.6 and conclusions are drawn in Section 8.7.

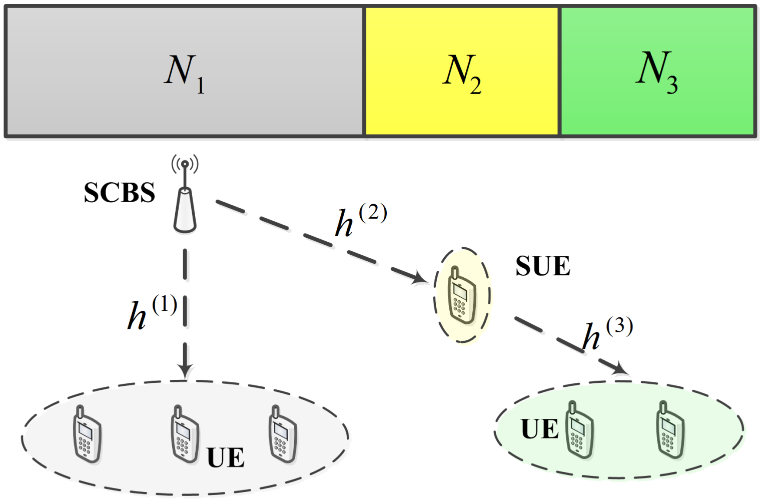


Figure 8.1: Physical model for interference management in a D2D enabled SCN.

8.2 System Model

Consider the downlink of an OFDMA small cell network with a set $\mathcal{L}$ of L SCBSs randomly distributed within the network. The total bandwidth B is divided into N RBs in the set $\mathcal{N}$ and there are a total of M active users with $\mathcal{M}$ being the set of all users. We consider a co-channel network deployment in which the total bandwidth is shared between all small cells. In this network, we assume that users can communicate directly via D2D communication links within the cellular band. Such D2D communications enhance the indoor coverage and helps to offload the small cell traffic. In our model, some users are chosen as a serving user equipment (SUE) that are allowed to serve other UEs via D2D communication. Let $\mathcal{M}_s$ be the set of M_s SUEs and $\mathcal{M}_u$ be the set of M_u non-serving UEs. Thus, $\mathcal{M} = \mathcal{M}_s \cup \mathcal{M}_u$ and $\mathcal{M}_s \cap \mathcal{M}_u = \emptyset$. The criterion for SUE selection is discussed further in Section 8.3.1. Moreover, let $\mathcal{K} = \mathcal{M}_s \cup \mathcal{L}$ be the joint set of all SCBSs and SUEs with $|\mathcal{K}| = M_s + L$. Hereinafter, we use the term “serving node (SN)” to refer to either an SCBS or an SUE. Moreover, we refer to cellular links and D2D links, respectively, as SCBSs to UEs and SUEs to UEs links.

For resource allocation in D2D-enabled SCNs, one simple approach is to allow the SCBSs to share all the RBs with the SUEs. However, in such a scheme, the D2D communication links will be dominated by the interference from the SCBSs. To overcome this problem, we propose to divide the spectrum in such a way that no mutual interference occurs between SCBSs and SUEs. Consequently, the sources of interference and the SINR relations will differ at each RB $n \in \mathcal{N}$, depending on whether RB n is reused by an SCBS or an SUE. In this model, D2D links, SCBS to SUE links, and cellular links are separated in the frequency domain. Hence, the set of resource blocks $\mathcal{N}$, is divided into non-overlapping sets, namely, $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$ as shown in Fig. 8.1. $\mathcal{N}_1$ and $\mathcal{N}_2$ represent the set of N_1 and N_2 RBs dedicated to the direct links from SCBSs to UEs and SCBSs to SUEs, respectively. In addition, $\mathcal{N}_3$ is the set of N_3 dedicated RBs that are shared by all SUEs for D2D transmission. We let h_{knm} be the channel state of subcarrier $n \in \mathcal{N}$ in the transmission from SN k to user m. In this model, SCBSs may interfere with one another, since they share subbands $\mathcal{N}_1$ and $\mathcal{N}_2$. However, SCBSs will not interfere with D2D links as SUEs and SCBSs transmit on two different orthogonal bands. This encourages UEs to be served via D2D links which can improve the offloading capabilities of the network.

The achievable rate for the transmission between an SN $k \in \mathcal{K}$ and a user $m \in \mathcal{M}$ over RB $n \in \mathcal{N}_i$ is

$$\Phi_{knm}(\gamma_{knm}^{(i)}) = w_n \log(1 + \gamma_{knm}^{(i)}), \tag{8.1}$$

where w_n is the bandwidth of RB n and $\gamma_{knm}^{(i)}$ is the instantaneous SINR for user m from SN k when using RB n. The superscript $i \in \{1, 2, 3\}$ indicates the set of RBs to which RB n belongs. For $i \in \{1, 2\}$ we have

$$\gamma_{lnm}^{(i)} = \frac{p_{ln} h_{lnm}}{\sum_{l' \in \mathcal{L}, l' \neq l} p_{l'n} h_{l'nm} + \sigma^2}, \tag{8.2}$$

where p_{ln} denotes the transmit power of SCBS l over RB n. Moreover, m corresponds to an arbitrary UE and SUE, respectively, for $i = 1$ and $i = 2$. For the transmissions over $\mathcal{N}_3$, the SINR is given by:

$$\gamma_{m_s nm}^{(3)} = \frac{p_{m_s n} h_{m_s nm}}{\sum_{m'_s \in \mathcal{M}_s, m'_s \neq m_s} p_{m'_s n} h_{m'_s nm} + \sigma^2}, \tag{8.3}$$

where $p_{m_s n}$ denotes the transmit power of SUE m_s over RB n and σ^2 is the variance of the receiver's Gaussian noise.

Given this model, one important problem is how to allocate the bandwidth resources to the wireless users. As discussed in Section 8.1, beyond power allocation and interference management techniques, we can boost the capacity of wireless networks by making the network better informed of its environment. Recent studies [171, 173, 174] have shown that friends in social networks, e.g. Facebook, have many common interests and activities that define their so-called *social tie*. Such social ties' strength could properly show how frequently people interact with their friends, share popular videos or pictures, or invite one another to activities of common interest. Therefore, such interrelationships can explain how often socially connected people request common contents [40, 51–54]. Observing such behavior is interesting for SCN resource allocation, since it motivates the possibility of serving a user directly by other users with shared interests over D2D communication, instead of requesting the content from SCBSs. Therefore, such scheme allows the network to decrease redundant transmissions and offload this traffic from the backhaul network. In fact, we are investigating a *content distribution* model that allows the network to use certain devices as SUEs, to serve as "caching points" whose storage can be used to cache popular content via overlay D2D. Therefore, our model is not a classical cooperative communication or relaying system.

For example, consider the scenario shown in Fig. 8.2, where a group of friends, e.g. students in a dorm or coworkers of a company spend a significant amount of time each day in neighboring

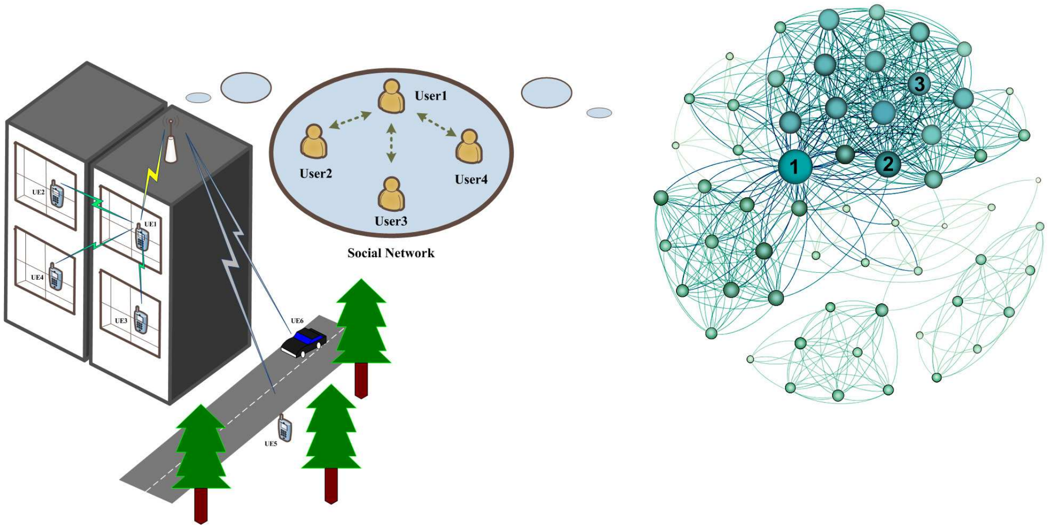


Figure 8.2: A schematic of a D2D enabled SCN which exploits the context information underneath the social network. The network in the right hand side shows the Facebook friendship graph of user 4.

rooms. Beyond this physical closeness, there is a social relationship between users at a higher layer that can underline their common interests in various topics such as sports or media. Since these users might have mutual interests, they are likely to be interested in common contents. Hence, although the formation of social ties is an application oriented metric, however, it strongly impacts how they access their wireless services, thus, directly impacting resource allocation. One illustrative example is the case in which one user, say user 1, shares a certain video on the social network, which, in turn, will be viewed by some of its friends, due to the mutual interests. Hence, those users could be served using data that may be cached at UE 1, directly through the D2D link. Clearly, by knowing the social ties between the users, the network will, on the one hand, be able to avoid multiple transmission of the same data and, on the other hand, will be able to allocate additional resources to the users outside the social group. In this work, we use the term *traffic offload* to refer to the reduction of redundant transmissions from SCBSs to UEs that can be obtained by exploiting D2D links between SUEs and UEs. Such traffic offload will alleviate the traffic on the backhaul-constrained SCBSs while also allowing to service additional users over the SCBSs' RBs [41].

With this in mind, we propose to exploit, jointly with conventional channel information, the users' social interrelationships in order to optimize resource allocation in D2D-enabled SCNs. The resource allocation can be posed as an optimization problem in which RBs are assigned to UEs ($\xi^\star : \mathcal{N} \to \mathcal{M}$) such that the overall sum utility of the network is maximized. Taking the social context into account, we can formulate the problem as

$$\underset{\xi^\star}{\operatorname{argmax}} \sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}} \xi_{knm}\varOmega_m(\varPhi_{knm}(\gamma_{knm}),\boldsymbol{Z}), \tag{8.4}$$

$$\text{subject to} \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} \xi_{knm} \leq 1, \quad \forall n \in \mathcal{N}, \tag{8.5}$$

$$\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}} \xi_{knm} \leq 1, \quad \forall m \in \mathcal{M}, \tag{8.6}$$

$$\xi_{knm} \in \{0,1\}, \tag{8.7}$$

where $\boldsymbol{Z}$ is a matrix that captures the social tie strength between every user pair and will be formally defined in Section 8.3. Moreover, $\varOmega_m(.)$ is the utility of user m which is a function of achievable rates and social ties. If user m is connected to an SCBS, $\varOmega_m(.)$ simply represents the achievable rate of the link. If user m is connected to an SUE, $\varOmega_m(.)$ is the sum of the link's achievable rate, plus a term that determines how much user m is socially connected to the cluster. The optimization problem in (8.4) aims to maximize the sum utility of all users. The constraint in (8.5) ensures that each RB is assigned to only one user, and the constraint in (8.6) ensures that each user is assigned to one RB.

Due to the unplanned deployment of backhaul-constrained SCBSs and the limited possibilities for SCBS coordination [5], our goal is to develop a *self-organizing, decentralized* resource allocation solution. This decentralized solution for the problem in (8.4)-(8.7) will be addressed in depth in Section 8.4. Before doing so, we formally define the social tie strength in the next Section and explain how such context information could be extracted from the social networks.

8.3 Modeling Relationship Strength in Social Networks

8.3.1 Social context in the proposed SCN model and SUE choice

Let z_{ij} denote the social tie strength between two UEs i and j. We define the social tie as a metric that determines how strong the relationship of two users is as inferred from the social network. This metric should then be incorporated into a proper utility function to be used in the context-aware resource allocation problem in (8.4). In order to benefit from caching at the edge, popular contents must be cached at UEs that are chosen to serve as SUEs. Here, we assume that a user is chosen as SUE m_s if its total social influence $I_{m_s} = \sum_{m\in\mathcal{M},m\neq m_s} z_{m_s m}$ is larger than other users[1]. I_{m_s} can be interpreted as a weighted degree of m_s in a social network graph where the edge weights are determined by the z term. We note that network operators need to provide some form of reward and incentive mechanisms to their users so that they act as SUEs. We can now define the notion of a *social cluster*

[1] Without loss of generality, other approaches for selecting SUE can also be accommodated.

Definition 15. A *social cluster (SC)* is defined as a set, $\mathcal{C}_{m_s}$, composed of an SUE m_s and all the UEs which are connected to m_s via D2D links.

We use the term *social* here to emphasize that the social relationships of the users affect the formation of the cluster, as will be elaborated in Section 8.4. Due to the social effects, we can make the following observations: 1) a UE m is encouraged by its friends to join the same SC in order to form socially stronger clusters, 2) SUEs with larger clusters (more assigned UEs) must get higher quality links from the $\mathcal{N}_2$ set, since the quality of the link from SCBS to SUE indirectly affects the quality of the direct D2D links, and 3) to improve offloading, SCBSs have an incentive to encourage UEs, with at least one friend as SUE, to use D2D links.

Such *peer effects* motivate the need for an advanced model that can accurately define the strength of ties, z_{ij}. In [175], a graphical model is proposed to learn the strength of ties among a set of Facebook users which is suitable for our model. In particular, based on the homophily property, it is observed that the stronger the tie, the higher the similarity [174]. Therefore, if two users have more attribute similarities in their profiles, e.g., the common groups that two UEs are members of, or the geographical locations, then their relationship is stronger. The relationship strength can be modeled as a *hidden effect* of profile similarities and inferred via statistical learning concepts [173, 174]. In the following, we review a learning model based on [175] that allows to understand how we can find the strength of ties, z from a given social network dataset.

8.3.2 Learning model

We note that the strength of the social relationship between two user impacts the nature and frequency of online interactions between a pair of users. Moreover, users naturally invest more of their resources (e.g., time) to build and maintain the relationships that they deem more important [175]. Hence, as the relationship becomes stronger, it is more likely that a certain type of interaction will take place between the pair of users. In this way, we can model the relationship strength as the hidden cause of user interactions.

Formally, let $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ be, respectively, the attribute vectors of two UEs i and j. An attribute vector is a vector that includes some of a user's social profile information, such as the user's age, political view, major, or the level of education. The relationship strength between i and j can be defined as a latent variable z_{ij} which will be inferred for every pair of users. In addition, let $\boldsymbol{y}_{ij}$ be the vector of interactions whose elements $y_{ij,f}, f = 1, ..., F$ are the F different interactions considered between i and j. Essentially, the interactions include the activities of UE i that involves UE j, e.g., tagging one another or posting on each others' walls. This variable is assumed to be binary such that $y_{ij,f} = 1$ if this interaction has occurred between UE i and UE j and $y_{ij,f} = 0$, otherwise. Furthermore, the vector $\boldsymbol{e}_{ij,f} = [e^1_{ij,f}, ..., e^{\vartheta}_{ij,f}]^T$ is defined for each interaction f occurred between users i and j. This vector can show how much user j is important for user i to interact. For instance, if user i has tagged user j and assuming $\vartheta = 1$, then $e^1_{ij,f}$ can be the overall number of users that user i usually tags. Thus, smaller $e^1_{ij,f}$ implies stronger tie between users i and

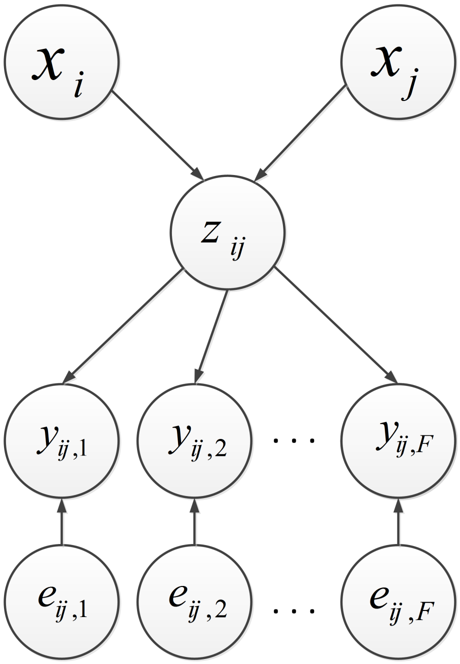


Figure 8.3: The graphical representation of the social tie strength model [175].

j. This social model can be represented by a directed graphical model as shown in Fig. 8.3. In this model, z_{ij} summarizes the profile similarities and interactions between users i and j. However, it is not observable from users' profiles. Hence, we need to estimate z_{ij} so as to maximize the overall observed data likelihood. To this end, the joint distribution of z and $\boldsymbol{y}$ can be represented using general factorization:

$$P(z_{ij}, \boldsymbol{y}_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j) = P(z_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j) \prod_{f=1}^{F} P(y_{ij,f}|z_{ij}). \tag{8.8}$$

In order to infer the latent variables, we need to adopt the conditional probability of the relationship strength given the attribute similarities, i.e., $P(z_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)$. In this regard, we consider the widely used Gaussian distribution [175]

$$P(z_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathcal{N}(\boldsymbol{w}^T \zeta(\boldsymbol{x}_i, \boldsymbol{x}_j), \upsilon), \tag{8.9}$$

where the similarity vector $\zeta(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the set of similarity measures taken on the pair of users (i, j) and $\boldsymbol{w}$ denotes the vector of parameters of the model in (8.9).

Now, in order to completely describe the joint distribution in (8.8), the conditional probability of $y_{ij,f}$ given z_{ij} and $e_{ij,f}$, can be modeled by using the logistic function:

$$P(y_{ij,f} = 1|\boldsymbol{u}_{ij,f}) = \frac{1}{1 + e^{-(\varrho_f^T \boldsymbol{u}_{ij,f})}}, \tag{8.10}$$

where $\boldsymbol{u}_{ij,f} = [\boldsymbol{e}_{ij,f}, z_{ij}]^T$. Moreover, $\boldsymbol{\varrho}_f = [\varrho_{f,1}, \varrho_{f,2}, ..., \varrho_{f,\vartheta+1}]^T$ are the parameters of the model in (8.10) that must be estimated. From Fig. 8.3 and given latent variable z_{ij}, all elements of $\boldsymbol{y}_{ij}$ become independent of each other. Let $\mathcal{D} = \{(i_1, j_1), (i_2, j_2), ..., (i_D, j_D)\}$ be the set of user sample pairs observed from the network. The variables $\boldsymbol{x}_i$, $\boldsymbol{x}_j$, $\boldsymbol{y}_{ij}$ and $\boldsymbol{e}_{ij,f}$ could all be extracted from the social network. Hence, conditioned to the attribute similarities and model parameters, we can write (8.8) as:

$$\begin{aligned}&P(z_{ij}, \boldsymbol{y}_{ij} | \boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{w}, \boldsymbol{\varrho}) = \\ &\prod_{(i,j)\in\mathcal{D}} \left(P(z_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{w}) \prod_{f=1}^{F} P(y_{ij,f}|z_{ij}, \boldsymbol{\varrho}_f) \right).\end{aligned} \tag{8.11}$$

Here, by substituting (8.9) and (8.10) in (8.11), the joint distribution can be written as:

$$\begin{aligned}&P(z_{ij}, \boldsymbol{y}_{ij} | \boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{w}, \boldsymbol{\varrho}) \propto \\ &\prod_{(i,j)\in\mathcal{D}} \left(e^{-\frac{1}{2\upsilon}\left(\boldsymbol{w}^T\zeta_{ij} - z_{ij}\right)^2} \prod_{f=1}^{F} \frac{e^{-(\boldsymbol{\varrho}_f^T \boldsymbol{u}_{ij,f})(1-y_{ij,f})}}{1 + e^{-(\boldsymbol{\varrho}_f^T \boldsymbol{u}_{ij,f})}} \right).\end{aligned} \tag{8.12}$$

8.3.3 Inference

Given the defined learning model, we can now infer the social tie strength between each arbitrary pair of users (i, j). One way to estimate z_{ij} is to adopt the approach of [175] in which z_{ij} is treated as a parameter. Essentially, we can find the point estimates $\hat{\boldsymbol{w}}, \hat{\boldsymbol{\varrho}}, \hat{z}$ that maximize the likelihood $P(y, \hat{z}, \hat{\boldsymbol{w}}, \hat{\boldsymbol{\varrho}}|x)$. To avoid overfitting the training dataset for the model in (8.9) and (8.10), regularizers λ_w and λ_ϱ will be used respectively for the parameters $\boldsymbol{w}$ and $\boldsymbol{\varrho}$ with Gaussian priors. Overfitting can occur if the size of the $\boldsymbol{w}$ vector is too large for the observed data from the attribute vector $\boldsymbol{x}$. Using (8.12), we have:

$$\begin{aligned}&P(z_{ij}, \boldsymbol{y}_{ij}, \boldsymbol{w}, \boldsymbol{\varrho} | \boldsymbol{x}_i, \boldsymbol{x}_j) = \\ &P(z_{ij}, \boldsymbol{y}_{ij} | \boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{w}, \boldsymbol{\varrho}) P(\boldsymbol{w}, \boldsymbol{\varrho} | \boldsymbol{x}_i, \boldsymbol{x}_j),\end{aligned} \tag{8.13}$$

and since the model parameters $\boldsymbol{\omega}$ and $\boldsymbol{\varrho}$ are independent of one another, as well as of the attributes of the users, we can write the joint conditional distribution in (8.13) as

$$P(z_{ij}, \boldsymbol{y}_{ij}, \boldsymbol{w}, \boldsymbol{\varrho} | \boldsymbol{x}_i, \boldsymbol{x}_j) = P(z_{ij}, \boldsymbol{y}_{ij} | \boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{w}, \boldsymbol{\varrho}) P(\boldsymbol{w}) P(\boldsymbol{\varrho}). \tag{8.14}$$

and hence,

$$
\begin{aligned}
\log P(z_{ij}, \boldsymbol{y}_{ij}, \boldsymbol{w}, \boldsymbol{\varrho} | \boldsymbol{x}_i, \boldsymbol{x}_j) = &\sum_{(i,j)\in\mathcal{D}} \left(-\frac{1}{2\upsilon} \left(\boldsymbol{w}^T \zeta_{ij} - z_{ij}\right)^2 \right) + \\
&\sum_{(i,j)\in\mathcal{D}} \left(\sum_{f=1}^{F} \left(-(1 - y_{ij,f})(\boldsymbol{\varrho}_f^T \boldsymbol{u}_{ij,f}) - \log\left(1 + e^{-(\boldsymbol{\varrho}_f^T \boldsymbol{u}_{ij,f})}\right)\right)\right) \\
&- \frac{\lambda_w}{2} \boldsymbol{w}^T \boldsymbol{w} - \sum_{f=1}^{F} \frac{\lambda_\varrho}{2} \boldsymbol{\varrho}_f^T \boldsymbol{\varrho}_f + C,
\end{aligned}
\tag{8.15}
$$

where C is a constant. (8.15) is a concave function and, thus, the latent variable and parameters can be derived by using gradients of this function. Using the iterative Newton-Raphson algorithm, we can maximize the function in (8.15) and find the optimum values for z_{ij}, $\boldsymbol{\varrho}_f$, and $\boldsymbol{w}$. We denote $\boldsymbol{Z}$ as a $M \times M$ matrix, where z_{ij} is the element of i-th row and j-th column.

Given the proposed wireless model of Section II along with the inferred social tie metric $\boldsymbol{Z}$, we are now able to account for the social interconnections among users in conjunction with the physical layer constrains of the D2D-enabled SCN. We next develop a matching-based approach to allocate resources in a social context aware SCN.

8.4 Context-Aware Resource Allocation As a Matching Game

Having defined the social context, our next goal is to solve the resource allocation problem in (8.4). The problem given by (8.4), subject to (8.5)-(8.7), is a 0-1 integer programming, which is a satisfiability problem as such one of Karp's 21 NP-complete problems [176]. Hence, it is difficult to solve this problem via classical optimization approaches. Moreover, for a large-scale SCN network with D2D communication, it is desirable to solve the context-aware resource allocation problem in (8.4)-(8.7) using a decentralized, self-organizing approach in which the SCBSs and devices can interact and make resource allocation decisions based on their local information without relying on a centralized entity for coordination. In addition, owing to the need for exploiting context information, it is of interest to define individual SN or UE utilities that capture the locally available context at each node.

To this end, we develop a decentralized resource management framework based on matching theory [164, 166, 167, 169, 177] to solve the proposed problem. The main benefit of matching theory is its ability to define individual utilities per UE and SNs as well as the available algorithmic implementations that allow to provide a largely decentralized and self-organizing solution to the resource allocation problem in (8.4)-(8.7) while accounting for all the nodes' information. In the studied context-aware model, the preference relation can be based on a variety of metrics related to the social and wireless realms. In this regard, we formulate the proposed resource allocation problem in SCNs as a two-sided one-to-one matching game in which each SN-controlled RB $n \in \mathcal{N}$ will be assigned to at most one user $m \in \mathcal{M}$ and vice versa. Therefore, we can formulate the problem as a one-to-one matching game given by the tuple $(\mathcal{M}, \mathcal{N}, \succ_{\mathcal{M}}, \succ_{\mathcal{N}})$. Here, $\succ_{\mathcal{M}} = \{\succ_m$

$\}_{m\in\mathcal{M}}$ and $\succ_{\mathcal{N}}=\{\succ_n\}_{n\in\mathcal{N}}$ denote, respectively, the set of the preference relation of users and RBs. We formally define the notion of a matching:

Definition 16. A *matching* μ is defined as a function from the set $\mathcal{M}\cup\mathcal{N}$ into the set of $\mathcal{M}\cup\mathcal{N}$ such that $m=\mu(n)$ if and only if $\mu(m)=n$.

Let $V_m(.)$ and $U_n(.)$ denote, respectively, the utility function of UE m and RB n. Given these utilities, we can say that a user m prefers RB n_1 to n_2, if $V_m(n_1)>V_m(n_2)$. This preference is denoted by $n_1\succ_m n_2$. Similarly, an RB n prefers UE m_1 to m_2, if $U_n(m_1)>U_n(m_2)$ and this is denoted by $m_1\succ_n m_2$.

8.4.1 Users' preferences

Depending on whether an RB n is offered by an SUE or an SCBS, the UEs will have different utility functions. Moreover, the utility of one player may depend on the matching of other players, due to the peer effects described in Section 8.3.1. In this regard, let $a_{mm_s;\mu}=\{a_{mm_s}|\mu\}$ be a variable used to determine the existence of a D2D link between UE m and SUE m_s, conditioned on the current matching μ. Similarly, we use $a_{mm_s;\mu\mu'}=\{a_{mm_s}|\mu,\mu'\}$, to indicate whether a certain UE m is a member of $\mathcal{C}_{m_s}$ in both matchings μ and μ'. The motivation for this definition will be explained in Section 8.5.2. Before computing the utilities, UEs first obtain the corresponding channel response of each RB, from all SNs and form a $K\times N$ channel coefficient matrix $\boldsymbol{H}_m$. Using $\boldsymbol{H}_m$, each user $m\in\mathcal{M}$ shapes its achievable data rate matrix, i.e. $\boldsymbol{\Phi}_m$, whose elements are given by (8.1)-(8.3). From (8.1) and (8.2), the utilities of a UE and an SUE, respectively, for RBs $n_1\in\mathcal{N}_1$ and $n_2\in\mathcal{N}_2$, are given by:

$$V_{mn_i,l}(\gamma_{lnm}^{(i)})=w_{n_i}\log\left(1+\gamma_{lnm}^{(i)}\right), \tag{8.16}$$

where m corresponds to a UE ($i=1$) or an SUE ($i=2$). From (8.3), the utility of a user m using an RB $n\in\mathcal{N}_3$ is given by:

$$\begin{aligned} &V_{mn,m_s}(\boldsymbol{Z},\gamma_{m_snm}^{(3)},\mu)= \\ &w_n\log\left(1+\gamma_{m_snm}^{(3)}\right)+\alpha_m\left(z_{mm_s}+\sum_{j\in\mathcal{M}_u\backslash m}a_{jm_s;\mu\mu'}z_{mj}\right), \end{aligned} \tag{8.17}$$

where α_m is a weighting parameter that allows to control the impact of the social ties on the overall decision of the user. The utilities given by (8.16) state that UEs and SUEs, respectively, rank RBs $n_1\in\mathcal{N}_1$ and $n_2\in\mathcal{N}_2$ based only on the achievable rates. However, (8.17) captures the peer effects on the UEs' preferences for RBs $n\in\mathcal{N}_3$. A UE can benefit more from the mutual interests among the SC members, if the SC members are more socially connected with one another and with the UE. The second term in (8.17) implies that the UE prefers to join an SC that has stronger ties with it.

8.4.2 SN-based RBs' preferences

The proposed matching game can be fully represented once the preference of each RB is defined. The decision of an RB $n \in \mathcal{N}$ is mainly controlled by the SCBS or SUE that is using it. Hence, in the proposed game, SCBSs and SUEs make decisions on behalf of their RBs. In the considered model, there are three groups of RBs, each of which has a different preference over the UEs. Similar to Section 8.4.1, we define a novel scheme at the RB side of the game, which is based on the information extracted from the corresponding social network. For the transmission between SCBS $l \in \mathcal{L}$ and UE $m \in \mathcal{M}_u$, over RB $n \in \mathcal{N}_1$, the utility of RB $n \in \mathcal{N}_1$ when choosing UE m is given by:

$$U_{nm,l}(\boldsymbol{Z}, \gamma_{lnm}^{(1)}) = w_n \log\left(1 + \gamma_{lnm}^{(1)}\right) - \beta_n\left(\sum_{j \in \mathcal{M}_s} z_{mj}\right). \tag{8.18}$$

Hence, $m_1 \succ_n m_2$ if and only if $U_{nm_1,l} > U_{nm_2,l}$. β_n is a weighting parameter that controls the impact of the social context. The second term in the right hand side of (8.18) implies that RBs $n \in \mathcal{N}_1$ give less utility to UEs who can be served by an SUE.

In addition, the utility achieved by RB $n \in \mathcal{N}_2$ when selecting SUE m_s, $U_{nm_s,l}$ is given by

$$U_{nm_s,l}(\boldsymbol{Z}, \gamma_{lnm_s}^{(2)}, \mu) = w_n \log\left(1 + \gamma_{lnm_s}^{(2)}\right) + \nu_n \cdot X_{m_s}, \tag{8.19}$$

where $X_{m_s} = \sum_{m \in \mathcal{M}_u} a_{mm_s;\mu} z_{mm_s}$ is the total cumulative social tie strength of a particular SUE m_s and ν_n is a weighting parameter that controls the importance of the SC that the SUE has formed in RB n's utility. The utility function in (8.19) promotes SUEs with higher social ties since they are likely to form larger SCs. Finally, the utility of RB $n \in \mathcal{N}_3$ for user m via D2D link from SUE m_s is given by

$$U_{nm,m_s}(\gamma_{m_snm}^{(3)}) = w_n \log\left(1 + \gamma_{m_snm}^{(3)}\right) + \kappa_n z_{mm_s}. \tag{8.20}$$

The utility function in (8.20) implies that UEs must be accepted based on both quality of the D2D link and social context. κ_n is a weighting parameter and implies that how important the role of social tie is for RB $n \in \mathcal{N}_3$ of an SUE $m_s \in \mathcal{M}_s$ in accepting a UE $m \in \mathcal{M}_u$.

8.5 Proposed Context-Aware Resource Allocation Algorithm

Given the formulated context-aware matching game, our goal is to find a *stable matching*, which is one of the key solution concepts in matching theory [177]. Let $\mathcal{A}(\mathcal{M}, \mathcal{N})$ denote the set of all possible matchings, and $\mu(m, n)$ denote a subset of $\mathcal{A}(\mathcal{M}, \mathcal{N})$, where m and n are matched together. Then, we define a stable matching as follows:

Definition 17. A pair $(m, n) \notin \mu$, where $m \in \mathcal{M}, n \in \mathcal{N}$ is said to be a *blocking pair* for the matching μ, if there is another matching $\mu' \in \mu(m, n)$, where $\mu' \succ_m \mu$ and $\mu' \succ_n \mu$. A matching μ^* is *stable* if and only if there is no blocking pair.

A stable matching solution for resource allocation problem in (8.4)-(8.7) ensures that after allocating resources to the users, no RB-UE or RB-SUE pair in SCN would benefit from replacing their current association with a new link. That is, no user can benefit by changing its assigned frequency resource and vice versa. For the proposed context-aware resource allocation game, we can make the following observation:

Remark 5. *The proposed SCN matching game has* peer *effects.*

The RBs and users in a context-aware resource allocation game may change their preferences as the game evolves. That is, the preference of one player may depend on the preferences of the other players. For instance, the peer effects introduced in Section 8.3.1 make the preferences of RBs and users interdependent, due to the social interrelationships among users. This type of game is known as the *matching game with peer effects*, in which players have preference ordering over the set of all possible matchings $\mathcal{A}(\mathcal{M}, \mathcal{N})$ [178]. This is in contrast with the traditional matching games in which players have fixed preference ordering [166–169, 177].

For traditional matching games such as in [166–169, 177], one can use the deferred acceptance algorithm, originally introduced in [72], to find a stable matching. However, such an algorithm may not be able to converge to a stable matching when the game has peer effects [177], such as in the proposed context-aware resource allocation model. Therefore, there is a need to develop new algorithms, that significantly differ from existing applications of matching theory in wireless such as [166–169], so as to find the solution of the studied matching game.

8.5.1 Proposed socially-aware resource allocation algorithm

To solve the formulated matching game, we propose a novel algorithm for resource allocation in D2D-enabled SCNs. Table 8.1 shows the various stages of this proposed socially-aware resource allocation (SARA) algorithm that allows to solve the SCNs' matching game.

The proposed algorithm is composed of four main stages: Stage 1 includes the matching of SUEs with RBs $n \in \mathcal{N}_2$, Stage 2 focuses on the matching of UEs with RBs $n \in \mathcal{N}_1 \cup \mathcal{N}_3$, Stage 3 focuses on updating the SC information, and Stage 4 during which the actual downlink transmission occurs. Initially, the SCBSs use the knowledge of social ties among users to choose the SUEs as discussed in 8.3.1. Each SCBS sends a proposal to its neighboring users that are deemed influential enough to be an SUE. UEs accept or reject the proposals and the SCBSs broadcast the set of SNs, $\mathcal{K}$, and the sets of RBs, $\mathcal{N}_i, i = 1, 2, 3$.

After initialization, each SUE applies for $n \in \mathcal{N}_2$ based on (8.16) and each RB accepts the most preferred UE and rejects other proposals based on the utilities defined in (8.19). Stage 1 terminates once each SUE is accepted by an RB or rejected by all its preferred RBs. This matching remains unchanged until SUEs update the cluster $\mathcal{C}_{m_s}, \forall m_s \in \mathcal{M}_s$, sets. However, the new context information changes the preferences of the RBs for SUEs based on (8.19). That is due to the fact that $X_{m_s} = \sum_{m \in \mathcal{M}_u} a_{mm_s;\mu} z_{mm_s}$ determines how much the members of an SUE's cluster are

Table 8.1: Proposed Social Context-Aware Resource Allocation Algorithm

```
**Inputs:** $\mathcal{L}, \mathcal{M}, \mathcal{N}, \boldsymbol{H}, \boldsymbol{Z}$
  *Initialize:* SCBSs send proposal to high influential UEs to act as SUE. SCBSs broadcast the set $\mathcal{K}$ and announce the set of available RBs in $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$. Initialize the set of $\mathcal{C}_{m_s}$ for each SUE as an empty set.
  *Stage 1:*
      (a) SUEs determine their preference ordering for RBs $n \in \mathcal{N}_2$, using (8.16).
      (b) RBs $n \in \mathcal{N}_2$ calculate utility of each SUE applicant using (8.19) for the current state of the matching.
      (c) SUEs apply for RBs $n \in \mathcal{N}_2$ and get accepted or rejected via the deferred acceptance algorithm.
  *Stage 2:*
      (a) UEs apply for RBs $n \in \mathcal{N}_1$ and $n \in \mathcal{N}_3$, using (8.16) and (8.17), respectively.
      (b) RBs $n \in \mathcal{N}_1$ and $n \in \mathcal{N}_3$ calculate utility of each UE applicant using (8.18) and (8.20), respectively.
      (c) UEs get accepted or rejected by RBs $n \in \mathcal{N}_1 \cup \mathcal{N}_3$ through deferred acceptance algorithm.
  *Stage 3:*
      (a) Update SC information, $\mathcal{C}_{m_s}$ for $\forall m_s \in \mathcal{M}_s$.
      (b) SUEs broadcast SC information of the current matching, i.e., $a_{mm_s;\mu}$ coefficients to their nearby UEs.
**while** $\mathcal{C}_{m_s}, \forall m_s \in \mathcal{M}_s$ remain unchanged for two consecutive matchings
*repeat Stage 1 to Stage 3*
  *Stage 4:*
      (a) For any cluster member, SUE determines if the current requested data exists in its directory.
      (b) Actual downlink transmission of data occurs from each RB to its matched SUE or UE.
**Output:** Stable matching $\mu^*$
```

socially connected. A larger X_{m_s} implies that the members can benefit more from D2D due to common interests in their requested data. Therefore, RBs $n \in \mathcal{N}_2$ prefer to be matched to an SUE with larger X_{m_s}. Due to the change in their preference ordering, SUEs and RBs $n \in \mathcal{N}_2$ need to repeat this stage once the context information is updated.

Following the first stage, UEs apply for $n \in \mathcal{N}_1$ or $n \in \mathcal{N}_3$, based on the utilities defined in (8.16) and (8.17), respectively. The SCBSs and SUEs controlling RBs $n \in \mathcal{N}_1$ and $n \in \mathcal{N}_3$, accept the UE that gives the higher utility based, respectively, on (8.18) and (8.20), and reject the rest of the applicants. As long as $\mathcal{C}_{m_s}, \forall m_s \in \mathcal{M}_s$, do not change, UEs have strict preference over RBs $n \in \mathcal{N}_1 \cup \mathcal{N}_3$ and vice versa. Stage 2 ends once each UE is accepted by one RB or rejected by all

RBs of its preference list.

All clusters are subject to change due to the peer effects in the matching game. Thus, players need to update their preferences based on the new $\mathcal{C}_{m_s}, \forall m_s \in \mathcal{M}_s$, information resulted from Stage 2. In Stage 3, each SUE m_s updates the SC information, i.e. $\mathcal{C}_{m_s}, m_s \in \mathcal{M}_s$, based on the results of the current matching and broadcasts $\mathcal{C}_{m_s}$ set to its nearby UEs and the corresponding SCBS. According to this information, players sort their preferences conditioned on the current matching. The algorithm terminates, once the $\mathcal{C}_{m_s}, m_s \in \mathcal{M}_s$ sets do not change for two consecutive matchings. In the final stage, once the matching is complete, the downlink transmission of the UEs occurs, using the allocated resource blocks. This stage is essentially the actual communication stage in the D2D-enabled SCN.

8.5.2 Convergence and stability of the proposed algorithm

In this Subsection, we prove the stability of the algorithm proposed in Table 8.1. Prior to doing so, we make the following definition:

Definition 18. Given the social interrelationship between UEs, an SC $\mathcal{C}_{m_s}$ is said to be *S-stable*, if both of the following conditions are satisfied:
1) No UE m outside the cluster $\mathcal{C}_{m_s}$ can join it. That is, for any $m \notin \mathcal{C}_{m_s}$ and $n \in \mathcal{N}_3$ belonging to m_s, there is no pair $(m, n) \notin \mu$ where $m \succ_n \mu(n)$ and $n \succ_m \mu(m)$.
2) No UE m inside the $\mathcal{C}_{m_s}$ can leave the cluster. That is, for any $m \in \mathcal{C}_{m_s}$ and $n \in \mathcal{N}_1 \cup \mathcal{N}_3$ that does not belong to m_s, there is no pair $(m, n) \notin \mu$ where $m \succ_n \mu(n)$ and $n \succ_m \mu(m)$.
A matching is S-stable, if and only if all the clusters are S-stable.

This notion of stability guarantees that the peer effects cannot make a UE outside the SCs join a cluster. In addition, the UEs inside SCs will not leave or change their clusters. However, it is not sufficient to ensure the required two-sided stability of the matching. Next, we discuss a property of the proposed game and show why the S-stability of SCs is non-trivial.

Proposition 5. Given the information on the social clusters, once a UE m is accepted by an RB of a particular SC, $\mu(m)$ will not reject m in favor of any new applicant. However, this does not imply UE m has no incentive to leave the cluster.

Proof. See Appendix E.1. □

Based on the Proposition 5, we can show the S-stability of the proposed context-aware resource allocation game as follows:

Theorem 8. *Each SC becomes S-stable after a finite number of iterations and, thus, the proposed algorithm in Table* 8.1 *is guaranteed to converge.*

Proof. See Appendix E.2. □

Given the results in Proposition 5 and Theorem 8, we can now state the main result with regard to the two-sided stability of the matching.

Theorem 9. *The proposed algorithm in Table* 8.1 *is guaranteed to reach a two-sided stable matching between users and RBs.*

Proof. See Appendix E.3. □

8.5.3 Complexity analysis of the proposed algorithm

In order to analyze the computational complexity of the proposed algorithm, we can start by investigating the simple case in which the matching game has no peer effect, i.e., users and RBs have strict preference ordering. Here, we consider two cases: 1) when the number of UEs is less than the total number of RBs, i.e., $M_u \leq N_T$, where $N_T = N_1 \times L + N_3 \times M_s$ and 2) when the number of UEs is greater than the total number of RBs, i.e., $M_u > N_T$. Our goal is to analyze the worst case scenario, i.e., the maximum number of iterations and the maximum number of matching proposals sent from UEs to either SCBSs or SUEs (which relate to the messaging overhead). In each iteration, UEs send a proposal to their most preferred RB, and RBs receive the proposals, accept the most preferred one and reject the other UEs. Therefore, it is clear that the number of unmatched UEs at each iteration is equal or less than the number of unmatched UEs at previous iterations.

For the first case, once the algorithm converges, all the UEs are matched, since RBs prefer any UE to being unallocated. We can easily observe that the worst case happens, if all UEs have the same preference ordering. Hence, at the end of each iteration t, there are $M_u - t$ unmatched UEs. Therefore, the maximum number of iterations, t_{max}, is obtained when all the users are matched, i.e., $M_u - t_{max} = 0$. Hence, the complexity is of the order $\mathcal{O}(M_u)$. Furthermore, at each iteration t, $M_u - t + 1$ proposals are sent. Hence, the messaging overhead, S_{max}, is equal to:

$$S_{max} = \sum_{t=1}^{t_{max}} (M_u - t + 1) = \frac{M_u(M_u + 1)}{2}. \tag{8.21}$$

Similarly for the second case, we know that once the algorithm converges, there are exactly $M_u - N_T$ unmatched users. Again, the worst case happens, if all UEs have the same preference ordering. Hence, at each iteration, only one UE gets accepted. Therefore, the maximum number of iterations, t_{max}, is obtained when there are $M_u - N_T$ unmatched users, i.e., the complexity is of the order $\mathcal{O}(N_T)$. The messaging overhead is equal to:

$$S_{max} = \sum_{t=1}^{t_{max}} (M_u - t + 1) = \tag{8.22}$$

$$\sum_{t=1}^{N_T} (M_u - t + 1) = (M_u + 1)N_T - \frac{N_T(N_T + 1)}{2}.$$

As we see from the above equations, for $M_u < N_T$, the complexity of the matching algorithm increases linearly with the number of users. In addition, the messaging overhead exhibits quadratic increase with respect to the number of users. Moreover, for $M_u > N_T$, the upperbound for complexity is independent of the number of users.

The complexity of the proposed context-aware algorithm will further depend on the social matrix $\boldsymbol{Z}$. However, for a given SCs, the complexity of our algorithm follows the above analysis. From Theorem 1, we know that SCs change only for a finite number of iterations. Hence, we can anticipate that the overall complexity of the context-aware approach be linearly proportional to the complexity of the context-unaware approach.

8.6 Simulation Results and Analysis

8.6.1 Social context dataset and simulation parameters

For the evaluation of our results, we first use the learning model introduced in Section 8.3. We have computed the social tie matrix $\boldsymbol{Z}$, for a set of 80 users from the Facebook network. We have used the real dataset released by Stanford University [179], which is generated by surveying a number of volunteer Facebook users, known as ego nodes. For the selected ego node, the dataset contains 224 anonymized attributes for each user, including education, gender, location, language, and work, among others. In addition, the dataset specifies 32 anonymized circles and determines which subset of users are in which circle. Each circle is a group, composed of a subset of users, which can be thought as a high school institution or a company. For our simulations, we assume that if two users i and j are within at least one common circle, then they interact with each other on Facebook. Finally, in order to determine the tendency of user i to interact with another user j, we consider the degree of user i in the ego network, i.e., the number of friends of user i. The motivation behind this assumption can be explained as follows: if user i picks user j to interact with from a larger number of friends, this implies that user j is more important to user i than other users.

In order to select the SUEs, we choose the four users with the highest weighted degree from the ego network. Thus, in our simulations, we have $M_s = 4$ and the weights are determined by the social tie strength z for each pair of users. We consider $L = 7$ SCBSs distributed randomly within a square area of $2\,\text{km} \times 2\,\text{km}$. The number of active RBs for SCBS to UE link, SCBS to SUE link, and SUE to UE link, respectively, are $N_1 = 5, N_2 = 3$, and $N_3 = 5$, unless stated otherwise. Here, we would like to note that depending on the channel state information and social ties among users, spectrum partitioning can be done dynamically and the proposed model is not limited to any specific resource partitioning. In this work, however, we assume that throughout the resource allocation, neither the social tie matrix $\boldsymbol{Z}$, nor the channel state information are changing and therefore, the partitions are static. Each RB is composed of 12 consecutive subcarriers, each of which having a 15 KHz bandwidth, according to 3GPP Rel-12 Standard [180]. The transmit power of SCBSs and SUEs are set to 2 W and 10 mW, respectively. The wireless channel experiences Rayleigh fading, with the propagation loss set to 3. The receivers' noise is assumed Gaussian with zero mean and with variance equals to -90 dBm. The weighting parameters, α_m, β_n, ν_n and κ_n are set to half of the RB bandwidth. Throughout the simulations, the unmatched users are assigned a zero utility. All statistical results are averaged over a large number of independent runs for different locations and channel gains.

For comparison purposes, we compare our proposed approach with two centralized approaches: 1) the context-aware centralized solution which aims to maximize the overall utilities of all UEs and RBs, 2) the context-unaware centralized approach, that maximizes the throughput of the users. In simulation results, the centralized solutions refer to the linear programming relaxation of the original 0-1 integer programming problem in (8.4)-(8.7) by letting $0 \leq \xi \leq 1$. To obtain centralized solutions in our simulations, we used the YALMIP toolbox of MATLAB. In addition, we compare our results with the context-unaware distributed algorithm that is based on the one-to-one matching game similar to our proposed algorithm, however, no social context is incorporated. That is, UEs and RBs rank one another only based on the maximum SINR values. This benchmark algorithm is in line with some existing works such as [167] and [166].

In addition, we show the effect of context-awareness by comparing the offloaded traffic of both approaches as one of the main performance metrics in our results. We define the offloaded traffic as the number of users who will be served directly by data that is cached in a directory or folder at the level of the SUE. The users obtain this offloaded traffic via D2D communications without having to use the SCN's infrastructure, due to the correlation in their requests for content as discussed in Section 8.3. To compute this offloaded traffic, we must find the probability $P_m(y_d = 1)$ of requesting content that already exists in the directory of SUE m_s by each UE $m \in \mathcal{C}_{m_s}$:

$$P_m(y_d = 1) = \sum_{d \in \mathcal{D}_{m_s}} P_m(y_d = 1 | d \in \mathcal{D}_{m_s}) P(d \in \mathcal{D}_{m_s}), \tag{8.23}$$

where d and $\mathcal{D}_{m_s}$ denote, respectively, the requested file and the set of files of the SUE m_s's directory. The Prior information, $P(d \in \mathcal{D}_{m_s})$ in (8.23), depends on both the history of requested files in previous time slots and on the mutual social tie between all members of the SC. Finding a closed-form solution for (8.23) to model the correlation between users in the time domain is difficult. Thus, for simplicity, we assume that requesting a file from the directory of an SUE can be modeled as an interaction y_d between the user and its SC set. Motivated by (8.10), we model $P_m(y_d = 1)$ as a function of cluster's average social tie $\bar{z}_{m_s} = \frac{1}{|\mathcal{C}_{m_s}|-1} \sum_{m \in \mathcal{C}_{m_s}} z_{mm_s}$, as follows

$$P_m(y_d = 1) = \frac{1}{1 + e^{(-\rho \bar{z}_{m_s})}}; \quad \forall m \in \mathcal{C}_{m_s}, \tag{8.24}$$

where ρ is a constant normalizing parameter. Equation (8.24) allows to relax the dependencies between users by considering average social tie of the cluster.

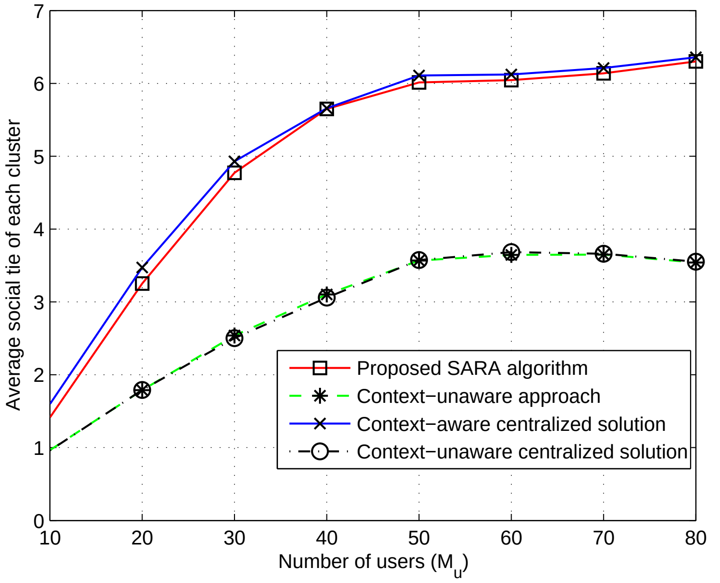


Figure 8.4: Comparison of the average social tie of SCs for proposed SARA algorithm with other three approaches.

8.6.2 Simulation results and discussions

In Fig. 8.4, we show the average social tie in the final clusters resulting from the proposed approach and the other three algorithms. Clearly, from this figure, we can observe that the members of the clusters in average are much more socially connected in SARA compared to the context-unaware approach. Fig. 8.4 shows that the average cluster social tie is up to 77% higher for the proposed SARA algorithm relative to the context-unaware scenario for $M_u = 70$ UEs. As the number of users increases and the SCBS resources become more scarce, UEs have no choice but to join the D2D clusters. However, Fig. 8.4 shows that by using the proposed SARA algorithm, the UEs will be more socially connected within their clusters and can benefit more from the social interconnections of one another.

In Fig. 8.5, the average sum rate of all users is compared for the SARA algorithm with the distributed context-unaware approach and centralized approaches. It is not surprising to see that the average sum rate of our proposed approach is slightly below that of an algorithm that is focused on optimizing only the data rate. Fig. 8.5 shows that the gap between the two algorithms does not exceed 3% for all network sizes. However, as will be shown in Fig. 8.6, this small loss in average rate will be compensated by having more offloaded traffic in the downlink of the SCN. Interestingly, the rate performance of the proposed approach is very close to the centralized solution and the gape between the two algorithms does not exceed 4% for all network sizes.

Fig. 8.6 compares the average offloaded traffic at each time slot for the proposed algorithm and other three approaches as the number of users varies. We note that for the context-unaware approaches, the contents stored in the SUE's directory do not depend on the average social tie of the social cluster. Considering $\rho = 0$ in (8.24) allows us to model this independence. The average

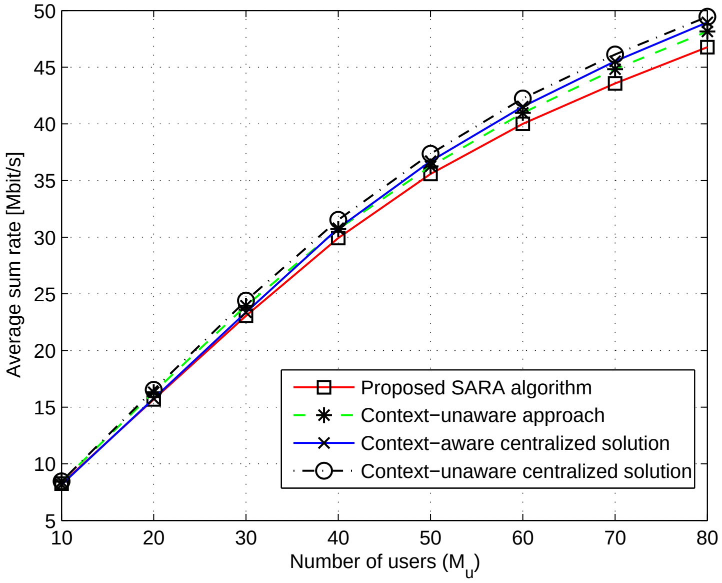


Figure 8.5: Comparison of the average sum rate of the proposed SARA algorithm and other three approaches.

offloaded traffic determines how many of UEs in average could be served directly by data that is cached in the SUEs' directories. The offloaded traffic increases with the number of users, since more UEs will move to the D2D tier. However, this metric will saturate for the large network sizes, since the number of RBs $n \in \mathcal{N}_3$ is limited. We can see that the proposed approach achieves a very close performance to the context-aware centralized solution, in terms of traffic offloads, for the network size with more than $M_u = 40$ UEs. Moreover, we observe that the performance of the context-unaware approach coincides with the central solution. This is primarily due to the fact that context-unaware approach follows the deferred acceptance algorithm which is known to achieve optimal solution for the proposing players (i.e. users) [72, 166]. In Fig. 8.6, we can see that the proposed SARA algorithm outperforms the context-unaware approach by increasing the offloaded traffic for different network sizes. Fig. 8.6 shows that the proposed SARA algorithm can offload up to 84% more traffic from the SCN's infrastructure when compared to the context-unaware approach for a network with $M_u = 70$ UEs.

In Fig. 8.7, we show the average offloaded traffic for different values of the ρ parameter in (8.24) and different number of users. One interesting observation from Fig. 8.7 is that the proposed SARA algorithm offloads more traffic as the network size grows and eventually saturates, since the number of RBs is fixed. In fact, Fig. 8.7 implies that, at the cost of a slight reduction in the average rate of the matched users (see Fig. 8.5), the proposed SARA algorithm allows to admit more users into the network. Hence, the proposed approach can be used to optimize the tradeoff between serving more users in a congested network and the average rate of current users. By properly setting the control parameters in the utility functions, this tradeoff can be balanced according to the traffic of the network.

In Fig. 8.8, we show the average offloaded traffic resulting from the proposed SARA algorithm

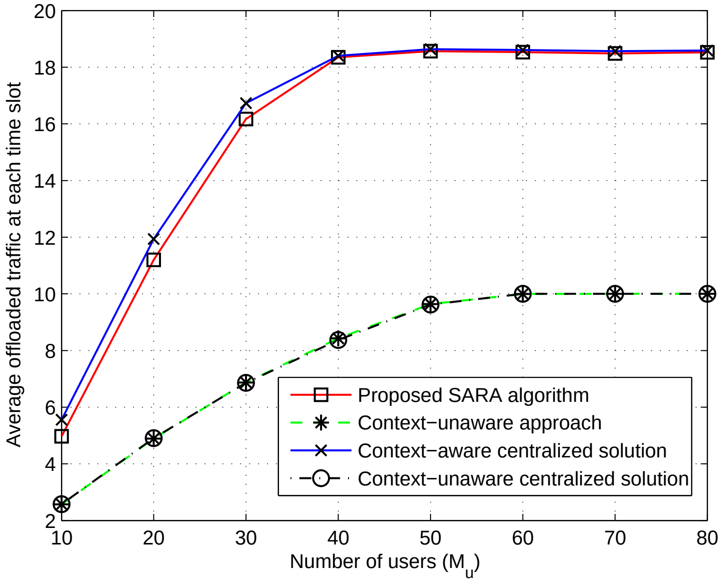


Figure 8.6: Comparison of the average offloaded traffic between the proposed SARA algorithm and other three approaches. The $\rho = 0.5$ and $\rho = 0$ is assumed, respectively, for context-aware algorithms and context-unaware approaches.

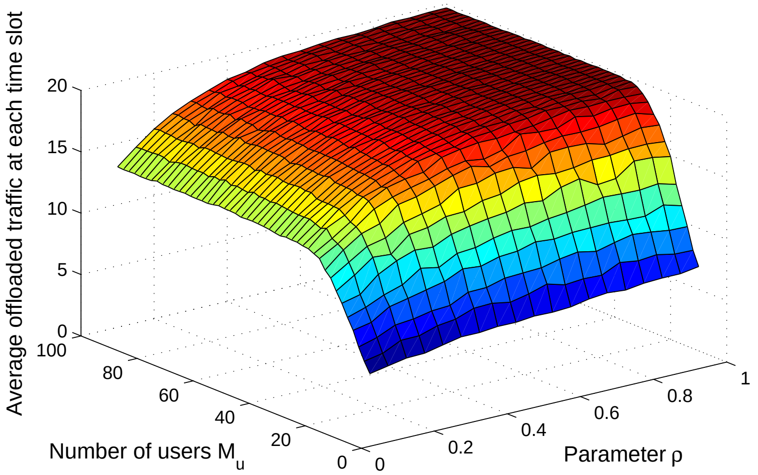


Figure 8.7: The average offloaded traffic resulted by proposed SARA algorithm, versus different M_u and different values of ρ.

and other approaches as the number of RBs $n \in \mathcal{N}_3$ varies for a network with $M_u = 30$ UEs. This figure shows that, as the number of SUE resources increases, the context-aware approach achieves better offload performance. That is due to the fact that more users with shared interests can join the same cluster which results to form stronger social clusters. In Fig. 8.8, we can see that the gap between the offloaded traffic of the proposed SARA algorithm and the context-unaware approach consistently increases as the network sizes grows. Fig. 8.8 shows that the proposed

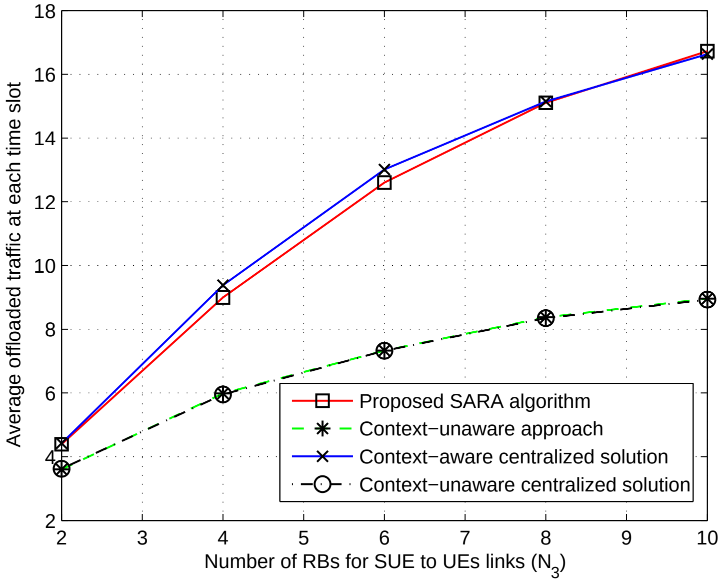


Figure 8.8: Comparison of the average offloaded traffic for the proposed SARA algorithm with other three approaches, versus different number of RBs $n \in \mathcal{N}_3$. The values $M_u = 30$ and $N_1 = 5$ is assumed and the parameter ρ is set to 0.1.

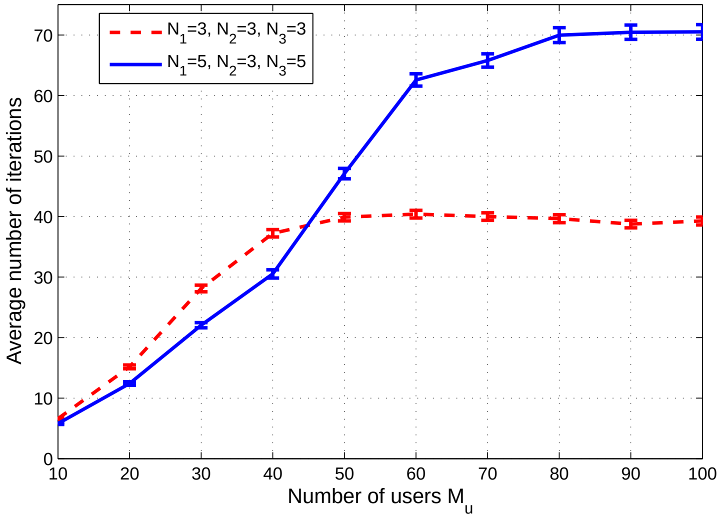


Figure 8.9: Average number iterations of the proposed SARA algorithm vs network size, for different number of RBs. The error bars indicate the 95% confidence.

SARA algorithm can offload up to 78% more traffic from the SCN's infrastructure when compared to the context-unaware approach for a network with $N_3 = 8$ RBs.

Fig. 8.9 shows the average number of iterations resulting from the proposed SARA algorithm versus the network size M_u, for two different RB numbers. In this figure, we can see that, as the number of UEs increases, the average number of iterations increases due to the increase in the number of players. Fig. 8.9 demonstrates that the proposed matching approach has a reasonable convergence time that does not exceed an average of 72 iterations for all network sizes and with $N_1 = 5$, $N_2 = 3$, and $N_3 = 5$ number of RBs. The maximum of the average number of iterations reduces to 40, when there are $N_1 = 3$, $N_2 = 3$, and $N_3 = 3$ RBs.

8.7 Summary

In this chapter, we have presented a novel approach for context-aware resource allocation in D2D-enabled small cell networks. We have formulated the context-aware resource allocation problem as a one-to-one matching game and we have shown that the game exhibits peer effects. To solve the game, we have proposed a distributed social-aware resource allocation algorithm that exploits the physical layer metrics of the wireless network along with the users' social ties from the underlaid social network. Then, we have shown that the proposed algorithm is guaranteed to converge to a two-sided stable matching between the users and the network's resource blocks. Simulation results have shown that the proposed matching-based algorithm yields socially well-connected cluster between D2D links, thus, allowing to offload significantly more traffic than conventional context-unaware approach. The results provide novel insights into the gains that future wireless networks can achieve from exploiting social context. The results show that with manageable complexity, the proposed context-aware approach can substantially improve the wireless resource utilization by offloading a large amount of traffic from the backhaul-constrained small cell network.

8.8 Appendix E

E.1 Proof of Proposition 5

First, consider the game at its initial state where no UE has been assigned to any SUE (no SC is formed). That is, $\sum_{j\in\mathcal{M}_u\setminus m} a_{jm_s;\mu\mu'} z_{mj} = 0$ and $V_{mn,m_s} = w_n \log\left(1+\gamma^{(3)}_{m_snm}\right) + \alpha_m z_{mm_s}$ for $\forall m \in \mathcal{M}_u$. Thus, from (8.17), a UE m that applies for RB $n \in \mathcal{N}_3$ during the first matching will do so only due to the higher achievable rates plus social tie with the corresponding SUE. Without loss of generality, the preference ordering of UE m is identified with two sets $\mathcal{U}$ and $\mathcal{W}$, where

$$\begin{aligned}
\mathcal{U} &= \{\forall n' \in \mathcal{N}_1 | V_{mn',l} > V_{mn,m_s}\} \\
&\quad \cup \{\forall n' \in \mathcal{N}_3 | V_{mn',m_s} > V_{mn,m_s}\}; \\
\mathcal{W} &= \{\forall n' \in \mathcal{N}_1 | V_{mn',l} < V_{mn,m_s}\} \\
&\quad \cup \{\forall n' \in \mathcal{N}_3 | V_{mn',m_s} < V_{mn,m_s}\}.
\end{aligned} \tag{8.25}$$

In (8.25), $\mathcal{U}$ and $\mathcal{W}$, represent the set of all RBs that are, respectively, more preferred and less preferred to UE m than RB n. Thus, the preference ordering of UE m can be denoted by $\mathcal{U} \succ_m$

$n \succ_m \mathcal{W}$. In Stage 2, the process of acceptance or rejection of applicants is done in a manner analogous to the conventional deferred acceptance algorithm [166]. Thus, we can ensure that, for a given SC setting, each $n \in \mathcal{N}_3$ has accepted, out of the applicants, the UE that

$$\begin{aligned}\mu(n) &= \underset{m}{\operatorname{argmax}}\, U_{nm,m_s}(\gamma^{(3)}_{m_s nm}) \\ &= \underset{m}{\operatorname{argmax}}\, w_n \log\left(1 + \gamma^{(3)}_{m_s nm}\right) + \kappa_n \cdot z_{mm_s}. \end{aligned} \tag{8.26}$$

Therefore, if another UE m' applies for the RB n during the next matching μ', then necessarily $n' = \mu(m') \in \mathcal{W}$, if UE m' is not unmatched. This is due to the fact that, if $n' = \mu(m') \in \mathcal{U}$, then it means that UE m' already satisfies (8.26) for RB n', and hence, it will be accepted by n', before applying to n.

Now, since $\mu(m') \in \mathcal{W}$, we can conclude that $U_{nm',m_s} < U_{nm,m_s}$; since otherwise, $\mu(m') = n$ which contradicts (8.26). In other words, the UEs who are accepted in the first iteration will not be rejected by their match as the game proceeds. We can hold the same argument for the next iterations. Nevertheless, this does not imply that UE m necessarily stays in its cluster forever. To show this, we give an example.

Example: With this example, we show that UEs may alter the S-stability of the clusters. Given two UEs m and m', two RBs $n \in \mathcal{N}_1$ and $n' \in \mathcal{N}_3$, and two SUEs m_{s1} and m_{s2}, assume that $V_{mn;m_{s1}} > V_{mn';m_{s2}}$ before any SC is formed and let the current matching be $\mathcal{C}_{m_{s1}} = \{m_{s1}, m\}$ with $\mu(m) = n$ and $\mathcal{C}_{m_{s2}} = \{m_{s2}, m'\}$ with $\mu(m') = n'$. Once the SC information is updated for UEs, the utility of UE m for RBs n and n' might change to $V_{mn;m_{s1}} < V_{mn';m_{s2}}$, due to its social tie with m'. Therefore, m will leave its current cluster $\mathcal{C}_{m_{s1}}$ and will join $\mathcal{C}_{m_{s2}}$. Due to this move, both clusters are not considered S-stable.

E.2 Proof of Theorem 8

From Proposition 5, each $n \in \mathcal{N}_3$ will not reject its current match $\mu(n)$, in favor of other applicants. Thus, a new UE m can join an SC, only if it applies for an unmatched $n \in \mathcal{N}_3$. After each matching is done, more UEs will join SCs, due to the non-negative social effect they impose on one another. Eventually, there is a stage where no more UEs prefer to join clusters.

There can be only two possibilities if a UE m with $\mu(m) \in \mathcal{N}_3$ can leave its cluster, and thus, alter the S-stability condition: *Case 1:* there is an unmatched $n \in \mathcal{N}_1$ such that $n \succ_m \mu(m)$, and *Case 2:* there is an unmatched $n \in \mathcal{N}_3$ corresponding to another cluster where, again, we have $n \succ_m \mu(m)$. Next, we show that even when either Case 1 or Case 2 occurs, the current SC will converge to an S-stable SC.

For the first case, we can check that UE m has surely been rejected in previous matchings by an RB $n \in \mathcal{N}_1$ such that $n \succ_m \mu(m)$. Here, since $\mu(n)$ is matched to another player and RB n is unmatched now, m will be accepted by RB n. If $\mu(n)$ never applies again for RB n, then we can ensure that UE m will not be rejected by n, since it satisfies $\mu(n) = \operatorname{argmax}_m U_{nm,l}(\gamma^{(1)}_{lnm})$. Thus, UE m will not get back to its SC again. If $\mu(n)$ applies again to RB n, then RB n will reject m. However, $\mu(n)$ cannot cycle between RB n and another RB n' for unlimited number of iterations. This is due to the fact that, if $\mu(n)$ oscillates between n and another RB $n' \in \mathcal{N}_3$, the reason is due to the peer effect by a current SC member m'' who also oscillates between n' and another RB. However, $\mu(n)$ will set $a_{\mu(n)m'';\mu\mu'} = 0$, if m'' does not stay in SC for two consecutive matchings. Therefore, $\mu(n)$ has to finally decide between n and n' after finite iterations. Moreover, if an RB $n' \in \mathcal{N}_1$ is the reason due to which $\mu(n)$ cycles, then similarly we can show that $\mu(n')$ will have to stop the oscillation after a finite number of iterations. Consequently, $\mu(n)$ will stop oscillation.

Case 2 implies that some of the friends of UE m has encouraged m to join their cluster. Here, if UE m joins the new SC, it will never prefer the previous SC to new one, unless some of its friends, say m', leave the cluster. Then, UE m ignores the peer effect of m' by setting $a_{mm';\mu\mu'} = 0$ and reorganizes its preference ordering. After a finite number of iterations, the friends' list who stay in the cluster will not change and, thus, UE m stays in the new cluster or comes back to the previous cluster and will never cycle between those two. In addition, if the incentive for UE m to join RB n' of the new cluster does not stem from the friendship relationship, then, this implies that $n' \succ_m \mu(m)$ and UE m has been rejected by n' in previous matchings. Thus, similar to the previous case, we can observe that after finite iterations, the $\mu(n')$ stops oscillating, and, hence, UE m will decide whether to stay with $\mu(m)$ or leave it.

E.3 Proof of Theorem 9

In order to prove the two-sided stability, we need to show that there is no blocking pair (m, n) or (m_s, n) that meets either of the following:

$$(m, n) \notin \mu \ \mid \mu(m) \in \mathcal{N}_3 \ , \ \ n \in \mathcal{N}_1 \cup \mathcal{N}_3; \tag{8.27}$$

$$(m_s, n) \notin \mu \mid m_s \in \mathcal{M}_s \, , \, n \in \mathcal{N}_2; \tag{8.28}$$

$$(m, n) \notin \mu \ \mid \mu(m) \in \mathcal{N}_1 \ , \ \ n \in \mathcal{N}_1 \cup \mathcal{N}_3. \tag{8.29}$$

All SCs are S-stable once the algorithm terminates since, otherwise, the SC sets will be different from those in previous matching which contradicts the termination condition.

From Theorem 8, we can see that no UE m having $\mu(m) \in \mathcal{N}_3$ would make a blocking pair with any RB $n_1 \in \mathcal{N}_1$ or $n_3 \in \mathcal{N}_3$ from another cluster, due to the S-stability of all clusters. In addition, the matching of RBs $n \in \mathcal{N}_3$ belonging to SUE m_s and UEs in $\mathcal{C}_{m_s}$ is done through the deferred acceptance algorithm. Therefore, these players will not form a blocking pair. Consequently, there is no blocking pair that satisfies (8.27).

In addition, the preferences of RBs $n \in \mathcal{N}_2$ become strictly fixed, once the S-stability is satisfied at all clusters. Then, followed by the deferred acceptance algorithm, we ensure that the given matching between SUEs and RBs $n \in \mathcal{N}_2$ is stable and there is no blocking pair that satisfies (8.28).

Finally, no RB $n \in \mathcal{N}_3$ makes a blocking pair with any m where $\mu(m) \in \mathcal{N}_1$, due to the S-stability of all clusters. Moreover, the matching of RBs $n \in \mathcal{N}_1$ and UEs m where $\mu(m) \in \mathcal{N}_1$ is followed by the deferred acceptance algorithm. Therefore, these players will not form a blocking pair and there is no blocking pair that satisfies (8.29). Hence, the proposed SARA algorithm is guaranteed to reach a two-sided stable matching for all D2D and cellular links.

Chapter 9

Conclusions

In this dissertation, we have addressed some of the challenging resource management problems for next-generation wireless networks. Among these problems include: 1) Context-aware scheduling of the user applications in presence of both mmW and sub-6 GHz frequency resources, 2) Finding a cost-effective mmW backhaul solution in dense small cell networks, 3) Optimizing the mobility management in future dense HetNets by leveraging the new capabilities of emerging cellular networks, such as mmW communications and caching, 4) Addressing the challenge of load balancing in HetNets with both mmW and sub-6 GHz radio access technologies, 5) Designing a novel MAC protocol to optimize the performance of WLANs by managing the traffic over both mmW and sub-6 GHz unlicensed bands, and 6) Exploiting the social context information to reduce the backhaul traffic in D2D-enabled small cell networks. Following, we present a summary of the research that has been carried out in this dissertation.

9.1 Summary

9.1.1 Context-aware scheduling of joint millimeter wave and microwave resources for dual-Mode base stations

Modern smartphones and tablets allow a user to simultaneously run multiple UAs, such as navigation, chat interfaces, or online games. However, each UA may belong to a variety of QoS classes, depending on the required data rates and tolerable delay. To meet the QoS requirement per UA, novel network protocols and scheduling algorithms are required that enable application decoupling, based on the QoS class, and manage network resources to maximize the number of satisfied UAs. Such strict QoS provisioning can be achieved by exploiting the capability of emerging wireless SCNs to operate at both mmW and sub-6 GHz frequencies. However, the intermittent nature of mmW links and other major differences in characteristics of communications over mmW and μW bands make the joint mmW-μW resource management challenging.

To address this problem, in Chapter 3, we have proposed a novel dual-mode scheduling framework to enable cellular networks to jointly perform UA selection and scheduling over μW and mmW bands. The proposed scheduling framework allows multiple UAs to run simultaneously on each UE and utilizes a set of context information, including the CSI per UE, the delay tolerance and required load per UA, and the uncertainty of mmW channels, to maximize the QoS per UA. The dual-mode scheduling problem has been formulated as a min-UR optimization problem which has been shown to be challenging to solve.

Consequently, a long-term scheduling framework, consisting of two stages, has been proposed. Within this framework, first, the joint UA selection and scheduling over μW band has been formulated as a one-to-many matching game between the μW resources and UAs. To solve this problem, a novel scheduling algorithm has been proposed and shown to yield a two-sided stable resource allocation. Second, over the mmW band, the joint context-aware UA selection and scheduling problem has been formulated as a 0-1 Knapsack problem and a novel algorithm that builds on the Q-learning algorithm was proposed to find a suitable mmW scheduling policy while adaptively learning the UEs' LoS probabilities.

Furthermore, we have shown that the proposed scheduling framework can find an effective scheduling solution, over both μW and mmW, in polynomial time. Simulation results have shown that, compared with conventional scheduling schemes, the proposed approach significantly increases the number of satisfied UAs while improving the statistics of QoS violations and enhancing the overall users' quality-of-experience.

9.1.2 Inter-operator resource management for millimeter wave, multi-hop backhaul networks

In Chapter 4, a novel framework has been proposed for optimizing the operation and performance of a large-scale, multi-hop mmW backhaul within a wireless SCN that encompasses multiple MNOs. The proposed framework has been shown to enable the SBSs to jointly decide on forming the multi-hop, mmW links over backhaul infrastructure that belongs to multiple, independent MNOs, while properly allocating resources across those links.

In this regard, the problem has been addressed using a novel framework based on matching theory that was decomposed to two, highly inter-related stages: a multi-hop network formation stage and a resource management stage. One unique feature of this framework was that it jointly accounts for both wireless channel characteristics and economic factors during both network formation and resource management. The multi-hop network formation stage has been formulated as a one-to-many matching game which is solved using a novel algorithm, that builds on the so-called deferred acceptance algorithm and was shown to yield a stable and Pareto optimal multi-hop mmW backhaul network. Then, a one-to-many matching game has been formulated to enable proper resource allocation across the formed multi-hop network. This game was then shown to exhibit peer effects and, as such, a novel algorithm has been developed to find a stable and optimal resource management solution that can properly cope with these peer effects.

Simulation results have shown that, with manageable complexity, the proposed framework yields substantial gains, in terms of the average sum rate, reaching up to 27% and 54%, respectively, compared to a non-cooperative scheme in which inter-operator sharing is not allowed and a random allocation approach. The results also have shown that our framework improves the statistics of the backhaul sum rate and provides insights on how to manage pricing and the cost of the cooperative mmW backhaul network for the MNOs.

9.1.3 Enhanced mobility management in 5G networks

In Chapter 5, a novel approach for analyzing and managing mobility in joint μW-mmW networks has been proposed. The proposed approach leverages device-level caching along with the capabilities of dual-mode SBSs to minimize handover failures, reduce inter-frequency measurement energy consumption, and provide seamless mobility in emerging dense heterogeneous networks.

First, fundamental results on the caching capabilities, including caching probability and cache duration are derived for the proposed dual-mode network scenario. Second, the average achievable rate of caching have been derived for mobile users. Moreover, the impact of caching on the number of HOs, energy consumption, and the average HOF has been analyzed. Then, the proposed cache-enabled mobility management problem was formulated as a *dynamic matching game* between MUEs and SBSs. The goal of this game was to find a distributed handover mechanism that, under network constraints on HOFs and limited cache sizes, allows each MUE to choose between: a) executing an HO to a target SBS, b) being connected to the MBS, or c) perform a transparent HO by using the cached content. The formulated matching game has inherently captured the dynamics of the mobility management problem caused by HOFs. To solve this dynamic matching problem, a novel algorithm has been proposed and its convergence to a two-sided dynamically stable HO policy for MUEs and target SBSs has been proved.

Numerical results have corroborated the analytical derivations and shown that the proposed solution will provides significant reductions in both the HOF and energy consumption of MUEs, resulting in an enhanced mobility management for heterogeneous wireless networks with mmW capabilities.

9.1.4 Downlink cell association and load balancing for joint millimeter wave-microwave cellular networks

In Chapter 6, a novel cell association framework has been proposed that considers both the blockage probability and the achievable rate to assign UEs to mmW-BSs or μW-BSs. The problem was formulated as a one-to-many matching problem with *minimum quota constraints* for the BSs that provides an efficient way to balance the load over the mmW and μW frequency bands.

To solve the problem, a distributed algorithm has been proposed that is guaranteed to yield a Pareto optimal and two-sided stable solution. Simulation results have shown that the proposed matching with MMQ algorithm outperforms the conventional max-RSSI and max-SINR cell association schemes. In addition, it has been shown that the proposed MMQ algorithm can effectively balance the number of UEs associated with the μW-BSs and mmW-BSs and achieve further gains, in terms of the average sum rate.

9.1.5 Performance analysis of integrated sub-6 GHz-millimeter wave wireless local area networks

Millimeter wave communications at the 60 GHz unlicensed band is expected to boost the capacity of WLANs. If properly integrated into legacy IEEE 802.11 standards, mmW communications can offer substantial gains by offloading traffic from congested sub-6 GHz unlicensed bands to the 60 GHz mmW frequency band. In Chapter 7, a novel MAC protocol has been proposed to dynamically manage the WLAN traffic over the unlicensed mmW and sub-6 GHz bands. The proposed protocol leverages the capability of advanced multi-band STAs to perform FST to the mmW band, while considering the intermittent channel at the 60 GHz band and the level of congestion observed over the sub-6 GHz bands.

The performance of the proposed scheme has been analytically studied via a new Markov chain model and the probability of transmissions over the mmW and sub-6 GHz bands, as well as the aggregated saturation throughput have been derived. In addition, analytical results were validated by simulation results. Simulation results have shown that the proposed integrated mmW-sub 6 GHz MAC protocol yields significant performance gains, in terms of maximizing the saturation throughput and minimizing the delay experienced by the STAs. The results also have provided insights on the tradeoffs between the achievable gains and the overhead introduced by the FST procedure.

9.1.6 Leveraging social context information to optimize resource allocation in D2D-enabled small cell networks

Chapter 8 has presented a novel approach for optimizing and managing resource allocation in wireless SCNs with D2D communication. In particular, we have developed a novel resource management framework to address the following challenges: 1) How to extract the social context information, 2) How to define social tie between a pair of users, based on the context information, 3) How to form D2D connections, based on the social ties, and 4) How to allocate the resources to both BS-connected and D2D users. The proposed approach allows to jointly exploit both the wireless and social *context* of wireless users for optimizing the overall allocation of resources and improving traffic offload in SCNs.

This context-aware resource allocation problem has been formulated as a matching game in which UEs and RBs rank one another, based on utility functions that capture both wireless and social metrics. Due to social inter-relations, this game has been shown to belong to a class of matching games with peer effects. To solve this game, a novel, self-organizing algorithm has been proposed, using which UEs and RBs can interact to decide on their desired allocation. The proposed algorithm has then been proven to converge to a two-sided stable matching between UEs and RBs. The properties of the resulting stable outcome have been studied and assessed.

Simulation results using real social data have shown that clustering of socially connected users allows to offload a substantially larger amount of traffic than the conventional context-unaware approach. These results have shown that exploiting social context has high practical relevance in saving resources on the wireless links and on the backhaul.

In light of aforementioned contributions, next, we introduce some of the main directions to extend the research carried out in this dissertation.

9.2 Open Problems

Future works can revolve around addressing the following important open problems:

- The proposed scheduling framework in Chapter 3 focuses on a single-cell scenario, and thus, it does not consider inter-cell interference management required in a multi-cell environment. It is interesting to study how dual-mode capability of BSs can be used to mitigate the inter-cell interference. In addition, we have considered a TDMA scheme over the mmW band with beamforming for a single user at a time. However, other schemes such as user clustering and multi-user MIMO can be employed.
- Addressing different challenges of mmW network at PHY and Link layers, including: 1) Adopting fast beam alignment techniques by proposing novel beam search algorithms based on location information, 2) Reducing the coverage gap between data and control planes, and 3) Leveraging spatial diversity and adopting practical multi-user MIMO transmissions for capacity improvements.
- Chapter 4 has considered a static pricing scheme in the resource allocation problem. This topic can be studied further by finding a comprehensive economical model with dynamic pricing policies, exploring other possibilities for cooperation among MNOs to reduce backhaul costs, and extending the theoretical results on the performance-cost compromises of backhaul solutions.
- The work presented in Chapter 5 can be studied in more details by considering the other challenges of mobility management in small cells, such as the effect of channel fading, Doppler effect (especially at mmW frequencies), and the impact of incomplete information for the user trajectory.
- The performance of the proposed MAC protocol in Chapter 7 can be analyzed further under

unsaturated and heterogeneous traffic scenarios. Moreover, the coexistence of the proposed integrated mmW-μW WLAN with LTE-Unlicensed can be studied.

- In Chapter 8, the proposed resource management scheme assumes a static and orthogonal partitioning of the available bandwidth to serve D2D and cellular links. This work can be extended to dynamic resource allocation in which the spectrum partitioning may vary, depending on the social tie strength among users.
- Extracting context information from dynamic settings of wireless networks is a challenging task, due to the variations of wireless channel, mobility of users, and stochastic network load. In spite of development of many applications for context information extraction, such as Netvizz to extract social context data from Facebook, most of these applications are suitable for offline tasks and cannot be used in online network optimization. In this regard, one can develop applications, specifically designed for fast data extraction, while taking into account the constraints of mobile users.

Bibliography

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020," *Whitepaper*, February 2016.

[2] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3gpp heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, June 2011.

[3] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. Reed, "Femtocells: Past, present, and future," *IEEE J. Select. Areas Commun.*, vol. 30, no. 3, pp. 497–508, April 2012.

[4] A. Ghosh, J. G. Andrews, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. A. Thomas, P. Xia, H. S. Jo, H. S. Dhillon, and T. D. Novlan, "Heterogeneous cellular networks: From theory to practice," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 54–64, June 2012.

[5] T. Q. S. Quek, G. de la Roche, I. Guvenc, and M. Kountouris, *Small Cell Networks: Deployment, PHY Techniques, and Resource Management*, Cambridge University Press, 2013.

[6] I. Hwang, B. Song, and S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Communications Magazine*, vol. 51, no.6, pp. 20–27, Jun. 2013.

[7] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, March 2014.

[8] F. Boccardi, R. W. Heath Jr, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[9] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4391–4403, October 2013.

[10] A. Ghosh, R. Ratasuk, P. Moorut, T. S. Rappaport, and S. Sun, "Millimeter-Wave enhanced local area systems: A high-data-rate approach for future wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1152 –1163, June 2014.

[11] Interdigital, “White paper: Small cell millimeter wave mesh backhaul,” *Interdigital, Inc*, February 2013.

[12] J. Hoadley and P. Maveddat, “Enabling small cell deployment with hetnet,” *IEEE Wireless Communications*, vol. 19, no. 2, pp. 4–5, April 2012.

[13] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schober, and L. Hanzo, “User association in 5G networks: A survey and an outlook,” *IEEE Communications Surveys and Tutorials*, vol. 18, no. 2, pp. 1018–1044, 2016.

[14] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, “Tractable model for rate in self-backhauled millimeter wave cellular networks,” *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2196–2211, Oct 2015.

[15] O. Semiari, W. Saad, Z. Daw, and M. Bennis, “Matching theory for backhaul management in small cell networks with mmwave capabilities,” in *2015 IEEE International Conference on Communications (ICC)*, London, UK, June 2015.

[16] MiWaveS consortium, “Heterogeneous wireless network with mmwave small cell access and backhauling,” *White Paper, available at http://www.miwaves.eu/publications.html*, Jan. 2015.

[17] I. F. Akyildiz, Jiang Xie, and S. Mohanty, “A survey of mobility management in next-generation all-IP-based wireless systems,” *IEEE Wireless Communications*, vol. 11, no. 4, pp. 16–28, August 2004.

[18] F. Giust, L. Cominardi, and C. J. Bernardos, “Distributed mobility management for future 5G networks: overview and analysis of existing approaches,” *IEEE Communications Magazine*, vol. 53, no. 1, pp. 142–149, January 2015.

[19] A. Prasad, O. Tirkkonen, P. LundÃľn, O. N. C. Yilmaz, L. Dalsgaard, and C. Wijting, “Energy-efficient inter-frequency small cell discovery techniques for LTE-advanced heterogeneous network deployments,” *IEEE Communications Magazine*, vol. 51, no. 5, pp. 72–81, May 2013.

[20] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, “Mobility management for femtocells in LTE-advanced: Key aspects and survey of handover decision algorithms,” *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 64–91, First 2014.

[21] A. Ahmed, L. M. Boulahia, and D. Gaiti, “Enabling vertical handover decisions in heterogeneous wireless networks: A state-of-the-art and a classification,” *IEEE Communications Surveys Tutorials*, vol. 16, no. 2, pp. 776–811, Second 2014.

[22] K. Vasudeva, M. Simsek, D. Lopez-Perez, and I. Guvenc, “Impact of channel fading on mobility management in heterogeneous networks,” in *2015 IEEE International Conference on Communication Workshop*, June 2015.

[23] M. Khan and K. Han, "An optimized network selection and handover triggering scheme for heterogeneous self-organized wireless networks," *Mathematical Problems in Engineering*, vol. 16, pp. 1–11, 2014.

[24] H. Zhang, N. Meng, Y. Liu, and X. Zhang, "Performance evaluation for local anchor-based dual connectivity in 5G user-centric network," *IEEE Access*, vol. 4, pp. 5721–5729, September 2016.

[25] I. Elgendi, K. S. Munasinghe, and A. Jamalipour, "Mobility management in three-tier sdn architecture for densenets," in *2016 IEEE Wireless Communications and Networking Conference*, April 2016, pp. 1–6.

[26] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 101–107, June 2011.

[27] F. Boccardi, R.W. Heath, A. Lozano, T.L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.

[28] T. S. Rappaport, W. Roh, and K. Cheun, "Smart antennas could open up new spectrum for 5G," *IEEE Spectrum, Available: http://spectrum.ieee.org/telecom/wireless/smart-antennas-could-open-up-new-spectrum-for-5g*, Aug. 2014.

[29] E. J. Violette, R. H. Espeland, R. O. DeBolt, and F. K. Schwering, "Millimeter-wave propagation at street level in an urban environment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 26, no. 3, pp. 368–380, May 1988.

[30] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3437–3458, October 2015.

[31] D. W. K. Ng, M. Breiling, C. Rohde, F. Burkhardt, and R. Schober, "Energy-efficient 5G outdoor-to-indoor communication: SUDAS over licensed and unlicensed spectrum," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3170–3186, May 2016.

[32] J. Qiao, X. Shen, J. Mark, Q. Shen, Y. He, and L. Lei, "Enabling device-to-device communications in millimeter-wave 5G cellular networks," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 209–215, January 2015.

[33] H. Singh, J. Oh, C. Kweon, X. Qin, H. Shao, and C. Ngo, "A 60 GHz wireless network for enabling uncompressed video communication," *IEEE Communications Magazine*, vol. 46, no. 12, pp. 71–78, December 2008.

[34] D. Wu, J. Wang, Y. Cai, and M. Guizani, "Millimeter-wave multimedia communications: challenges, methodology, and applications," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 232–238, January 2015.

[35] J. Park, S. L. Kim, and J. Zander, “Tractable resource management with uplink decoupled millimeter-wave overlay in ultra-dense cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 4362–4379, June 2016.

[36] M. Wang, A. Dutta, S. Buccapatnam, and M. Chiang, “Smart exploration in hetnets: Minimizing total regret with mmwave,” in *Proc. IEEE International Conference on Sensing, Communication and Networking*, London, UK, June 2016.

[37] F. Boccardi, R. W. Heath Jr., A. Lozano, T. L. Marzetta, and P. Popovski, “Five disruptive technology directions for 5G,” *IEEE Communications Magazine*, vol. 52, no.2, pp. 74–80, Feb. 2014.

[38] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklos, and Z. Turanyi, “Design aspects of network assisted device-to-device communications,” *IEEE Communications Magazine*, vol. 50, no.3, pp. 170–177, Mar. 2012.

[39] B. Kaufman and B. Aazhang, “Cellular networks with an overlaid device to device network,” in *Proc. of 42nd Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Oct. 2008.

[40] Y. Li, T. Wu, P. Hui, D. Jin, and S. Chen, “Social-Aware D2D communications: qualitative insights and quantitative analysis,” *IEEE Communications Magazine*, vol. 52. No. 6, pp. 150–158, June 2014.

[41] H. Ishii, X. Cheng, S. Mukherjee, and B. Yu, “A LTE offload solution using small cells with D2D links,” in *Proc. of IEEE International Conference on Communications*, Budapest, Hungary, Jun 2013.

[42] S. Krishnan and H. S. Dhillon, “Effect of user mobility on the performance of device-to-device networks with distributed caching,” *IEEE Wireless Communications Letters*, vol. 6, no. 2, pp. 194–197, April 2017.

[43] M. Afshang, H. S. Dhillon, and P. H. Joo Chong, “Modeling and performance analysis of clustered device-to-device networks,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4957–4972, July 2016.

[44] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, “Cache in the air: exploiting content caching and delivery techniques for 5G systems,” *IEEE Communications Magazine*, vol. 52, no. 2, pp. 131–139, February 2014.

[45] Y. Rao, H. Zhou, D. Gao, H. Luo, and Y. Liu, “Proactive caching for enhancing user-side mobility support in named data networking,” in *2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, July 2013, pp. 37–42.

[46] K. Poularakis and L. Tassiulas, “Exploiting user mobility for wireless content delivery,” in *2013 IEEE International Symposium on Information Theory*, July 2013, pp. 1017–1021.

[47] Andre S. Gomes, Bruno Sousa, David Palma, Vitor Fonseca, Zhongliang Zhao, Edmundo Monteiro, Torsten Braun, Paulo Simoes, and Luis Cordeiro, "Edge caching with mobility prediction in virtualized LTE mobile networks," *Future Generation Computer Systems*, 2016.

[48] H. Abou-zeid, S. Valentin, and H. S. Hassanein, "Long-term proportional fairness over multiple cells," in *Proc. of IEEE 37th Conference on Local Computer Networks Workshops (LCN Workshop)*, Clearwater, FL, Oct. 2012.

[49] M. Proebster, M. Kaschub, T. Werthmann, and S. Valentin, "Context-aware resource allocation for cellular wireless networks," *EURASIP Journal on Wireless Communications and Networking*, vol. no. 216, July 2012.

[50] O. Riva, T. Nadeem, C. Borcea, and L. Iftode, "Context-aware migratory services in Ad Hoc networks," *IEEE Transactions on Mobile Computing*, vol. 6, no.12, pp. 1313–1328, Dec. 2007.

[51] C. Anagnostopoulos, S. Hadjiefthymiades, and E. Zervas, "Information dissemination between mobile nodes for collaborative context awareness," *IEEE Transactions on Mobile Computing*, vol. 10, no.12, pp. 1710–1725, Dec. 2011.

[52] V. Kulkarni and M. Devetsikiotis, "Social distance aware resource allocation in wireless networks," in *Proc. of IEEE Global Telecommunications Conference*, Honolulu, HI, Dec. 2009.

[53] G. I. Tsiropoulos, D. G. Stratogiannis, N. Mantas, and M. Louta, "The impact of social distance on utility based resource allocation in next generation networks," in *Proc. of 3rd International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Budapest, Hungary, Oct. 2011.

[54] D. G. Stratogiannis, G. I. Tsiropoulos, and P. G. Cottis, "Bandwidth allocation in wireless networks employing social distance aware utility functions," in *Proc. of IEEE Global Telecommunications Workshops*, Houston, TX, Dec. 2011.

[55] Prodromos Makris, Dimitrios N Skoutas, and Charalabos Skianis, "A survey on context-aware mobile and wireless networking: On networking and computing environments' integration," *Communications Surveys & Tutorials, IEEE*, vol. 15, no. 1, pp. 362–386, 2013.

[56] W.N. Schilit, *A system architecture for context-aware mobile computing*, Ph.D. thesis, Columbia University, New York, 1995.

[57] X. Chen, F. Meriaux, and S. Valentin, "Predicting a user's next cell with supervised learning based on channel states," in *Proc. IEEE Int. Workshop on Signal Processing Advances for Wireless Commun. (SPAWC)*, June 2013.

[58] H. Abou-zeid, H. Hassanein, and S. Valentin, "Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2013–2026, June 2014.

[59] Nicola Bui, *Mobile network resource optimization under imperfect prediction*, Ph.D. thesis, Universidad Carlos III de Madrid, Spain, 2014.

[60] H. Abou-zeid and H. Hassanein, "Towards green media delivery: location-aware opportunities and approaches," *IEEE Wireless Communications*, vol. 21, no. 4, pp. 38–46, August 2014.

[61] O. Semiari, W. Saad, S. Valentin, M. Bennis, and H. V. Poor, "Context-aware small cell networks: How social metrics improve wireless resource allocation," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 5927–5940, Nov 2015.

[62] Dusit Niyato and Ekram Hossain, "Competitive pricing in heterogeneous wireless access networks: Issues and approaches," *Network, IEEE*, vol. 22, no. 6, pp. 4–11, 2008.

[63] Dusit Niyato and Ekram Hossain, "A game theoretic analysis of service competition and pricing in heterogeneous wireless access networks," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 12, pp. 5150–5155, 2008.

[64] N. A. Bradley and M. D. Dunlop, "Towards a user-centric and multidisciplinary framework for designing context-aware applications," *First International Workshop on Advanced Context Modeling, Reasoning And Management*, 2004.

[65] A. Kumar V. Goyal and V. Sharma, "Power constrained and delay optimal policies for scheduling transmission over a fading channel," in *Proc. of IEEE Infocom*, San Francisco, USA, April 2003.

[66] R. A. Berry and E. M. Yeh, "Cross-layer wireless resource allocation," *IEEE signal processing magazine*, vol. 21, no. 5, pp. 59 – 68, August 2004.

[67] M. Bennis, M. Simsek, W. Saad, S. Valentin, M. Debbah, and A. Czylwik, "When cellular meets wifi in wireless small cell networks," *IEEE communications magazine*, vol. 51, no. 6, pp. 44–50, June 2013.

[68] Francesco Pantisano, Mehdi Bennis, Walid Saad, Stefan Valentin, and Mérouane Debbah, "Matching with externalities for context-aware user-cell association in small cell networks," in *Globecom Workshops (GC Wkshps), 2013 IEEE*. IEEE, 2013, pp. 4483–4488.

[69] Ahmad M El-Hajj and Zaher Dawy, "A context-aware resource management approach for heterogeneous connectivity in next generation wireless networks," .

[70] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Match to cache: Joint user association and backhaul allocation in cache-aware small cell networks," in *Proc. of the IEEE International Conference on Communications (ICC), Mobile and Wireless Networks Symposium*, London, UK, June 2015.

[71] Mikhail Gerasimenko, Nageen Himayat, Shu-ping Yeh, Shilpa Talwar, Sergey Andreev, and Yevgeni Koucheryavy, "Characterizing performance of load-aware network selection in multi-radio (wifi/lte) heterogeneous networks," in *Globecom Workshops (GC Wkshps), 2013 IEEE*. IEEE, 2013, pp. 397–402.

[72] D. Gale and L Shapley, "College admissions and the stability of marriage," *American Mathematical Monthly*, vol. 69, pp. 9–15, Jan. 1962.

[73] Meriem Kassar, Brigitte Kervella, and Guy Pujolle, "An overview of vertical handover decision strategies in heterogeneous wireless networks," *Computer Communications*, vol. 31, no. 10, pp. 2607 – 2620, 2008.

[74] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: fundamentals and applications," *IEEE Communications Magazine*, vol. 53, no. 5, pp. 52–59, May 2015.

[75] O. Semiari, W. Saad, and M. Bennis, "Context-aware scheduling of joint millimeter wave and microwave resources for dual-mode base stations," in *Proc. of IEEE International Conference on Communications, Mobile and Wireless Networks Symposium*, Kualalumpur, Malaysia, May 2016.

[76] E. Jorswieck, "Stable matchings for resource allocation in wireless networks," in *Proc. of 17th International Conference on Digital Signal Processing (DSP)*, Corfu, Greece, July 2011.

[77] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: Fundamentals and applications," *arXiv:1410.6513*, March 2015.

[78] J. Wells, *Multi-Gigabit Microwave and Millimeter-Wave Wireless Communications*, Artech House, 2010.

[79] M. Mezzavilla, A. Dhananjay, S. Panwar, S. Rangan, and M. Zorzi, "An MDP model for optimal handover decisions in mmwave cellular networks," *arXiv:1507.00387*, Feb. 2016.

[80] S. Singh, S. Yeh, N. Himayat, and S. Talwar, "Optimal traffic aggregation in multi-RAT heterogeneous wireless networks," *arXiv:1603.08062*, Mar. 2016.

[81] Z. Wei, X. Zhu, S. Sun, and Y. Huang, "Energy-efficiency-oriented cross-layer resource allocation for multiuser full-duplex decode-and-forward indoor relay systems at 60 GHz," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3366–3379, Dec 2016.

[82] Z. Pi, F. Khan, and J. Zhang, "Techniques for millimeter wave mobile communication," July 2011, US Patent App. 12/916,019.

[83] A. Ghosh and R. Ratasuk, "Multi-antenna systems for LTE eNodeB," in *Proc. of IEEE 70th Vehicular Technology Conference*, Anchorage, Alaska, Sept 2009.

[84] H. Moon, "Waterfilling power allocation at high SNR regimes," *IEEE Transactions on Communications*, vol. 59, no. 3, pp. 708–715, March 2011.

[85] T. Baykas, C. S. Sum, Z. Lan, J. Wang, M. A. Rahman, H. Harada, and S. Kato, "IEEE 802.15.3c: the first ieee wireless standard for data rates over 1 Gb/s," *IEEE Communications Magazine*, vol. 49, no. 7, pp. 114–121, July 2011.

[86] T. Nitsche, A. B. Flores, E. W. Knightly, and J. Widmer, "Steering with eyes closed: Mm-wave beam steering without in-band measurement," in *Proc. of 2015 IEEE Conference on Computer Communications*, Hong Kong, April 2015.

[87] M. K. Samimi and T. S. Rappaport, "Local multipath model parameters for generating 5G millimeter-wave 3GPP-like channel impulse response," in *Proc. of 10th European Conference on Antennas and Propagation (EuCAP)*, Davos, Switzerland, April 2016.

[88] E. Amaldi and V. Kann, "On the approximation of minimizing non zero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, pp. 237–260, December 1998.

[89] K. Seong, M. Mohseni, and J. M. Cioffi, "Optimal resource allocation for OFDMA downlink systems," in *Proc. of IEEE International Symposium on Information Theory*, Seattle, Washington, July 2006.

[90] A. E. Roth and M. A. O. Sotomayor, *Two-sided matching: A study in game-theoretic modeling and analysis*, Cambridge University Press, 1992.

[91] O. Semiari, W. Saad, S. Valentin, M. Bennis, and B. Maham, "Matching theory for priority-based cell association in the downlink of wireless small cell networks," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[92] O. Semiari, W. Saad, S. Valentin, M. Bennis, and H. V. Poor, "Context-aware small cell networks: How social metrics improve wireless resource allocation," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 5927–5940, Nov 2015.

[93] B. C. Dean, *Approximation algorithms for stochastic scheduling problems*, Ph.D. thesis, Massachusetts Institute of Technology, Boston, 2005.

[94] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[95] M. Simsek, M. Bennis, and A. Czylwik, "Dynamic inter-cell interference coordination in HetNets: A reinforcement learning approach," in *Proc. of IEEE Global Communications Conference*, Anaheim, USA, Dec 2012.

[96] T. Bai, R. Vaze, and R. W. Heath, "Analysis of blockage effects on urban cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 9, pp. 5070–5083, Sept 2014.

[97] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, pp. 2318–2328, June 2008.

[98] L. Qian, N. Song, D. R. Vaman, X. Li, and Z. Gajic, "Power control and proportional fair scheduling with minimum rate constraints in clustered multihop TD/CDMA wireless ad hoc networks," in *Proc. of IEEE Wireless Communications and Networking Conference*, Las Vegas, USA, April 2006.

[99] S. Collonge, G. Zaharia, and G. E. Zein, "Influence of the human activity on wide-band characteristics of the 60 GHz indoor radio channel," *IEEE Transactions on Wireless Communications*, vol. 3, no. 6, pp. 2396–2406, Nov 2004.

[100] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2010-2015," *Whitepaper*, February 2011.

[101] T. Q. S. Quek, G. de la Roche, I. Guvenc, and M. Kountouris, *Small Cell Networks: Deployment, PHY Techniques, and Resource Management*, Cambridge University Press, 2013.

[102] G. Liebl, T. Martins de Moraes, A. Soysal, and E. Seidel, "Fair resource allocation for the relay backhaul link in LTE-advanced," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, Shanghai, China, April 2012.

[103] S.Yi and M. Lei, "Backhaul resource allocation in LTE-advanced relaying systems," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, Shanghai, China, April 2012.

[104] I. V. Loumiotis, E. F. Adamopoulou, K. P. Demestichas, T. A. Stamatiadi, and M. E. Theologou, "Dynamic backhaul resource allocation: An evolutionary game theoretic approach," *IEEE Transactions on Communications*, vol. 62, no. 2, pp. 691–698, February 2014.

[105] M. Peter, R. J. Weiler, T. Kuhne, B. Goktepe, J. Serafimoska, and W. Keusgen, "Millimeter-wave small-cell backhaul measurements and considerations on street-level deployment," in *2015 IEEE Globecom Workshops (GC Wkshps)*, San Diego, USA, Dec. 2015.

[106] X. Ge, S. Tu, G. Mao, C. X. Wang, and T. Han, "5g ultra-dense cellular networks," *IEEE Wireless Communications*, vol. 23, no. 1, pp. 72–79, February 2016.

[107] Q. Li, H. Niu, G. Wu, and R. Q. Hu, "Anchor-booster based heterogeneous networks with mmwave capable booster cells," in *Proc. of Globecom 2013 Workshop-Emerging Technologies for LTE-Advanced and Beyond-4G*, Atlanta, GA, 2013.

[108] Honglei Miao and M. Faerber, "Self-organized multi-hop millimeter-wave backhaul network: Beam alignment and dynamic routing," in *Proc. of 2015 European Conference on Networks and Communications (EuCNC)*, Paris, June 2015.

[109] A. Ahmed and D. Grace, "A dual-hop backhaul network architecture for 5g ultra-small cells using millimetre-wave," in *2015 IEEE International Conference on Ubiquitous Wireless Broadband (ICUWB)*, Montreal, Canada, Oct. 2015.

[110] I. Siaud, A. M. Ulmer-Moll, M. A. Bouzigues, and N. Cassiau, "Adaptive and spatial processing for millimeter wave backhaul architectures," in *Proc. of 2015 IEEE International Conference on Ubiquitous Wireless Broadband (ICUWB)*, Montreal, Canada, Oct 2015.

[111] B. Malila, O. Falowo, and N. Ventura, "Millimeter wave small cell backhaul: An analysis of diffraction loss in NLOS links in urban canyons," in *Proc. of IEEE AFRICON*, Addis Ababa, Sept 2015.

[112] X. Xu, W. Saad, X. Zhang, X. Xu, and S. Zhou, "Joint deployment of small cells and wireless backhaul links in next-generation networks," *IEEE Communications Letters*, vol. 19, no. 12, pp. 2250–2253, Dec 2015.

[113] Girija Narlikar, Gordon Wilfong, and Lisa Zhang, "Designing multihop wireless backhaul networks with delay guarantees," *Wireless Networks*, vol. 16, no. 1, pp. 237–254, 2010.

[114] Interdigital, "White paper: Street lightsmall cells âĂŞa revolution in mobile operator network economics," *Interdigital, Inc*, Oct. 2014.

[115] K. Taga, G. Peres, B. Grau, and C. Schwaiger, "Network cooperation, making it work and creating value," *White Paper, available at* `http://www.adlittle.com/viewpoints.html?&view=632`, 2013.

[116] S. Sengupta and M. Chatterjee, "An economic framework for dynamic spectrum access and service pricing," *IEEE/ACM Transactions on Networking*, vol. 17, no. 4, pp. 1200–1213, August 2009.

[117] M. Mahloo, P. Monti, J. Chen, and Lena Wosinska, "Cost modeling of backhaul for mobile networks," in *Proc. of IEEE International Conference on Communications (ICC) workshops*, Sydney, NSW, June 2014.

[118] D. Li, W. Saad, and C. S. Hong, "Decentralized renewable energy pricing and allocation for millimeter wave cellular backhaul," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1140–1159, May 2016.

[119] M. K. Samimi, T. S. Rappaport, and G. R. MacCartney, "Probabilistic omnidirectional path loss models for millimeter-wave outdoor communications," *IEEE Wireless Communications Letters*, vol. 4, no. 4, pp. 357–360, Aug 2015.

[120] S. Biswas, S. Vuppala, J. Xue, and T. Ratnarajah, "An analysis on relay assisted millimeter wave networks," in *Proc. of IEEE International Conference on Communications (ICC)*, Kuala Lumpur,Malaysia, May 2016.

[121] M. K. Samimi and T. S. Rappaport, "3-D millimeter-wave statistical channel model for 5G wireless system design," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 7, pp. 2207–2225, July 2016.

[122] M. K. Samimi and T. S. Rappaport, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 4, pp. 1850–1859, April 2013.

[123] Myung-Don Kim, Jinyi Liang, Juyul Lee, Jaejoon Park, and Bonghyuk Park, "Directional multipath propagation characteristics based on 28GHz outdoor channel measurements," in *Proc. of 10th European Conference on Antennas and Propagation (EuCAP)*, Davos, Switzerland, April 2016.

[124] P. B. Papazian, C. Gentile, K. A. Remley, J. Senic, and N. Golmie, "A radio channel sounder for mobile millimeter-wave communications: System implementation and measurement assessment," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 9, pp. 2924–2932, Sept 2016.

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

[126] M. Kim, Y. Kim, and W. Lee, "Resource allocation scheme for millimeter wave-based WPANs using directional antennas," *ETRI Journal*, vol. 36, no. 3, pp. 385–395, June 2014.

[127] H. Kwon, Y. H. Kim, and B. D. Rao, "Uniform power allocation with thresholding over rayleigh slow fading channels with QAM Inputs," in *Proc. of IEEE 78th Vehicular Technology Conference (VTC Fall)*, Las Vegas, USA, Jan 2013.

[128] M. Sikora, J. N. Laneman, M. Haenggi, D. J. Costello, and T. E. Fuja, "Bandwidth- and power-efficient routing in linear wireless networks," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2624–2633, June 2006.

[129] Q. Ye, M. A. Shalash, C. Caramanis, and J. G. Andrews, "Distributed resource allocation in device-to-device enhanced cellular networks," *IEEE Trans. Comm.*, vol. 63, no. 2, pp. 441–454, December 2015.

[130] D. Gale and L. Shapley, "College admissions and the stability of marriage," *Amer. Math. Monthly*, vol. 69, pp. 9 –15, 1962.

[131] Z. Zhou, M. Dong, K. Ota, and Z. Chang, "Energy-efficient context-aware matching for resource allocation in ultra-dense small cells," *IEEE Access*, vol. 3, pp. 1849–1860, 2015.

[132] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *Proc. of the 4th International Conference on Algorithmic Game Theory*, Amalfi, Italy, July 2011.

[133] D. Lopez-Perez, I. Guvenc, and X. Chu, "Mobility management challenges in 3GPP heterogeneous networks," *IEEE Communications Magazine*, vol. 50, no. 12, pp. 70–78, December 2012.

[134] S. G. Park and Y. S. Choi, "Mobility enhancement in centralized mmwave-based multi-spot beam cellular system," in *2015 International Conference on Information and Communication Technology Convergence*, October 2015, pp. 200–205.

[135] J. Qiao, X. S. Shen, J. W. Mark, and L. Lei, "Video quality provisioning for millimeter wave 5G cellular networks with link outage," *IEEE Transactions on Wireless Communications*, vol. 14, no. 10, pp. 5692–5703, October 2015.

[136] O. Semiari, W. Saad, M. Bennis, and B. Maham, "Mobility management for heterogeneous networks: Caching meets millimeter wave to provide seamless handover," in *submitted to the 2017 IEEE International Symposium on Information Theory*, June 2017.

[137] A. Prasad, O. Tirkkonen, P. Lunden, O. N.C. Yilmaz, L. Dalsgaard, and C. Wijting, "Energy-efficient inter-frequency small cell discovery techniques for LTE-Advanced heterogeneous network deployments," in *IEEE Communications Magazine*, May 2013, pp. 72–81.

[138] A. Ravanshid, P. Rost, D. S. Michalopoulos, V. V. Phan, H. Bakker, D. Aziz, S. Tayade, H. D. Schotten, S. Wong, and O. Holland, "Multi-connectivity functional architectures in 5g," in *2016 IEEE International Conference on Communications Workshops*, May 2016, pp. 187–192.

[139] 3GPP, "E-UTRA: Mobility enhancements in heterogeneous networks," *3rd Generation Partnership Project*, vol. Rel 11, September 2012.

[140] C. H. M. de Lima, M. Bennis, and M. Latva-aho, "Modeling and analysis of handover failure probability in small cell networks," in *Proc. of IEEE Conference on Computer Communications Workshops*, April 2014, pp. 736–741.

[141] S.V. Kadam and M. H. Kotowski, "Multi-period matching," *Available online at `http://scholar.harvard.edu/kadam/publications/multi-period-matching`*, April 2016.

[142] H. Elshaer, M. N. Kulkarni, F. Boccardi, J. G. Andrews, and M. Dohler, "Downlink and uplink cell association with traditional macrocells and millimeter wave small cells," *arXiv:1601.05281*, Jan 2016.

[143] J. Park, S. L. Kim, and J. Zander, "Tractable resource management with uplink decoupled millimeter-wave overlay in ultra-dense cellular networks," *arXiv:1507.08979*, 2016.

[144] I. Guvenc, "Capacity and fairness analysis of heterogeneous networks with range expansion and interference coordination," *IEEE Communications Letters*, vol. 15, no. 10, pp. 1084–1087, October 2011.

[145] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[146] D. Niyato and E. Hossain, "Dynamics of network selection in heterogeneous wireless networks: An evolutionary game approach," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 4, pp. 2008–2017, May 2009.

[147] G. Athanasiou, P. C. Weeraddana, C. Fischione, and L. Tassiulas, "Optimizing client association for load balancing and fairness in millimeter-wave wireless networks," *IEEE/ACM Transactions on Networking*, vol. 23, no. 3, pp. 836–850, June 2015.

[148] D. E. Fragiadakis, A. Iwasaki, P. Troyan, S. Ueda, and M. Yokoo, "Strategyproof matching with minimum quotas," in *Proc. of 11th International Conference on Autonomous Agents and Multiagent Systems*, Valencia, Spain, June 2012.

[149] D. Wang and C. H. Chan, "Multiband antenna for WiFi and WiGig communications," *IEEE Antennas and Wireless Propagation Letters*, vol. 15, pp. 309–312, 2016.

[150] S. Choi, J. d. Prado, S. Shankar N, and S. Mangold, "IEEE 802.11e contention-based channel access (EDCF) performance evaluation," *Proc. of IEEE International Conference on Communications,*, vol. 2, pp. 1151–1156, May 2003.

[151] O. Tickoo and B. Sikdar, "Queueing analysis and delay mitigation in IEEE 802.11 random access MAC based wireless networks," *Proc. of IEEE INFOCOM*, vol. 2, pp. 1404–1413, March 2004.

[152] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, March 2000.

[153] D. Malone, K. Duffy, and D. Leith, "Modeling the 802.11 distributed coordination function in nonsaturated heterogeneous conditions," *IEEE/ACM Transactions on Networking*, vol. 15, no. 1, pp. 159–172, February 2007.

[154] P. Liu, Z. Tao, S. Narayanan, T. Korakis, and S. S. Panwar, "CoopMAC: A cooperative MAC for wireless LANs," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 2, pp. 340–354, February 2007.

[155] I. Inan, F. Keceli, and E. Ayanoglu, "Saturation throughput analysis of the 802.11e enhanced distributed channel access function," *Proc. IEEE International Conference on Communications*, pp. 409–414, June 2007.

[156] Hua Zhu, Ming Li, I. Chlamtac, and B. Prabhakaran, "A survey of quality of service in IEEE 802.11 networks," *IEEE Wireless Communications*, vol. 11, no. 4, pp. 6–14, August 2004.

[157] IEEE 802.11 WG, "IEEE 802.11ad, Amendment 3: Enhancements for very high throughput in the 60 GHz band," December 2012.

[158] K. Chandra, V. Prasad, and I. Niemegeers, "Performance analysis of IEEE 802.11ad MAC protocol," *IEEE Communications Letters*, vol. PP, no. 99, pp. 1–1, 2017.

[159] Q. Chen, J. Tang, D. T. C. Wong, X. Peng, and Y. Zhang, "Directional cooperative MAC protocol design and performance analysis for IEEE 802.11ad WLANs," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 6, pp. 2667–2677, July 2013.

[160] X. Zhu, A. Doufexi, and T. Kocak, "Throughput and coverage performance for IEEE 802.11ad millimeter-wave WPANs," *Proc. of IEEE 73rd Vehicular Technology Conference (VTC Spring)*, pp. 1–5, May 2011.

[161] IEEE 802.11 WG, "Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," March 2012.

[162] S. Lakshminarayana and M. Assaad, "H-infinity control based scheduler for the deployment of small cell networks," in *Proc. of IEEE International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, Princeton, NJ, May 2011.

[163] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE J. Select. Areas Commun.*, vol. 28, no. 9, pp. 1479–1489, Dec. 2010.

[164] V. N. Ha and L. B. Le, "Fair resource allocation for OFDMA femtocell networks with macrocell protection," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 3, pp. 1–14, March 2014.

[165] P. Semasinghe, E. Hossain, and K. Zhu, "An evolutionary game for distributed resource allocation in self-organizing small cells," *IEEE Transactions on Mobile Computing*, vol. 99, pp. 1–15, April 2014.

[166] E. Jorswieck, “Stable matchings for resource allocation in wireless networks,” in *Proc. of 17th International Conference on Digital Signal Processing (DSP)*, Corfu, Greece, July 2011.

[167] B. Holfeld, R. Mochaourab, and T. Wirth, “Stable matching for adaptive cross-layer scheduling in the LTE downlink,” in *Proc. of IEEE Vehicular Technology Conference*, Dresden, Germany, Jun 2013.

[168] O. Semiari, W. Saad, S. Valentin, M. Bennis, and B. Maham, “Matching theory for priority-based cell association in the downlink of wireless small cell networks,” in *Proc. of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 2014.

[169] A. M. El-Hajj, Z. Dawy, and W. Saad, “A stable matching game for joint uplink/downlink resource allocation in ofdma wireless networks,” in *Proc. of IEEE International Conference on Communications*, Ottawa, ON, June 2012.

[170] C. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, “Resource sharing optimization for D2D communication underlaying cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2752–2763, Aug. 2011.

[171] N. Yu and Q. Han, “Context-aware communities and their impact on information influence in mobile social networks,” in *Proc. of IEEE Pervasive Computing and Communications Workshops*, Lugano, Switzerland, Mar. 2012.

[172] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora, “Distance matters: geo-social metrics for online social networks,” in *Proc. of the 3rd workshop on online social networks (WOSN’10)*, Boston, MA, June 2010.

[173] M. McPherson, L. Smith-Lovin, and J. Cook, “Homophily in social networks,” *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[174] M. Granovetter, “The strength of weak ties: a network theory revisited,” *Sociological Theory*, vol. 1, pp. 201–233, 1983.

[175] R. Xiang, J. Neville, and M. Rogati, “Modeling relationship strength in online social networks,” in *Proc. of Workshop on Analyzing Networks and Learning with Graphs*, Dec. 2009.

[176] R. M. Karp, R. E. Miller, and J. W. Thatcher, “Reducibility among combinatorial problems,” *Complexity of Computer Computations*, pp. 85–103, 1972.

[177] A. E. Roth and M. A. O. Sotomayor, *Two-sided matching: A study in game-theoretic modeling and analysis*, Cambridge University Press, 1992.

[178] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, “Peer effects and stability in matching markets,” in *Proc. of the 4th Symposium on Algorithmic Game Theory*, 2011.

[179] J. Leskovec, "Stanford large network dataset collection," http://snap.stanford.edu/data, 2012.

[180] 3GPP, "E-UTRA physical channels and modulation (release 12)," Tech. Rep. TS 36.211 V12.1.0, 3GPP, March 2014.