NORTHWESTERN UNIVERSITY

Centralized Radio Resource Management for Metropolitan Area Networks

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Electrical Engineering and Computer Science

By

Zhiyi Zhou

EVANSTON, ILLINOIS

September 2018

ABSTRACT

Centralized Radio Resource Management for Metropolitan Area Networks

Zhiyi Zhou

In the last decade, global mobile data traffic increased by more than a hundred fold times while maintaining essentially the same monthly charge to the average mobile user. Cisco predicts that overall mobile data traffic will continue to grow rapidly at a compound annual growth rate of 60 percent between 2017 and year 2021. Anticipating future demands, the wireless industry set an ambitious goal to increase the capacity per unit area by three orders of magnitude through the deployment of next generation technologies, often referred to as 5G or IMT-2020.

The capacity gain will be achieved mainly using three means: 1) increased spectral efficiency, primarily through physical-layer improvements and the use of efficient resource management; 2) extreme network densification to improve the area spectral efficiency; and 3) increased bandwidth, primarily by exploiting the millimeter wave band. There has been broad consensus that next-generation networks are going to be heterogeneous with dense deployment of small cells under the umbrella of macro cells.

The objective of this thesis is to formulate the general centralized resource management problems as a class of optimization problems and to provide computationally tractable resource management methods. In this thesis, the general resource management problems are addressed in three aspects: 1) spectrum allocation and user association over multiple radio access technologies. 2) scalable large-scale resource management with guaranteed near-optimal performance. 3) joint spectrum allocation, user association, and power control for large networks.

First, we study spectrum allocation in downlink heterogeneous networks (HetNets) with multiple radio access technologies (RATs) over different bands using the average packet sojourn time as the performance metric. In addition to the licensed band, a queueing model with vacation has been proposed to model the additional delay associated with the unlicensed band. Two optimization-based schemes have been proposed and shown to be highly effective through simulation results.

The thesis then focuses on the design of scalable centralized resource allocation algorithms for large-scale networks consisting many hundred access points (APs) and user devices. Instead of solving a convex optimization problem with an exponential number of variables in the network size, a scalable reformulation is obtained by exploiting the geometric graph nature of the network and provable sparsity of the optimal solution. A pattern pursuit algorithm with low complexity is proposed to solve the reformulated problem with guaranteed gap to the global optimum.

Finally, the joint spectrum allocation, user association, and power control problem for large-scale networks is studied. We develop a scalable reformulation by exploiting the hidden sparse structure of the optimal solution. An efficient algorithm is proposed to

solve the reformulated problem with guaranteed convergence. Moreover, each iteration is performed in closed form, which makes centralize resource allocation practically feasible even for a very large network.

Acknowledgements

Above all, I would like to express my sincere gratitude to my advisors, Professor Dongning Guo and Professor Michael Honig, for their insightful guidance, enlightening advising, and continuous support and encouragement throughout my Ph.D. study. Their enthusiasm for research and dedication to their students are truly inspiring. I feel fortunate and proud that I made the wisest decision when I chose to join this wonderful research group, and had them as my Ph.D. advisors.

I also want to thank Dr. Weimin Xiao, Dr. Jialing Liu, Professor Ermin Wei and Professor Randall Berry, for giving me many valuable suggestions and advice that have benefited me a lot.

I am very thankful to all the colleagues and friends in the Communications and Networking Laboratory here at Northwestern University, including Binnan Zhuang, Fei Teng, Khalid Zeineddine, Xu Chen, Chang Liu, Tho Ngoc Le, Su Yan, Liwen Liang, Ryan Keating, Hao Ge, Ruijie Xu, Zeyu Zhang, Xu Wang, Hao Zhou, Jing Li, Ding Xiang, Yining Zhu, Yasar Sinan Nasir, and Haoran Yu. With all of you, the Commnet group is like a warm family, and I have enjoyed five years of happiness. Our friendship never fades.

Finally, and most importantly, I want to thank my parents and my wife, for their endless love and support; I could not have accomplished any of this without them. To them I dedicate this dissertation.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

The evolution to the fifth-generation (5G) wireless networks is expected to enable a host of new services and applications that will place increasing pressure on physical resources. The desire to increase capacity and enhance coverage is motivating trends to deploy denser cells with more antennas, and to seek new spectrum opportunities by moving to higher frequencies and sharing with different applications (e.g., radar) [**1**–**3**]. Those trends will introduce a high degree of network heterogeneity: macro/micro/pico access points (APs) will vary according to spectrum assignments, energy requirements, cost, and technology (e.g., WiFi open access, cellular, cm/mm-wave). Demand for services will also be heterogenous, mixing different rate, latency, and mobility requests from many user devices over time and spatial locations. The industry (3GPP) is currently developing 5G New Radio, which uses a unified, scalable, slot-based air interface across all frequency bands to support a wide range of services from massive Internet of Things to enhanced mobile broadband to mission-critical services [**4**, **5**].

A wireless network operator needs to plan, design, and deploy an infrastructure, and then allocate appropriate physical resources (including frequency, time, and power) to maximize user experience and minimize cost. The allocation problem is very challenging due to their scale and dynamics. In particular, a large number of APs need to be deployed densely to provide coverage and capacity; yet nearby transmitters will interfere

with each other unless they avoid using the same time-frequency slot. Especially in heterogeneous networks (HetNets), with many closely located small cells serving dynamic traffic, strong time-varying inter-cell interference is to be expected. Hence traditional cell planning, frequency reuse and power control can be ineffective. New interference models and interference management schemes are highly desired for HetNets.

Resource allocation in wireless networks has been extensively studied over the last few decades. Spectrum allocation schemes are basically designed either from an economic perspective (e.g., see [**8**, **9**]) or from a technical perspective (e.g., see [**11**–**25**, **27**, **32**–**36**, **42**]). As for the user association problem, NP-hard nonconvex integer programming problems are often formulated [**20**, **21**], which are typically hard to obtain a local optimum, not to mention a global optimum. Similar to the user association problem, power control is also known to be a hard non-convex optimization problem due to the complicated interference coupling between links. Most importantly, there have been only a few works studying the resource allocation problem in a large-scale network with hundreds or even thousands of APs. Because centralized algorithms are usually not scalable to large-scale networks, most resource allocation algorithms are designed in a distributed manner. However, this is at the sacrifice of system performance and it is even worse for large-scale networks.

To summarize, the urgent need for redesigning resource allocation schemes and interference management motivates the following fundamental questions. For large-scale networks: (1) How to jointly optimize resource allocation? (2) What is the gap between such optimized allocation scheme and a simple one (e.g., full spectrum reuse)? (3) What

is the performance gain from joint consideration of spectrum allocation and user association? (4) How to design scalable centralized resource allocation scheme (5) How should the interference be managed if power control is added?

The objective of this thesis is to formulate the preceding general challenge as a class of optimization problems and to provide computationally tractable resource management methods. A distinguishing feature here is that we propose to use a central controller or cloud to coordinate a metropolitan-scale network with many hundred APs. This architecture can fully harness the power of cloud computing, big data and large-scale optimization methods. Specifically, APs measure/estimate traffic and channel conditions and send the cloud regular updates. The cloud periodically solves a large-scale optimization problem for the whole network and returns the resulting allocation scheme to the APs.

We next summarize the contributions to large-scale resource management, which are presented in Chapters 2 to 4. Conclusion and future research directions will be discussed in Chapter 5.

1.1. Licensed and Unlicensed Spectrum Allocation and User Association

In order to support fast growing mobile data traffic, wireless operators have started to use multiple radio access technologies (RATs) over multiple licensed and unlicensed frequency bands. A widely used method is WiFi off-loading of cellular network traffic. A recent proposal is side-by-side deployment of Long Term Evolution (LTE) in licensed spectrum and LTE in unlicensed spectrum (LTE-U). In this chapter, we studied the spectrum allocation problem in HetNets. For practical reasons, the allocation is conceived to be on a relatively slow timescale. Such centralized control is on a relatively slow

timescale to allow information exchange and joint optimization over multiple cells. This is in contrast and complementary to distributed scheduling on a fast timescale.

A queueing model is introduced for the unlicensed band to capture its lower spectral efficiency, reliability, and additional delay due to contention and/or listen-before-talk requirements. Under mild assumptions, the spectrum allocation problem is formulated as a bi-convex optimization problem. Solving this problem gives an effective and computationally efficient solution for both user association and spectrum allocation over multiple RATs.

Two optimization-based spectrum allocation schemes are proposed along with efficient algorithms for computing the allocations. The proposed solutions take into account traffic loads, network topology, as well as external interference levels in the unlicensed bands.

Simulation results show that in the heavy-traffic regime, the proposed scheme significantly outperforms both orthogonal and full-frequency-reuse allocations. In addition, the solution to the optimization problem matches the intuition that users with relatively higher traffic demand are mostly assigned to the licensed spectrum, while those with lower traffic demand and less exogenous interference from the unlicensed band are assigned to the unlicensed spectrum.

1.2. Spectrum Allocation and User Association for Large-Scale Networks

As mentioned previously, Due to irregularities of network topology and sophisticated interference conditions, efficient joint spectrum allocation and user association becomes extremely crucial for harnessing the full power of the infrastructure.

In [**34**, **35**], Zhuang et al developed a centralized optimizationbased framework for allocating downlink spectrum resources on a relatively slow timescale. The spectrum allocation and user association problem was formulated as a convex optimization problem where the global optimal solution can be obtained using a standard solver. However, the number of variables in the problem grows exponentially with the number of APs. The space and time complexities of solving the problem for a large network of hundreds of APs become prohibitive.

To address the preceding challenges, we derive an equivalent reformulation of the fundamental resource allocation and user association problem from the viewpoints of user devices. Such user-centric reformulation captures the fact that each user device's performance depends only on the interference pattern of no more than a constant number of APs in the user device's neighborhood. This allows a low-complexity reformulation of the global problem, which reduces the total number of variables from exponential to quadratic in the number of user devices. Moreover, a scalable reformulation is obtained by exploiting the geometric graph nature of the network and provable sparsity of the optimal solution. A pattern pursuit algorithm with low complexity is proposed to solve the reformulated problem with guaranteed gap to the global optimum. Efficient algorithms are developed to obtain near-optimal allocations for a network with up to 1,000 APs and 2,500 active users. Numerical results show that the proposed solution achieves significant gains in terms of delay and throughput over existing schemes and is within 7% to the global optimum in a typical scenario.

1.3. Joint Spectrum Allocation, User Association, and Power Control for Large-Scale Networks

Due to limited resources in current wireless networks, efficient resource allocation (e.g., spectrum allocation, power control, link scheduling, routing, and congestion control) is crucial to achieving high performance and providing satisfactory quality-of-service (QoS). Conventional spectrum allocation schemes include full spectrum reuse, partial frequency reuse, and fractional frequency reuse [**28**, **30**, **31**]. However, the spectrum is either under-utilized or over-utilized under these schemes, resulting in either low spectral efficiency or strong link interference. Moreover, these schemes are not traffic-aware, that is, they are not adaptive to different traffic conditions. Considering the user association schemes used in the industry, most of them are based on simple heuristics which waste precious system capacity and offer no performance guarantees. For example, by default, in today's cellular and WiFi networks devices simply associate with the base station from which they receive the strongest signal. Power control is another way of mitigating inter-cell interference as well as saving energy. Instead of letting each AP always transmit using full power, some eICIC techniques, such as the almost blank subframe (ABS) control, have been proposed in LTE and LTE-A. However, the performance of such kind of distributed algorithms is far from optimal especially for large networks.

There have been a lot of works addressing the preceding challenges. To use spectrum more efficiently, opportunistic dynamic spectrum allocation methods are discussed [**46**, **47**]. Schemes based on game theory [**48**–**51**] or optimization [**52**, **53**] are also well explored. However, the maximum spectrum flexibility or utilization is hard to achieve. In most formulations, subbands are predefined, APs can not dynamically split the spectrum

according to user device's traffic demand or the channel state. Also, distributed methods are proposed to allocate spectrum at the expense of optimal solution. User association is also a challenging problem in wireless network. In practice, users are associated to the AP with maximum reference SINR. As for power management, proper assignments not only save power but also avoid unnecessary interference. Numerous power saving models have been proposed [**?**, **54**, **55**]. However, power management problems are often non-convex. Their solutions are either computationally expensive or far from optimal.

In this chapter, we extends the work in Chapter 3 to jointly solve the problems of spectrum allocation, user association, and power control for super large networks within small amount of time. An efficient algorithm with low complexity is proposed to obtain a locally-optimal allocation solution. One key distinction of the problem formulation in this chapter is that, instead of assuming all APs always transmit using full power, we allow each AP to apply a continuous power spectral density over the entire band, i.e., the size of the problem of power allocation at each AP has already grown to infinity. Moreover, we also generalize the problem formulation to accommodate various classes of network optimization problems.

CHAPTER 2

Licensed and Unlicensed Spectrum Allocation and User Association

In future networks, an operator may employ a wide range of APs using diverse RATs over multiple licensed and unlicensed frequency bands. This chapter studies centralized user association and spectrum allocation across many APs in such a HetNet. Such centralized control is on a relatively slow timescale to allow information exchange and joint optimization over multiple cells. This is in contrast and complementary to distributed scheduling on a fast timescale. A queueing model is introduced to capture the lower spectral efficiency, reliability, and additional delays of data transmission over the unlicensed bands due to contention and/or listen-before-talk requirements. Two optimization-based spectrum allocation schemes are proposed along with efficient algorithms for computing the allocations. The proposed solutions take into account traffic loads, network topology, as well as external interference levels in the unlicensed bands. Packet-level simulation results show that the proposed schemes significantly outperform orthogonal and full-frequency-reuse allocations under all traffic conditions.

2.1. Introduction

The wireless industry has set an ambitious goal to increase the area capacity (in bits per second per square meter) by three orders of magnitude in the next five to ten years. In addition to densely deploying small cells and improving the spectral efficiency [**7**], another

avenue is to exploit all available spectrum, which is a relatively scarce resource [**1**]. Future generation cellular networks are likely to involve multiple RATs over multiple frequency bands (including millimeter wave). Such RATs include LTE, WiFi, and LTE-U. The prime bands today are licensed frequency bands under 3 GHz and over 500 MHz of unlicensed spectrum in the 2.4 GHz and 5 GHz frequency bands.

Current 4G cellular networks generally use regular frequency reuse patterns, including full frequency reuse and fractional frequency reuse. In the former scheme, every cell uses all available frequency bands. The latter scheme is similar, except that cells use orthogonal frequency bands at the cell edge to reduce mutual interference. Such simple methods are unlikely to be as effective in emerging HetNets with highly irregular topologies and and widely varying traffic conditions across the cells.

There have been many studies on spectrum allocation in cellular networks. Some work studies the allocation problem from an economic perspective (e.g., see [**8**, **9**]), whereas other work investigates this problem from a technical perspective (e.g., see [**10**–**13**, **15**–**22**, **24**, **25**]). Most authors formulate the allocation problem as that of deciding, for each slice of the spectrum, which APs and/or links should use it. In addition, user association is considered in [**20**, **21**]. In general, NP-hard nonconvex integer programming problems are formulated, which typically have many local optima. Furthermore, the figures of merit used in most work are physical layer performance measures such as sum rate and outage probability. These traffic-independent metrics may not reflect a user device's relevant QoS in HetNets with large traffic variations in overlapping cells with complicated interference conditions.

To address the preceding issues, an alternative optimization-based framework was developed in [**34**] and [**35**] for allocating downlink spectrum resources in a HetNet. This framework allows arbitrary user association and flexible spectrum allocation with input parameters given by traffic loads over the geographic area. This framework is well matched to emerging centralized remote processing architectures such as cloud radio access network (C-RAN) and cell-free massive multiple-input multiple-output systems [**40**]. In contrast to most existing work [**13**, **15**–**18**], which considers resource allocation on the timescale of a frame, here the timescale of resource adaptation is conceived to be once every few seconds or minutes. This timescale is, on the one hand, fast enough for tracking aggregate traffic variations and large-scale fading, and, on the other hand, slow enough to allow information exchange and joint optimization of many APs with a large number of user devices. Another advantage is the spectrum resources can be assumed to be homogeneous over this timescale, namely, every segment in the same band has about the same spectral efficiency when averaged over many frames. The approach in [**34**] has been generalized to incorporate energy-efficient allocation via cell activation [**35**] as well as the effect of opportunistic scheduling on a fast timescale [**41**, **42**].

In this work we generalize the framework of [**34**] and [**35**] to the scenario in which there are multiple RATs over multiple frequency bands. Resource allocation over multiple RATs has been studied in [**43**–**45**]. In [**43**], the authors presented a scheme for balancing licensed and unlicensed traffic in the case of a single femto user device and single WiFi user device. In [**44**, **45**], joint licensed and unlicensed resource allocation algorithms were proposed for licensed-assisted access systems for throughput and energy efficiency maximization, respectively. However, the distinct characteristics of the unlicensed bands were

not modeled in those papers. In this work, one of the main contributions is to consider different queueing models to distinguish the characteristics of licensed and unlicensed bands. The multiple-band allocation problem is formulated as a bi-convex optimization problem. A conservative allocation scheme is first proposed following the approach of [**34**] and [**35**]. This is followed by a utilization-dependent allocation scheme which incorporates the utilizations of the APs into the formulation to more accurately account for dynamic inter-cell interference. An iterative algorithm is designed to solve the allocation problem with manageable computational complexity for small systems. In each step, we solve a convex problem with a unique optimal solution.

Another feature of this work is that the packet length is allowed to have a general probability distribution. In prior work [**34**, **35**, **41**, **42**], the packet length is assumed to be exponentially distributed to yield a simple analytic form in the objective representing the QoS. We show that the proposed formulation and algorithm apply to general packet length distributions and hence a broader class of traffic conditions.

The performance of the resource allocation methods is evaluated by packet-level simulations. It is shown that both the conservative allocation scheme and the utilization-based allocation scheme significantly reduce the average packet delay in the heavy traffic regime compared to orthogonal allocation and full-frequency-reuse allocation. The large performance gain is observed mainly because the proposed schemes are traffic-aware and also exploit the particular characteristics of each RAT. In addition, the utilization-dependent allocation scheme attains the best performance over all traffic conditions due to its accurate modeling of the dynamic interference among APs.

The centralized approach to resource allocation presented here is compatible with the emerging C-RAN, which allows many remote radio function units to connect to a centralized network controller. The total overhead for the network controller to perform the proposed resource allocation scheme includes collecting the spectral efficiencies of all links and user devices' traffic information. Since the timescale of resource adaptation is considered to be once every a few seconds or minutes, the overhead is quite small. For example, the rate for sending 30,000 parameters (16 bits each) every minute is only 8 kilobits per second (kbps).

The remainder of this chapter is organized as follows. The system model is introduced in Section 2.2. An optimization problem using the conservative allocation is formulated in Section 2.3. A utilization-dependent allocation scheme is presented in Section 2.4. The extension to general packet length distributions is given in Section 2.5. Simulation results are presented in Section 2.6 and concluding remarks are given in Section 2.7. All technical proofs are relegated to the appendices.

2.2. System Model

In this section, we introduce models for user traffic, spectrum allocation, and link throughput. The models extend those in [**35**] to accommodate multiple RATs. A new queueing model is then introduced for traffic over unlicensed bands. Table 2.1 summarizes all the symbols used throughout the chapter.

Table 2.1. Descriptions of symbols.

Symbol	Description
$w^{(l)}$	Bandwidth of the spectrum used by RAT l
$y^{A,l}$	Fraction of spectrum assigned to RAT l shared by APs in set A
$x_{i\to j}^{A,l}$	Fraction of spectrum assigned to the link $i \to j$ over pattern A under RAT l
$s_{i\to j}^{A,l}$	Spectral efficiency of link $i \to j$ over pattern A under RAT l
L	Average packet length in bits
$\alpha^{(l)}$	Discount factor for RAT l
$p_i^{(l)}$	Power spectral density of AP i over RAT l
$I_{i\to j}^{A,l}$	Total noise plus interference power spectral density from APs other than AP i in A to user device group j
$r_j^{(l)}$	Service rate of user device group j over RAT l using the conservative allocation
$\lambda_j^{(1)}$	Packet arrival rate of user device group j under LTE
$\lambda_j^{(2)}$	Packet arrival rate of user device group j under LTE-U
$t_j^{(1)}$	Average packet delay of user device group j under LTE
$t_j^{(2)}$	Average packet delay of user device group j under LTE-U
ν_j	Expected square vacation duration of user device group j under LTE-U
$r_j^{I,l}$	Service rate of user device group j for interfering APs I over RAT l using the utilization-dependent model
$\rho_i^{(l)}$	Utilization of AP i over RAT l
$p^{I,l}$	Probability of an active interfering AP set I over RAT l
$\sigma_j^{(l)}$	Average utilization of user device group j

2.2.1. Spectrum Allocation

We consider the downlink of a HetNet consisting of n APs and many user devices. Without loss of generality, suppose each AP can use all m different RATs, with each RAT on a separate frequency band. Since user devices located near each other often have similar channel conditions, these user devices can be treated as a group on the slow timescale. This is a generalization of the extreme case where each user device group contains only one single user device. Moreover, it suffices to carry out the slow timescale allocation to the user device groups rather than individual user devices.

Denote the set of all APs by $N = \{1, ..., n\}$, the set of all user device groups by $K = \{1, ..., k\}$, and the set of RATs by $M = \{1, ..., m\}$. RAT l employs its separate homogeneous spectrum of bandwidth $w^{(l)}$. We allow arbitrary association so that each

AP can simultaneously serve any subset of user device groups and each user device can be simultaneously served by any subset of APs. Furthermore, we allow flexible resource allocation in that each AP-user link can use an arbitrary number of RATs, where each RAT uses an arbitrary (possibly discontinuous) subset of the available spectrum. Despite the enormous number of possibilities, we will show that the actual AP-user association and spectrum allocation is extremely sparse.

The key to total spectrum agility is the notion of *pattern* [**34**,**35**]. In general, a pattern simply refers to a subset of transmitters. A time-frequency resource is said to be reserved for pattern A if the resource is to be shared by transmitters in A.[1] In the downlink, a pattern A is a subset of N, and all APs in A have access to the time-frequency resources associated with pattern A. Assuming known transmit power spectral densities (PSDs), the pattern of a resource determines the signal-to-interference-plus-noise ratio (SINR) and hence the spectral efficiencies of all links over the resource. The allocation problem can then be formulated as how to divide the spectrum of each RAT among all 2^n-1 nonempty patterns. To illustrate the concept of pattern, Fig. 2.1 shows an example with three APs operating over one frequency band. The spectrum can be divided into $2^3-1=7$ segments, where one segment is used by AP 1 exclusively (the pattern is $\{1\}$), a second is used by AP 2 exclusively (the pattern is $\{2\}$), a third is used by AP 3 exclusively (the pattern is $\{3\}$) and the remaining four segments include three shared by the two APs (the patterns are $\{1,2\}$, $\{2,3\}$, and $\{1,3\}$, respectively) as well as one segment shared by all three APs (the pattern is $\{1,2,3\}$). If the transmit pattern of a certain spectrum resource is $\{1,2\}$,

[1]The notion of pattern finds its root in the concept of independent set defined in the special case where the network is described by a conflict graph, and where nodes/links in the independent set share the same resources [**58**].

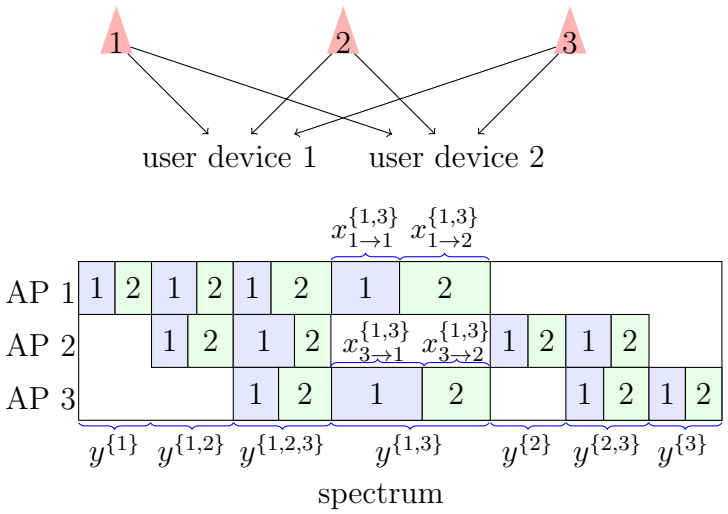


Figure 2.1. Illustration of a general spectrum allocation in a 3-AP 2-user network. The y variables corresponding to all 7 non-empty patterns are shown. The x variables under pattern $\{1, 3\}$ are also shown explicitly.

i.e., both APs transmit, then that determines the associated spectral efficiencies of all links, as discussed in what follows.

A network controller collects traffic load and channel/interference information from all the APs. Because the time period of slow timescale allocation is much longer than the channel coherence time, the channel conditions are modeled using path loss and the statistics of fading. Given the average channel and traffic conditions, the task of the central controller is to determine which spectrum segments to allocate to each AP, and subsequently, which sub-segments to allocate to specific user devices associated with that AP. To be precise, we need to solve the following three subproblems:

(1) Decide which RAT or RATs should be used to serve each user device group.

(2) Allocate the bandwidth across all 2^n patterns for each RAT. This is denoted by a $2^n \times m$-tuple: $y = (y^{A,l})_{A\subset N, l\in M}$, where $y^{A,l} \in [0, 1]$ is the fraction of spectrum

assigned to RAT l shared by APs in set A. Clearly,

$$\sum_{A \subset N} y^{A,l} = 1, \quad \forall l \in M \tag{2.1}$$

and an efficient allocation does not use the empty pattern, so that $y^{\varnothing,l} = 0$, for every $l \in M$.

(3) For every pattern $A \subset N$, for every AP $i \in A$, divide the bandwidth $y^{A,l}$ to the user device groups. Denote the bandwidth allocated to the link $i \to j$ over pattern A under RAT l as $x_{i \to j}^{A,l}$. As shown in Fig. 2.1, $y^{\{1,3\}}$ (colored yellow) is further divided by AP 1 into two parts to serve user device group 1 and user device group 2, respectively, while AP 3 divides the same shared spectrum differently. Then we have:

$$\sum_{j \in K} x_{i \to j}^{A,l} \leq y^{A,l}, \quad \forall i \in A, A \subset N, l \in M. \tag{2.2}$$

The user association is indirectly determined by the amount of spectrum resources assigned by each AP to each group. Specifically, user device j is assigned to AP i over RAT l if and only if $x_{i \to j}^{A,l} > 0$ for some pattern A.

2.2.2. From Spectrum to Transmission Rates

An important measure of user device QoS is the average packet delay in the system. The delay is determined by the packet arrival statistics and the service rates. The instantaneous service rate of a user device group's queue in turn depends on the spectral efficiencies along with the spectrum allocated to that specific user device group. For

simplicity, it is assumed that when AP i transmits over RAT l, it employs all patterns available to it and applies a flat PSD $p_i^{(l)}$ over the allocated spectrum. At any frequency designated for RAT l, the instantaneous spectral efficiency achievable by the link from AP i to user device j depends on the set of active APs $A \subset N$ using that frequency. Let this (spectral) efficiency be denoted by $s_{i\to j}^{A,l}$. Clearly, $s_{i\to j}^{A,l} = 0$ if $i \notin A$. Moreover, the spectral efficiency decreases as more APs become active, i.e., $s_{i\to j}^{A,l} \geq s_{i\to j}^{B,l}$ if $i \in A \subset B$. On the slow timescale considered, the spectral efficiencies are either known *a priori* or can be measured and sent to the central controller. For later convenience, we normalize the spectral efficiency over RAT l using factor $L/w^{(l)}$ (bits/second/Hz) so that the unit of $s_{i\to j}^{A,l}$ is packets/second.

For concreteness in obtaining numerical results, we use Shannon's formula for the link efficiencies:

$$s_{i\to j}^{A,l} = \frac{\alpha^{(l)} w^{(l)} \mathbf{1}(i \in A)}{L} \log_2 \left(1 + \frac{p_i^{(l)}}{I_{i\to j}^{A,l}}\right) \quad \text{packets/s}, \tag{2.3}$$

where L is the average packet length in bits,[2] $\alpha^{(l)} \in (0, 1]$ is the discount factor for RAT l, $\mathbb{1}(i \in A) = 1$ if $i \in A$ and $\mathbb{1}(i \in A) = 0$ otherwise, and $I_{i\to j}^{A,l}$ is the total noise plus interference PSD from other APs in A to user device j, which depends on their transmit PSDs and path loss. The discount factor of a licensed band is typically closer to one than that of an unlicensed band due to external interference from other operators in the unlicensed band. The discount factor is not crucial and is merely included for flexibility of the model. The average effect of small-scale fading can be included by considering the ergodic capacity in lieu of (2.3), which does not change the main developments.

[2]Packet lengths are assumed to be i.i.d. with exponential distribution; later we will show that the proposed methods can be applied to general packet length distributions.

Over a resource reserved for pattern A, only the subset of APs in A with data to transmit will be using the resource at any given time. In general, the instantaneous service rate depends on the set of transmitting APs and the rate adaptation scheme. Two service rate models are described next.

Under the conservative allocation introduced in [**34**], if some of RAT l's spectrum resources reserved for pattern A are allocated to link $i \to j$, then AP i transmits user device j's packets at spectral efficiency $s_{i\to j}^{A,l}$ using RAT l over pattern A. This rate is achievable even if all APs in A have data to transmit (hence the name conservative). This is equivalent to assuming that other APs' traffic is always backlogged. The rate contributed by spectrum reserved for pattern A under RAT l is the product of the spectral efficiency and the bandwidth: $s_{i\to j}^{A,l} x_{i\to j}^{A,l}$. The total service rate for the queue of user device j over RAT l is the sum rate of all APs and patterns, expressed as:

$$r_j^{(l)} = \sum_{A \subset N} \sum_{i \in N} s_{i\to j}^{A,l} x_{i\to j}^{A,l} \quad \text{packets/s.} \tag{2.4}$$

The advantage of the conservative allocation is that there is no need for the scheduler (or whichever unit that performs rate adaptation) to know the state of the other APs included in the pattern. However, the conservative rate (2.4) is a lower bound on the actual rate because the interference is overestimated.

To adapt the allocation to the actual interference pattern, we introduce a new model, referred to as the *utilization-dependent model*. Under this model, the service rate contributed by the spectrum reserved for pattern A depends on the subset of transmitting APs in A. Specifically, the service rate for user device group j over RAT l when the set

of active interfering APs is I is expressed as:

$$r_j^{I,l} = \sum_{A \subset N} \sum_{i \in N} s_{i \to j}^{A \cap I, l} x_{i \to j}^{A,l} \quad \text{packets/s.} \tag{2.5}$$

2.2.3. Queueing Model

Without loss of generality, we restrict the subsequent investigation to the scenario with two RATs, where one RAT is over the licensed band (conceived as LTE) and the other is over the unlicensed band (conceived as LTE-U). This can be easily generalized to more than two RATs.

The traffic to user device group j is modeled as an independent Poisson point process with rate λ_j packets per second. The packet lengths are independent random variables whose average is L bits. Packets intended for each user device group are transmitted according to the first-in-first-out (FIFO) discipline. The buffer in each queue is assumed to be unlimited for simplicity. The traffic load of user device group j is further divided into two streams served by the two RATs, respectively. This may be implemented by dividing the group of user devices for association with different RATs, so that each user device is only served by one RAT. The resulting packet arrival rates of two streams are denoted as $\lambda_j^{(1)}$ and $\lambda_j^{(2)}$, respectively, which are variables to be optimized subject to the constraint:

$$\lambda_j^{(1)} + \lambda_j^{(2)} \geq \lambda_j. \tag{2.6}$$

For each RAT, the corresponding k user device groups form a system of k interactive queues where the instantaneous service rate of each queue in general depends on which other queues are empty.

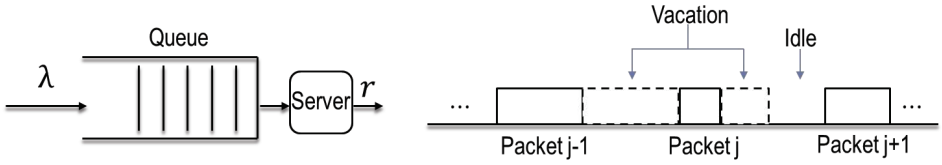


Figure 2.2. Illustration of queueing model for LTE-U.

The physical and/or medium access control layers of LTE-U are designed to facilitate coexistence with other RATs in unlicensed bands, such as WiFi.[3] In [**61**], a listen-before-talk scheme is proposed in which carrier sensing is embedded in a deterministic portion of a LTE subframe. In [**62**], the authors propose the use of LTE uplink power control to improve LTE/WiFi coexistence. References [**63**] and [**64**] propose to enable LTE/WiFi coexistence by muting LTE transmission on certain subframes following a pre-determined pattern.

In this chapter, we assume that LTE-U has a listen-before-talk feature, which is likely to be a dominant mode in emerging LTE-U standards [**59**]. In this mode, an AP with data to send over LTE-U first performs carrier sensing before transmitting. We model this as a queue with vacation and non-exhaustive service [**65**]. The queueing scheme is depicted in Fig. 2.2. Specifically, the server of queue j takes a single vacation after completing the service of each packet. The vacation duration V_j (in seconds) is a random variable with a given distribution. The mean and variance of the vacation time depend on the level of local interference,[4] such as interference from WiFi user devices. The higher the interference level, the more time a queue has to wait before being served.

[3]Challenges of such coexistence are discussed in [**59**, **60**].

[4]Although coexistence issues in the unlicensed band are very important over fast timescales, over slow timescales, effective interference levels and associated packet delays suffice to characterize how LTE-U interacts with other devices in the unlicensed band.

2.3. Queueing Delays and A Conservative Allocation Scheme

In this section, we adopt the conservative service rate model (2.4) to develop a scheme for allocating licensed and unlicensed spectrum. For each RAT, the corresponding k user device groups form k independent $M/M/1$ queues, because the service rate of each queue is independent of the states of other queues under the model (2.4). Let RAT 1 represent LTE and RAT 2 represent LTE-U.

Under LTE, the average packet delay (in seconds/packet) of user device group j is given by the average delay for the $M/M/1$ queue:

$$t_j^{(1)} = \frac{1}{\left(r_j^{(1)} - \lambda_j^{(1)}\right)^+}, \tag{2.7}$$

where $(x)^+ = x$ if $x > 0$ and $(x)^+ = 0$ if $x \leq 0$. If $r_j^{(1)} \leq \lambda_j^{(1)}$, the queueing delay is infinite, i.e., the queue becomes unstable. It is important to note that the right hand side of (7) is convex in the pair $(r_j^{(1)}, \lambda_j^{(1)})$ on $\mathbb{R}^2$.

Under LTE-U, the k user device groups form k independent $M/M/1$ queues with single vacation. We have:

Proposition 2.1. *: The average packet delay (in seconds/packet) of a queue with Poisson arrival rate λ and a single server rate r with single vacation V having expected squared duration $\nu = E[V^2]$ is given by:*

$$t = \frac{2 + r\lambda\nu}{2\,(r - \lambda)^+}. \tag{2.8}$$

For fixed ν, the function defined by 2.8 is convex in the pair (r, λ) on $\mathbb{R}^2$.

The proof of Proposition 2.1 is given in Appendix A.1.

The joint user association and conservative spectrum allocation problem is formulated as Problem P2.1:

$$\underset{\boldsymbol{r},\boldsymbol{x},\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{t}}{\text{minimize}} \quad \frac{1}{\sum_{j\in K}\lambda_j}\sum_{j\in K}(\lambda_j^{(1)}t_j^{(1)}+\lambda_j^{(2)}t_j^{(2)}) \tag{P2.1a}$$

$$\text{subject to} \quad t_j^{(1)} = \frac{1}{\left(r_j^{(1)}-\lambda_j^{(1)}\right)^+}, \qquad j\in K \tag{P2.1b}$$

$$t_j^{(2)} = \frac{2+r_j^{(2)}\lambda_j^{(2)}\nu_j}{2\left(r_j^{(2)}-\lambda_j^{(2)}\right)^+}, \qquad j\in K \tag{P2.1c}$$

$$r_j^{(l)} = \sum_{A\subset N}\sum_{i\in N} s_{i\to j}^{A,l}x_{i\to j}^{A,l}, \quad j\in K, l\in\{1,2\} \tag{P2.1d}$$

$$\lambda_j^{(1)}+\lambda_j^{(2)} \geq \lambda_j, \qquad j\in K \tag{P2.1e}$$

$$y^{A,l} \geq \sum_{j\in K} x_{i\to j}^{A,l}, \quad l\in\{1,2\}, i\in N, A\subset N \tag{P2.1f}$$

$$\sum_{A\subset N} y^{A,l} = 1, \qquad l\in\{1,2\} \tag{P2.1g}$$

$$x_{i\to j}^{A,l} \geq 0, \quad j\in K, l\in\{1,2\}, i\in N, A\subset N. \tag{P2.1h}$$

The objective (P2.1a) is the average packet delay of all queues over the entire network. (P2.1b) is the average packet delay of each user device group served by LTE as derived in (2.7). (P2.1c) is the average packet delay of each user device group served by LTE-U as derived in (2.8). (P2.1d) is the service rate of each divided user device group using the conservative model as given in (2.4). (P2.1e) is the total traffic constraint for each user

type (2.6). (P2.1f) guarantees the consistency of the spectrum allocation as given in (2.2). (P2.1g) constrains the total bandwidth of each RAT to be one unit. (P2.1h) precludes negative bandwidth. The variables in P2.1 are $\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{t}$ which are the vector forms of $(r_j^{(l)})_{j\in K, l\in\{1,2\}}$, $(x_{i\to j}^{A,l})_{j\in K, i\in N, A\subset N, l\in\{1,2\}}$, $(y^{A,l})_{A\subset N, l\in\{1,2\}}$, $(\lambda_j^{(l)})_{j\in K, l\in\{1,2\}}$, $(t_j^{(l)})_{j\in K, l\in\{1,2\}}$, respectively.

P2.1 is a bi-convex optimization problem because when variable $\boldsymbol{\lambda}$ (resp. $\boldsymbol{x}$) is fixed, all constraints are linear and the objective is a linear combination of convex functions in $\boldsymbol{x}$ (resp. $\boldsymbol{\lambda}$). Therefore, P2.1 can be solved by alternating optimization over $\boldsymbol{\lambda}$ and $\boldsymbol{x}$. Given $\boldsymbol{\lambda}$, there are $2(2k+nk2^n+2^n)$ variables, and given $\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t}$, there are $2k$ variables. Since the sequence of objective values obtained at each iteration is lower-bounded and non-increasing, the objective converges although it may not be a global minimum. As shown in Section 2.6, the alternating method achieves good performance.

Theorem 2.1. *: The (global) minimum average delay can be achieved by a sparse allocation, where the spectrum of each RAT is divided into at most k segments. That is, there exists an optimal solution that satisfies*

$$|\{A \mid y^{A,l} > 0, A \subset N\}| \leq k, \qquad \forall l \in M \tag{2.9}$$

where $|\cdot|$ denotes the cardinality of a set. The same applies to the delay achieved by the alternating optimization method.

Theorem 2.1 is proved in Appendix A.2. The theorem guarantees that although the number of all possible patterns grows exponentially with the number of APs in the network, using a small number of patterns achieves the optimal performance.

Theorem 2.2. *: Assume the channel gains are jointly continuous random variables. The (global) minimum average delay can be achieved with at most $n-1$ user device groups served by multiple APs on each RAT. That is, the optimal solution satisfies: for every $l \in M$,*

$$\left|\{j \mid x_{i_1\to j}^{A_1,l}, x_{i_2\to j}^{A_2,l} > 0, \text{ for some } i_1, i_2 \in N \text{ and } A_1, A_2 \subset N\}\right| \leq n-1. \tag{2.10}$$

The same applies to the delay achieved by the alternating optimization method.

Theorem 2.2 is proved in Appendix A.3. The theorem guarantees that although we allow a user device group to be served by multiple APs, most user device groups will be associated with only one AP in the optimal solution.

2.4. A Utilization-Dependent Allocation Scheme

In this section, we adopt the utilization-dependent service rate model (5). Unlike in the conservative scheme, we now let each AP adapt its transmission rates to the instantaneous set of active interfering APs. Thus the service rate is in general higher than the conservative rate.

In a stable interactive queueing system, each AP transmits (over each RAT) for a fraction of the time, referred to as the *utilization*. Denote the utilization of AP i over RAT l as $\rho_i^{(l)} \in [0,1]$. The analysis of interactive queueing system is difficult. As an approximation, we assume that different APs transmit independently over each RAT. The probability of an active interfering AP set I over RAT l is then:

$$p^{I,l} = \left(\prod_{i\in I} \rho_i^{(l)}\right)\left(\prod_{i'\notin I}(1-\rho_{i'}^{(l)})\right). \tag{2.11}$$

When user device group j is served, its instantaneous service rate is one of 2^n possible values depending on the set of active APs. Thus we use a certain "average" of 2^n independent $M/M/1$ queues to approximate the queueing behavior of user device group j with interactive queues. Such an approximation is reasonable under the premise that for any AP, the influence of other APs is adequately represented by their steady state probability distribution. Under each RAT, the average delay of user device group j is calculated as the expected delay over 2^n possible rates:

$$t_j^{(1)} = \sum_{I \subset N} \frac{p^{I,1}}{(r_j^{I,1} - \lambda_j^{(1)})^+} \tag{2.12}$$

$$t_j^{(2)} = \sum_{I \subset N} \frac{p^{I,2}(2 + r_j^{I,2}\lambda_j^{(2)}\nu_j)}{2(r_j^{I,2} - \lambda_j^{(2)})^+}. \tag{2.13}$$

The utilization of user device group j is calculated as its expected utilization over all possible sets of interfering APs. Specifically, when the active set of APs is I, the fraction of time that group j is served is $\frac{\lambda_j^{(l)}}{r_j^{I,l}}$, and the average utilization is obtained as:

$$\sigma_j^{(l)} = \sum_{I \subset N} p^{I,l} \frac{\lambda_j^{(l)}}{r_j^{I,l}}, \qquad l \in \{1, 2\}. \tag{2.14}$$

Since AP i may serve multiple user device groups over each RAT, the utilization of AP i depends on the utilization of its associated user device groups. We approximate the utilization of AP i over RAT l as its average utilization over the spectrum used. Specifically, the average amount of spectrum used by AP i to serve its user devices is

$\sum_{j\in K}\sigma_j^{(l)}\sum_{A:i\in A}x_{i\to j}^{A,l}$. Hence AP i's utilization is approximated as

$$\rho_i^{(l)}=\frac{1}{\sum_{A:i\in A}y^{A,l}}\sum_{A:i\in A}\sum_{j\in K}\sigma_j^{(l)}x_{i\to j}^{A,l}, \tag{2.15}$$

where $\sum_{A:i\in A}y^{A,l}$ is the total bandwidth used by AP i over RAT l.

With the preceding approximation, the joint user association and utilization-dependent spectrum allocation problem is formulated as P2.2:

$$\underset{\boldsymbol{r},\boldsymbol{x},\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{t},\boldsymbol{\sigma},\boldsymbol{\rho},\boldsymbol{p}}{\text{minimize}} \quad \frac{1}{\sum_{j\in K}\lambda_j}\sum_{j\in K}(\lambda_j^{(1)}t_j^{(1)}+\lambda_j^{(2)}t_j^{(2)}) \tag{P2.2a}$$

$$\text{subject to} \quad t_j^{(1)}=\sum_{I\subset N}\frac{p^{I,1}}{\left(r_j^{I,1}-\lambda_j^{(1)}\right)^+}, \qquad j\in K \tag{P2.2b}$$

$$t_j^{(2)}=\sum_{I\subset N}\frac{p^{I,2}(2+r_j^{I,2}\lambda_j^{(2)}\nu_j)}{2\left(r_j^{I,2}-\lambda_j^{(2)}\right)^+}, \qquad j\in K \tag{P2.2c}$$

$$r_j^{I,l}=\sum_{A\subset N}\sum_{i\in N}s_{i\to j}^{A\cap I,l}x_{i\to j}^{A,l}, \tag{P2.2d}$$

$$I\subset N, j\in K, l\in\{1,2\} \tag{P2.2e}$$

$$\lambda_j^{(1)}+\lambda_j^{(2)}\geq\lambda_j, \qquad j\in K \tag{P2.2f}$$

$$y^{A,l}\geq\sum_{j\in K}x_{i\to j}^{A,l}, \quad l\in\{1,2\}, i\in N, A\subset N \tag{P2.2g}$$

$$\sum_{A\subset N}y^{A,l}=1, \qquad l\in\{1,2\} \tag{P2.2h}$$

$$x_{i\to j}^{A,l}\geq 0, \quad j\in K, l\in\{1,2\}, i\in N, A\subset N \tag{P2.2i}$$

$$p^{I,l}=\left(\prod_{i\in I}\rho_i^{(l)}\right)\left(\prod_{i'\notin I}(1-\rho_{i'}^{(l)})\right), \tag{P2.2j}$$

$$l \in \{1,2\}, I \subset N \tag{P2.2k}$$

$$\rho_i^{(l)} = \frac{1}{\sum\limits_{A:i\in A} y^{A,l}} \sum_{A:i\in A} \sum_{j\in K} \sigma_j^{(l)} x_{i\to j}^{A,l}, \tag{P2.2l}$$

$$l \in \{1,2\}, i \in N \tag{P2.2m}$$

$$\sigma_j^{(l)} \geq \sum_{I\subset N} p^{I,l} \frac{\lambda_j^{(l)}}{r_j^{I,l}}, \qquad l \in \{1,2\}, j \in K. \tag{P2.2n}$$

The objective (P2.2a) is again the average packet delay of all queues of the entire network. (P2.2d) is the service rate of each divided user device group using the utilization model given by (2.5). The variables in P2.2 are $\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{p}$ which are of the vector forms of $(r_j^{I,l})_{I\subset N, j\in K, l\in\{1,2\}}$, $(x_{i\to j}^{A,l})_{j\in K, i\in N, A\subset N, l\in\{1,2\}}$, $(y^{A,l})_{A\subset N, l\in\{1,2\}}$, $(\lambda_j^{(l)})_{j\in K, l\in\{1,2\}}$, $(t_j^{(l)})_{j\in K, l\in\{1,2\}}$, $(\sigma_j^{(l)})_{j\in K, l\in\{1,2\}}$, $(\rho_i^{(l)})_{i\in N, l\in\{1,2\}}$, and $(p^{I,l})_{I\subset N, l\in\{1,2\}}$, respectively. P2.2 would be equivalent to P2.1 if we let $p^{N,l} = 1$, $p^{I,l} = 0$ for every $I \neq N$, and $\sigma_j^{(l)} = 1$ for every l and j. (This is a feasible suboptimal solution.) Unlike P2.1, P2.2 is not bi-convex because new variables and nonlinear constraints are introduced. However, when fixing the variables $\boldsymbol{\sigma}, \boldsymbol{\rho}$ and $\boldsymbol{p}$, it becomes bi-convex. Therefore, we divide P2.2 into two subproblems and solve them alternately to update all the variables.

Given $\boldsymbol{\sigma}, \boldsymbol{\rho}$ and $\boldsymbol{p}$, we update $\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{t}$ by solving subproblem P2.3:

$$\underset{\boldsymbol{r},\boldsymbol{x},\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{t}}{\text{minimize}} \quad \frac{1}{\sum\limits_{j\in K}\lambda_j} \sum_{j\in K} (\lambda_j^{(1)} t_j^{(1)} + \lambda_j^{(2)} t_j^{(2)}) \tag{P2.3a}$$

$$\text{subject to} \quad \sum_{I\subset N} p^{I,l} \frac{\lambda_j^{(l)}}{r_j^{I,l}} \leq \sigma_j^{(l)}, \quad l \in \{1,2\}; j \in K \tag{P2.3b}$$

$$(P2.2b) - (P2.2i). \tag{P2.3c}$$

(P2.3b) constrains the utilization of user device groups. The structure of subproblem P2.3 is similar to that of P2.1. It can be solved by alternating optimization over $\boldsymbol{\lambda}$ and $\boldsymbol{x}$.

Given $\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{t}$, we update $\boldsymbol{\sigma}, \boldsymbol{\rho}$ and $\boldsymbol{p}$. There are $2(2^n + n + k)$ variables and equations. To solve them with low complexity, Algorithm 2.1 updates $\boldsymbol{\sigma}, \boldsymbol{\rho}$ and $\boldsymbol{p}$ iteratively as in [**41**,**42**]. Here m denotes the iteration. Convergence of the algorithm can be established similarly as in [**41**, **42**]. In addition, the solution obtained by Algorithm 2.1 is feasible for (P2.2j)-(P2.2n).

Algorithm 2.1 Update $\boldsymbol{\sigma}, \boldsymbol{\rho}$ and $\boldsymbol{p}$

Step 1. Initialization: set initial utilization of user device groups $\boldsymbol{\sigma}(0)$;
Step 2. Update:

$$\sigma_j^{(l)}(m) \leftarrow \sum_{I \subset N} p^{I,l}(m-1) \frac{\lambda_j^{(l)}}{r_j^{I,l}},$$

$$\rho_i^{(l)}(m) \leftarrow \frac{1}{\sum_{A: i \in A} y^{A,l}} \sum_{A: i \in A} \sum_{j \in K} \sigma_j^{(l)}(m) x_{i \to j}^{A,l},$$

$$p^{I,l}(m) \leftarrow \left(\prod_{i \in I} \rho_i^{(l)}(m) \right) \left(\prod_{i' \notin I} (1 - \rho_{i'}^{(l)}(m)) \right),$$

$m \leftarrow m + 1$;
Step 3. if $\| \boldsymbol{\sigma}(m) - \boldsymbol{\sigma}(m-1) \| \geq \epsilon$, where ϵ is a fixed threshold, repeat Step 2; otherwise terminate with
$\sigma_j^{(l)} \leftarrow \sigma_j^{(l)}(m)$, $\rho_i^{(l)} \leftarrow \rho_i^{(l)}(m)$, $p^{I,l} \leftarrow p^{I,l}(m)$, $l \in \{1, 2\}, j \in K, i \in N, I \subset N$.

P2.2 is then solved using Algorithm 2.2, which alternates between solving the two corresponding subproblems. Each step in Algorithm 2.2 is an easier problem. Therefore, P2.2 can be solved iteratively with low complexity. Convergence of Algorithm 2.2 can also be similarly established as in [**41**, **42**].

Algorithm 2.2 Iterative algorithm for solving P2.2

Initialization: $\boldsymbol{x} \leftarrow 0; \boldsymbol{x}' \leftarrow 1; \sigma_j^{(l)} \leftarrow 1, \forall j \in K, l \in \{1,2\}$;
$\rho_i^{(l)} \leftarrow 1, \forall i \in N, l \in \{1,2\}; p^{N,l} \leftarrow 1$, $p^{I,l} \leftarrow 0$, $\forall I \subset N$, $I \neq N$.
while $||\boldsymbol{x} - \boldsymbol{x}'|| > \epsilon$ **do**
1. $\boldsymbol{x}' \leftarrow \boldsymbol{x}$;
2. Update $\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{t}$ by solving P2.3;
3. Update $\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{p}$ by using Algorithm 2.1.

end while

2.5. Extension

So far, the packet lengths have been assumed to be i.i.d with exponential distribution, which results in exponentially distributed service times for the packets. In this section, we show that the proposed schemes and algorithms also apply to general packet length distributions as long as the first and second moments of the service time can be written in the form:

$$E[X] = \frac{\beta}{r}, \tag{2.16}$$

$$E[X^2] = \frac{\eta}{r^2}, \tag{2.17}$$

where r is the service rate and β and η are positive numbers. A wide class of distributions have such characteristics. For example, $\beta = 1$ and $\eta = 2$ correspond to exponential service time; $\beta = 1$ and $\eta = 1$ correspond to constant service time. In general, the queues in the system are not $M/M/1$ queues. In contrast to (2.7) and (2.8), the formulas for calculating the average packet delay of user device group j become:

$$\hat{t}_j^{(1)} = \frac{(\frac{1}{2}\eta - \beta^2)\lambda_j^{(1)} + \beta r_j^{(1)}}{r_j^{(1)}\left(r_j^{(1)} - \beta\lambda_j^{(1)}\right)^+}, \tag{2.18}$$

Table 2.2. Parameter Values.

Parameters	Value/Function
AP transmit power	23 dBm
Total bandwidth for each RAT	10 MHz
Average packet length	0.5 Mbits
AP to user device pathloss	$140.7 + 36.7 \log_{10}(R)$

$$\hat{t}_j^{(2)} = \frac{(\frac{1}{2}\eta - \beta^2)\lambda_j^{(2)} + \beta r_j^{(2)} + \frac{1}{2}\nu_j \lambda_j^{(2)} \left(r_j^{(2)}\right)^2}{r_j^{(2)} \left(r_j^{(2)} - \beta \lambda_j^{(2)}\right)^+}. \tag{2.19}$$

Then the objective function in P2.1 is rewritten as:

$$\hat{U} = \sum_{j \in K} (\lambda_j^{(1)} \hat{t}_j^{(1)} + \lambda_j^{(2)} \hat{t}_j^{(2)}). \tag{2.20}$$

Proposition 2.2. *: The objective function (2.20) is bi-convex in* $\boldsymbol{\lambda}$ *and* $\boldsymbol{r}$*.*

The proof is in Appendix A.4. Due to Proposition 2.2, the techniques proposed in Section 2.3 and Section 2.4 still apply to the problem of user association and spectrum allocation with general packet length distribution.

2.6. Numerical Results

Simulations were performed using the network topology depicted in Fig. 2.3. The HetNet is deployed over a 100×200 m^2 area. Five APs, denoted by triangles, and 15 user device groups, denoted by the squares, are randomly dropped within the area. The spectral efficiency is calculated from (2.3) assuming $\alpha^{(1)} = 1$ and $\alpha^{(2)} = 1/2$ with a 30 dB cap on the received SINR. (An SINR greater than 30 dB is regarded as 30 dB.) Other parameters used in the simulation are given in Table 2.2, and are compliant with the LTE

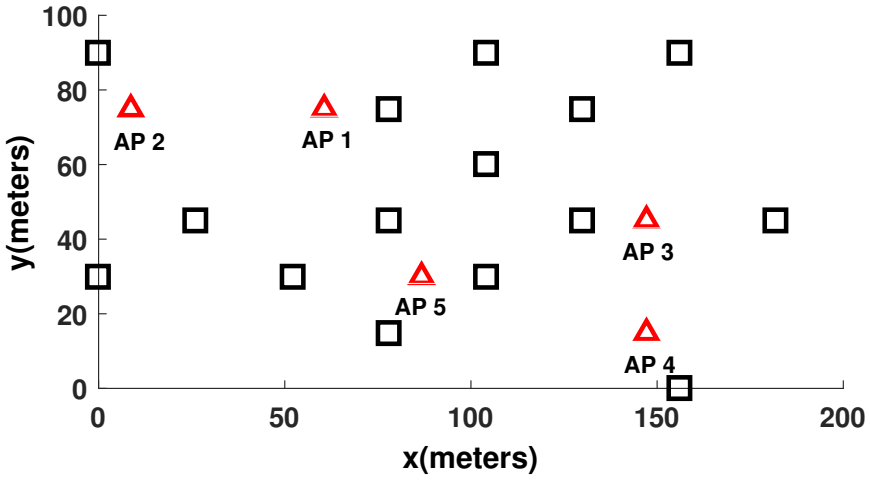


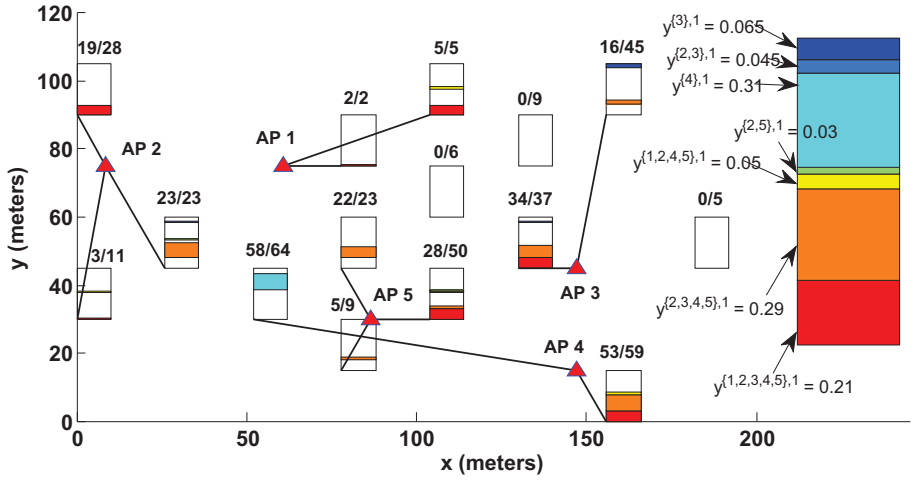
Figure 2.3. Topology of the 5-AP network.

standard[5] [**66**]. The results for actual packet delay with the utilization service rate model were obtained through a packet-level simulator, which determines the transmission time of each packet given the instantaneous set of active interfering APs.

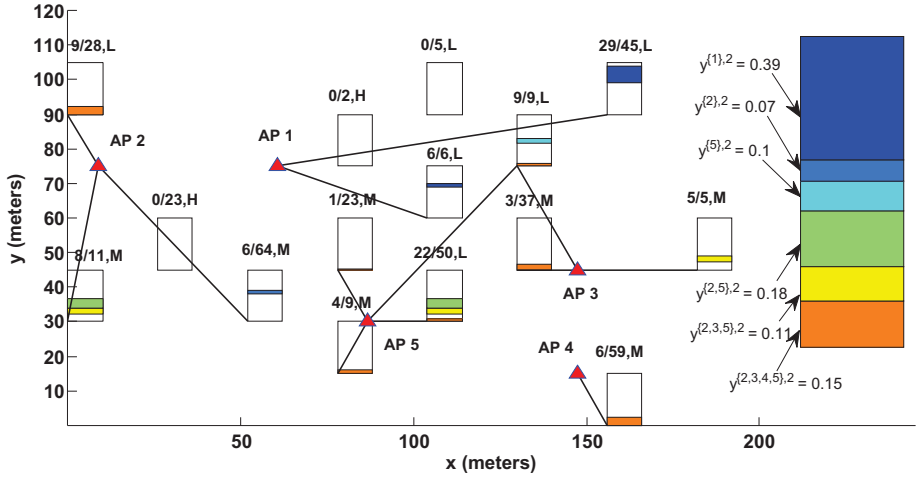
2.6.1. The Conservative Allocation Scheme

The conservative spectrum allocation and user AP-user association according to the solution of P2.1 under different traffic loads are shown in Fig. 2.4 and Fig. 2.5. Each smaller rectangle indicates the spectrum allocation at the corresponding user device group. The different colors represent different patterns. The amount of spectrum resources allocated to each user device group under each pattern is denoted by the size of the corresponding

[5]Note that the pathloss models for different RATs need not be the same.



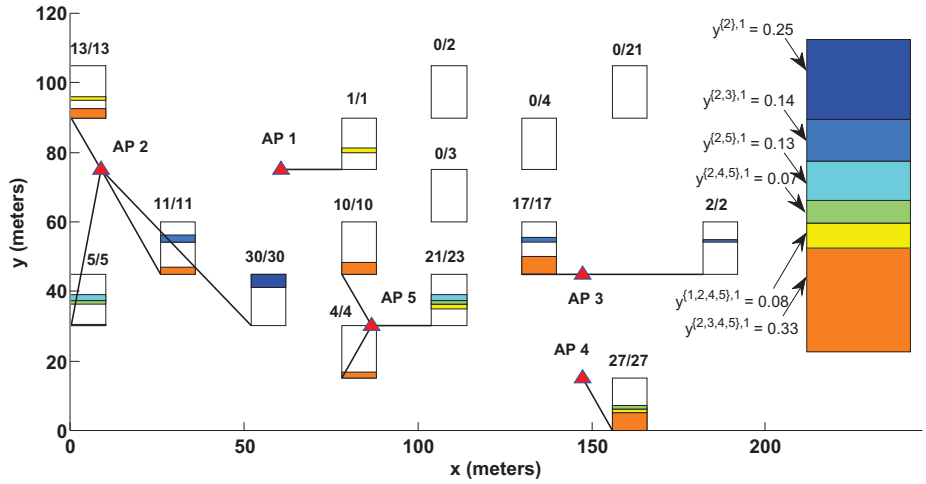
(a) Spectrum allocation pattern for LTE



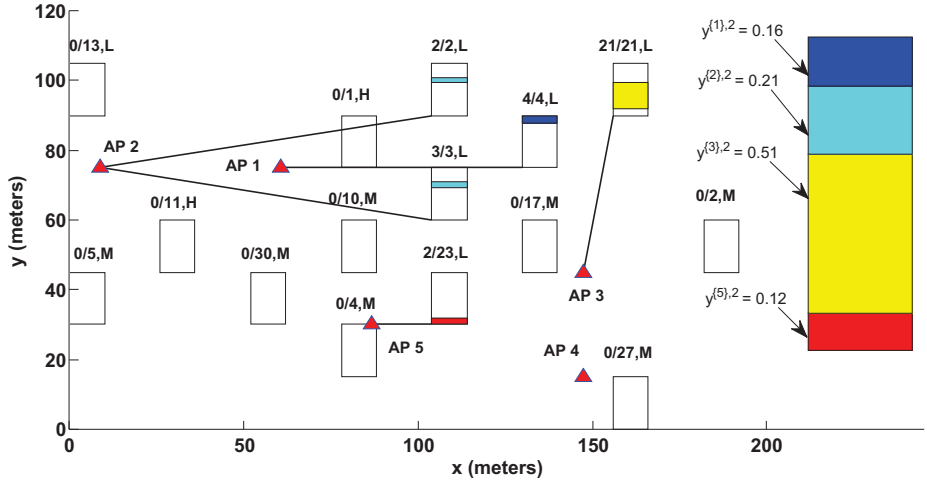
(b) Spectrum allocation pattern for LTE-U

Figure 2.4. Spectrum allocation patterns under heavy traffic, $\nu_H = 1, \nu_M = 0.01, \nu_L = 0.0025$.

bar inside the rectangle. The normalized average packet arrival rate relative to the total (sum) average arrival rate of each user device group as well as the intensity of nearby interference from the unlicensed band are shown above the rectangle. For example, for the user device group at the top left, 19/28 of its traffic load is served by LTE while



(a) Spectrum allocation pattern for LTE



(b) Spectrum allocation pattern for LTE-U

Figure 2.5. Spectrum allocation patterns under light/medium traffic, $\nu_H = 1, \nu_M = 0.01, \nu_L = 0.0025$.

the remaining 9/28 is served by LTE-U. There are three interference levels, referred to as "High", "Medium" and "Low", according to the intensity of the interference from the unlicensed band. For concreteness, we let the corresponding second moments of the vacation duration in 2.8 be $\nu_H = 1, \nu_M = 0.01, \nu_L = 0.0025$. When the second moment

$\nu_L = 0.0025$, the average vacation time is similar to the average packet service time, which is about 0.05 seconds/packet. Each line segment joining an AP and a user device group describes their association. The color bars on the right of each figure shows the actual spectrum partition into different patterns. Theorems 2.1 and Theorem 2.2 can be verified by counting the number of pieces in the partition and the number of rectangles connecting multiple triangles, respectively. For instance, in Fig. 2.4, the licensed band is partitioned into 7 segments and the unlicensed band is partitioned into 6 segments. Both numbers of partitions are smaller than the number of user device groups. Moreover, there is only one user device group served by multiple APs in the unlicensed band.

Fig. 2.4 shows the spectrum allocation patterns for the licensed and unlicensed bands in a high traffic scenario. The licensed band is partitioned into 7 segments and the unlicensed band is partitioned into 6 segments. The results show that the spectrum resources allocated to each user device group is roughly proportional to the corresponding traffic demand. Since the licensed band alone cannot support all the user device groups, some user device groups with low traffic demand are served by the unlicensed band only and most licensed spectrum is allocated to the user device groups with high traffic. In addition, even though some user device groups have high traffic demand (e.g., the top right one), more unlicensed spectrum is allocated to them because they do not have much nearby interference from other RATs in the unlicensed band. In contrast, some user device groups with low traffic demand (e.g., the one in the middle with arrival rate 2 packets/second) may also be allocated licensed spectrum if the nearby interference from the unlicensed band is severe.

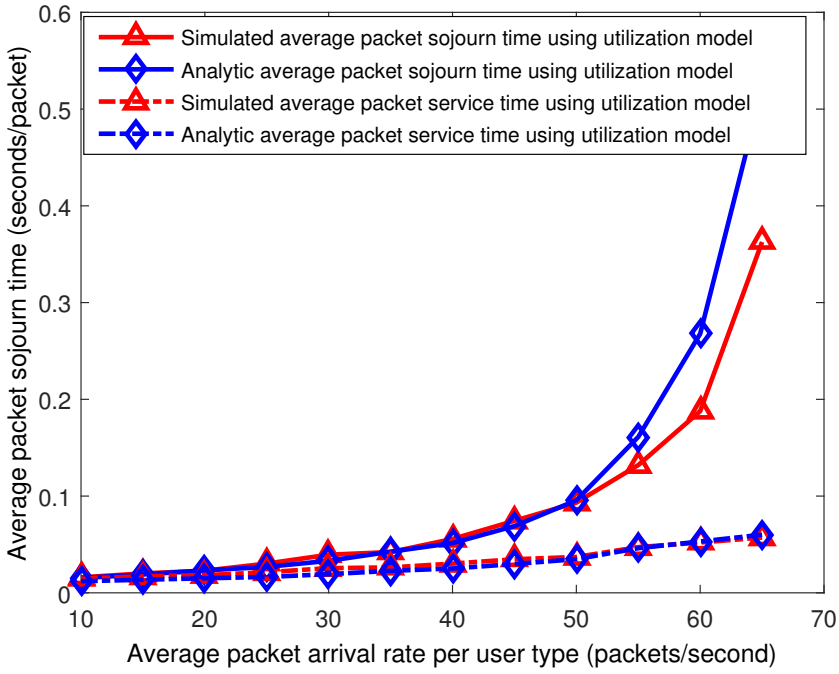


Figure 2.6. Analytic and simulated delay for the utilization model.

Fig. 2.5 shows the spectrum allocation patterns for the licensed and unlicensed bands in a medium/low traffic scenario. The licensed band is partitioned into 6 segments and the unlicensed band is partitioned into 4 segments. Since the licensed band alone is almost enough to support all the user device groups, most of them are served by the licensed band, and only a few of the user device groups are served by the unlicensed band because of the low traffic demand and lower interference generated within that band. In addition, most spectrum is allocated to the user device groups with high traffic.

2.6.2. The Utilization-Dependent Allocation Scheme

Clearly, the delay obtained by using the conservative allocation is an upper bound on the delay based on the utilization-dependent model. To validate the utilization-dependent

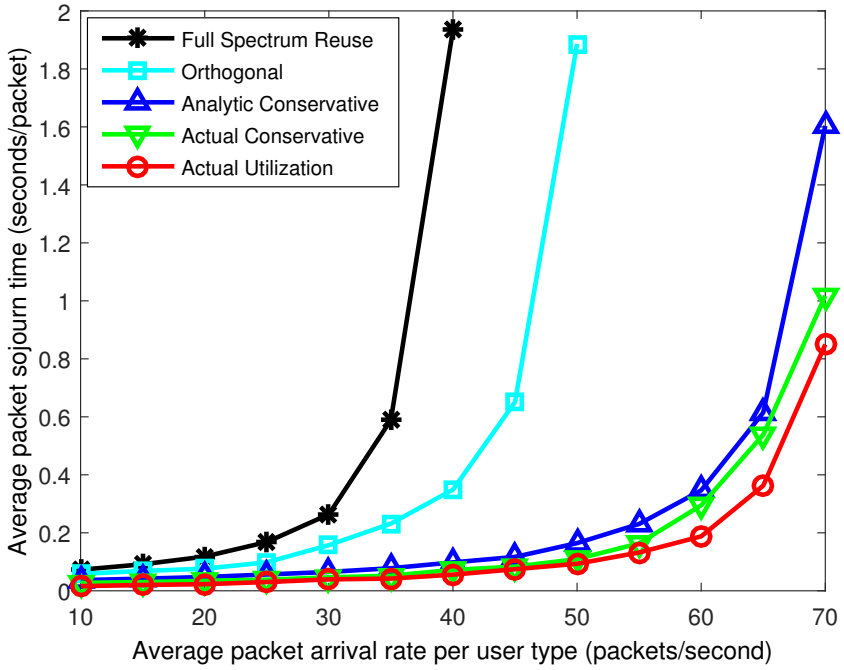
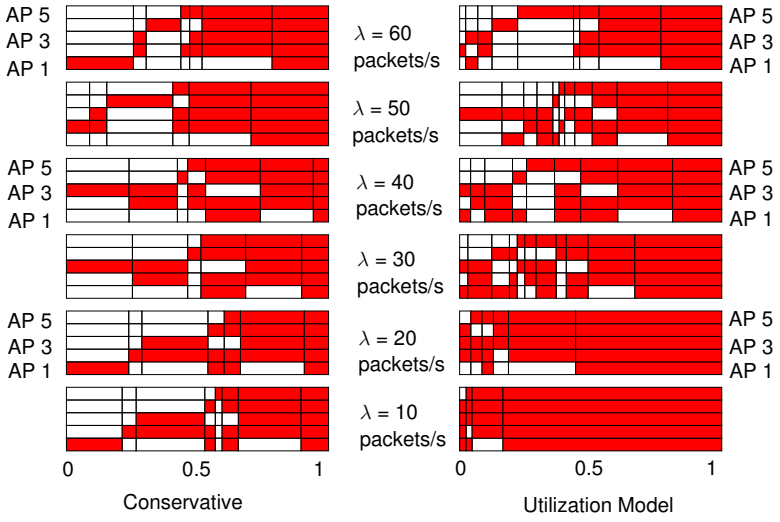


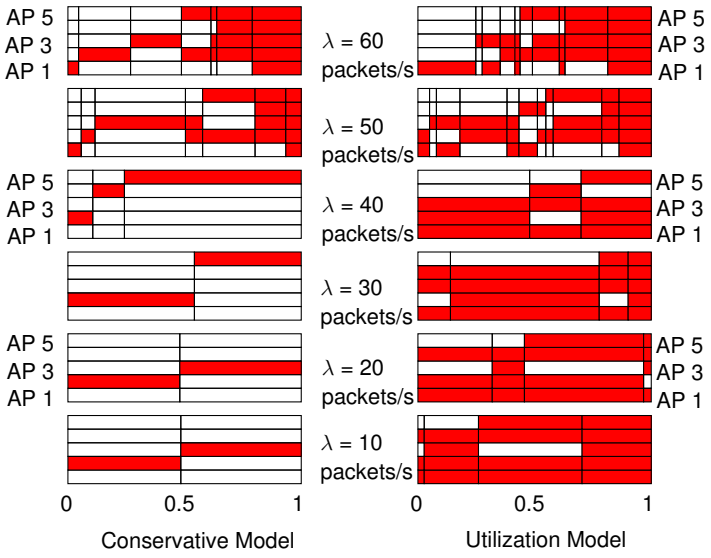
Figure 2.7. Comparison of the proposed schemes with benchmark schemes.

model, a packet-level simulator was used to compare the theoretical delay with the actual delay, shown in Fig. 2.6 The service (transmission) time of a packet is part of the packet sojourn time excluding the queueing delay. The horizontal axis is the average traffic load of each user device group. All curves in Fig. 2.6 are based on the same spectrum allocation. The utilization approximation is quite accurate with light traffic. With heavy traffic the approximation becomes coarser due to more interactions among APs. In addition, compared with delay, the service time increases much more slowly with traffic load, indicating that the spectrum is efficiently allocated to mitigate interference among APs.

Fig. 2.7 compares the optimization objective, average packet sojourn time, for the conservative and the utilization-dependent allocation schemes with orthogonal and full-reuse allocations. As the traffic increases to 40 and 50 packets/s, respectively, full-reuse



(a) Spectrum allocation patterns for LTE.



(b) Spectrum allocation patterns for LTE-U.

Figure 2.8. Spectrum allocation patterns for different loads.

and orthogonal allocations fail to support all the user device groups, suggesting these are the maximum throughputs under these two allocation schemes, respectively. Hence, the proposed conservative and utilization-dependent allocation schemes achieve significantly larger throughput regions than the other schemes. In addition, both conservative and utilization-dependent allocation schemes yield significant gain especially in the high traffic scenario. The reason is that the proposed spectrum allocation schemes are traffic-aware, i.e., in the low traffic regime, it is better to reuse most of the spectrum, while in the high traffic regime, it is better to adaptively allocate the spectrum across APs. Furthermore, the proposed schemes exploit the particular characteristics of the RATs. That is, user device groups receiving less external interference from the unlicensed band are more likely to be allocated spectrum from that band. Furthermore, the actual delay based on the conservative allocation is also shown, which is between the theoretical delay using conservative allocation and the actual delay using the utilization-dependent allocation scheme. The utilization-dependent allocation scheme always outperforms the conservative allocation scheme since it accurately accounts for dynamic inter-cell interference. In light traffic, the allocation based on the utilization-dependent model reduces the average delay by about 40% compared to the conservative allocation, and this improvement becomes smaller as traffic loads increase.

Fig. 2.8 shows the optimal allocations for the licensed and unlicensed bands under different traffic loads using the conservative and utilization-dependent allocation schemes. Rectangles represent frequency bands and solid ones are used by the corresponding APs. A vertical stack of rectangles then represents a reuse pattern. The length of each rectangle

Table 2.3. Runtimes for solving P2.1 and P2.2

Arrival rate (packets/second)	10	20	30	40	50	60	70
Runtime of solving P2.1 (seconds)	3.47	3.44	6.01	7.09	8.19	5.98	6.22
Runtime of solving P2.2 (seconds)	14.63	11.37	18.61	25.62	29.42	35.23	32.75

is proportional to the fraction of the whole spectrum occupied by the corresponding pattern. In both Fig. 2.8(a) and Fig. 2.8(b), when traffic loads are low, the allocation scheme using the utilization-dependent model tends to be full reuse, and as traffic loads increase, the scheme tends to orthogonalize the spectrum. In comparison with the conservative allocation, which remains orthogonal under different traffic loads, the utilization-dependent allocation uses the spectrum more efficiently since it approximates interactions among APs more accurately.

2.6.3. Runtime considerations

The optimization problems P2.1 and P2.2 were solved in Matlab using CVX [**67**] and the default convex program solver on an Intel Core i5 2.6 GHz quad-core computer with 8 GB RAM. Runtimes for solving P2.1 and P2.2 are shown in Table 2.3. In light traffic the runtimes are relatively short, indicating few iterations (alternate optimizations). The short runtimes indicate that the proposed schemes can be applied on the slow timescale considered for small to medium-size networks.

Although measuring and exchanging the side information needed for the optimizations requires only a modest amount of overhead, especially for a network cluster with about 20 APs, the number of variables in P2.1 and P2.2 increase exponentially with n. Hence the proposed schemes do not scale to large networks with hundreds of APs. Scalable

algorithms for solving similar optimization problems (with a single RAT) are presented in a recent paper [**32**]. The approach presented there, which can be applied to large networks with hundreds of APs, can also be applied to the type of multi-RAT networks considered here.

2.7. Summary

We have studied spectrum allocation in downlink HetNets with multiple RATs over different bands using the average packet sojourn time as the performance metric. In addition to the licensed band, a queueing model with vacation has been proposed to model the additional delay associated with the unlicensed band. Two optimization-based schemes have been proposed and shown to be highly effective. A sparse allocation (in terms of both the number of spectrum segments and user association) achieves the optimal network utility. Simulation results show that the proposed allocation schemes yield user associations and spectrum allocations, which utilize the spectrum of each RAT more efficiently compared with the benchmark allocation schemes, namely, orthogonal frequency reuse and full-frequency-reuse.

CHAPTER 3

Spectrum Allocation and User Association for Large-Scale Networks

This work studies centralized slow-timescale spectrum management in metropolitan area networks with a very large number of APs and user devices. The joint spectrum allocation and user association problem is first formulated as a convex optimization problem with an exponential number of variables in the network size. A scalable reformulation is obtained by exploiting the geometric graph nature of the network and provable sparsity of the optimal solution. A pattern pursuit algorithm with low complexity is proposed to solve the reformulated problem with guaranteed gap to the global optimum. Efficient algorithms are developed to obtain near-optimal allocations for a network with up to 1,000 APs and 2,500 active users. Numerical results show that the proposed solution achieves significant gains in terms of delay and throughput over existing schemes and is within 7% to the global optimum in a typical scenario.

3.1. Introduction

With increasing number of smart terminals and widening use of mobile Internet applications, we are witnessing an explosion of mobile traffic in commercial networks. Dense deployment of APs or small cells in addition to macro cells over a large area has been considered as a promising candidate for future 5G networks. The flexible multi-tier architecture can better match highly dynamic traffic demands of user devices to possible

serving APs. Due to irregularities of network topology and sophisticated interference conditions, efficient joint spectrum allocation and user association becomes extremely crucial for harnessing the full power of the infrastructure.

There have been many studies of resource management in cellular networks [**12**, **13**, **16**, **20**, **23**, **24**, **26**, **27**, **33**]. In [**13**], a dynamic fractional frequency reuse scheme was proposed to combat the inter-sector interference. In [**16**], a heuristic greedy search was proposed for user association. In [**23**, **24**], a utility maximization framework and pricing-based association methods were proposed. The association problem was jointly considered with resource allocation in [**12**, **20**]. In general, most of these work gives sub-optimal solutions either by solving a non-convex optimization problem or by running a distributed algorithm, which is far from optimal.

In [**34**, **35**], Zhuang et al developed a centralized optimization-based framework for allocating downlink spectrum resources on a relatively slow timescale. The spectrum allocation and user association problem was formulated as a convex optimization problem where the global optimal solution can be obtained using a standard solver. However, the number of variables in the problem grows exponentially with the number of APs. The space and time complexities of solving the problem for a large network of hundreds of APs become prohibitive.

To address the preceding challenges, we derive an equivalent reformulation of the fundamental resource allocation and user association problem from the viewpoints of user devices. Such user-centric reformulation captures the fact that each user device's performance depends only on the interference pattern of no more than a constant number of APs in the user device's neighborhood. This allows a low-complexity reformulation

of the global problem, which reduces the total number of variables from exponential to quadratic in the number of user devices. In related work [**36**, **37**], Zhuang et al proposed a scalable solution using convex relaxation and a heuristic coloring algorithm to compute a global spectrum allocation. In contrast, the treatment here is simpler, and we provide a more efficient algorithm with guaranteed performance. Specifically, we design a pattern pursuit algorithm and prove that it can yield a solution within any given gap from the global optimum. The framework here applies to all concave utility functions.

The underlying practical problem we wish to address is how to allocate resources in a metropolitan area network consisting of a very large number of APs and user devices. The total overhead for the network controller to perform the proposed resource allocation scheme is easily manageable if the timescale of resource adaptation is considered to be once every a few seconds or minutes. For example, the rate for sending 30,000 parameters (16 bits each) every minute is only 8 kilobits per second (kbps). To validate the performance of the proposed scheme, packet-level simulations are carried out. We demonstrate the proposed solution for networks with up to 1,000 APs and 2,500 user devices. It is observed that the proposed scheme significantly outperforms other conventional schemes. The performance is within 7% gap from (an upper bound of) the globally optimal utility in a typical scenario.

The remainder of this chapter is organized as follows. The system model is introduced in Section 3.2. The optimization problem is formulated in Section 3.3. A tractable solution is given in Section 3.4. Simulation results are presented in Section 3.5 and concluding remarks are given in Section 3.6. Most technical proofs are relegated to the appendices.

3.2. System Model

We consider the downlink of a network consisting of n APs and k user devices. A network controller is informed of the intensity of independent homogeneous Poisson traffic intended for every user device. It also receives sufficiently accurate reports of channel/interference information from all the APs. The resource allocation is performed on a slow timescale, e.g., once every a few seconds or minutes, which makes information exchange (channel state information feedback) and joint resource allocation viable at the central controller. In addition, since the period of slow-timescale resource allocation is much longer than the channel coherence time, the average channel conditions can be accurately modeled and measured using path loss and the statistics of small scale fading. The frequency resources are assumed to be homogeneous on a slow timescale. Given spectrum resource of bandwidth W Hz, the task of the central controller is to determine which spectrum segment(s) to allocate to each AP-user link in order to maximize the long-term network utility.

Denote the set of AP indexes by $N = \{1, \ldots, n\}$ and the set of user device indexes by $K = \{1, \ldots, k\}$. We allow arbitrary association so that each AP can simultaneously serve any subset of user devices and each user device can be simultaneously served by any subset of APs. Furthermore, we allow flexible resource allocation in that each AP-user link can use an arbitrary (possibly discontinuous) parts of the spectrum.

The key to total spectrum agility is the notion of *pattern* [**34**, **35**]. In general, a pattern simply refers to a subset of transmitters. A resource is said to be reserved for pattern A if the resource is to be shared by transmitters in A. We restrict our attention to frequency resources in this chapter, but this can be generalized to time, frequency,

and other resources (e.g., spatial resources). In the downlink, a pattern A is a subset of N, and APs in A are allowed simultaneous access to the frequency bands reserved for pattern A. The pattern uniquely determines the interference condition and henceforth also the efficiency of every AP-user link under the pattern. There are 2^n distinct patterns in total, including the empty one. Because the spectrum is regarded as homogeneous on the timescale of interest, the spectrum allocation problem can be formulated as how to divide the spectrum among all 2^n patterns and how to allocate these patterns to the links. We illustrate the concept of pattern in Fig. 2.1 using a small network with three APs and two user devices. The spectrum is first divided into $2^3 - 1 = 7$ segments, where one segment is used by AP 1 exclusively (the pattern is $\{1\}$), a second is used by AP 2 exclusively (the pattern is $\{2\}$), a third is used by AP 3 exclusively (the pattern is $\{3\}$), and the remaining four segments include three shared by the pairs of APs (the patterns are $\{1, 2\}$, $\{2, 3\}$, and $\{1, 3\}$, respectively), as well as one segment shared by all three APs (the pattern is $\{1, 2, 3\}$). Each AP then divides each pattern it is allocated into two piece to serve the two user devices.

The notion of pattern is related to the concept of *independent set* defined in the special case where the network is described by a weighted/unweighted conflict graph. In a conflict graph, since adjacent links cannot succeed simultaneously, it suffices to schedule only patterns corresponding to independent sets. The classical problem is to find independent sets of links that maximize the network utility. In this work, nearby links cause "soft" interference rather than a "hard" conflict. The solution space consists of all 2^n patterns (as shown shortly, the optimal solution consists of a very small subset of patterns).

The allocation problem can be divided into the following two subproblems:

(1) Allocate bandwidths to all 2^n patterns, denoted by a 2^n-dimensional vector: $y = (y^A)_{A \subset N}$, where $y^A \in [0, 1]$ is the fraction of bandwidth shared by APs in A. Clearly,

$$\sum_{A \subset N} y^A = 1. \tag{3.1}$$

An efficient allocation allocates no resource to the empty pattern, yielding $y^{\varnothing} = 0$.

(2) For every pattern $A \subset N$, every AP in A divides the spectrum reserved for A to serve all its associated user devices using orthogonal spectrum segments. Denote the bandwidth allocated to the link $i \to j$ (the link from AP i to user device j) over pattern A as $w^A_{i \to j}$. Consequently,

$$\sum_{j \in K} w^A_{i \to j} \leq y^A, \quad \forall A \subset N, i \in A. \tag{3.2}$$

As $w^A_{i \to j}$ is only defined for $i \in A$, we have exactly $kn2^{n-1}$ such variables.

Although the y variables specify the pattern bandwidths only, they directly imply a physical allocation as illustrated in Fig. 2.1. Finer allocation to AP-user links is then straightforward. As illustrated in Fig. 2.1, a physical spectrum allocation can be easily assembled from the set of w variables satisfying (3.1) and (3.2). Also, user device j is associated to AP i if and only if $w^A_{i \to j} > 0$ for some pattern A with $i \in A$.

For simplicity, it is assumed that each AP applies a flat power spectral density (PSD) over the allocated spectrum. The spectral efficiency of link $i \to j$ over pattern A is denoted by $s^A_{i \to j}$. It suffices to define $s^A_{i \to j}$ only for $i \in A$ as we shall not use $s^A_{i \to j}$ with

$i \notin A$ in problem formulations. To preempt any concern, we let

$$s^A_{i\to j} = 0, \quad \forall i \in N \setminus A. \tag{3.3}$$

Usually, the exclusive spectrum has higher spectral efficiency than shared spectrum. In general, $s^A_{i\to j} \geq s^B_{i\to j}$ if $i \in A \subset B$. The spectral efficiency $s^A_{i\to j}$ can either be calculated based on pathloss and other impairments or be measured over time. For concreteness in obtaining numerical results, Shannon's formula is used for the link efficiencies:

$$s^A_{i\to j} = \frac{W}{L}\log_2\left(1 + \frac{p_i g_{i\to j}}{n_0 + \sum\limits_{l\in A: l\neq i} p_l g_{l\to j}}\right) \quad \text{packets/second} \tag{3.4}$$

if $i \in A$ and $s^A_{i\to j} = 0$ if $i \notin A$, where L is the average packet length in bits, p_i is the transmit PSD of AP i, n_0 is the noise PSD, $g_{i\to j}$ is the gain of link $i \to j$ which captures the effects of path loss and shadowing, and $\sum_{l\in A: l\neq i} p_l g_{l\to j}$ is the interference received from other APs operating over the same pattern A. Here the link efficiency is normalized using factor W/L so that the units of $s^A_{i\to j}$ are packets/second.

The service rate to user device j contributed by AP $i \in A$ over pattern A is $s^A_{i\to j} w^A_{i\to j}$. The total service rate of user device j denoted as r_j can be calculated by summing over all APs over all patterns:

$$r_j = \sum_{A\subset N}\sum_{i\in A} s^A_{i\to j} w^A_{i\to j}. \tag{3.5}$$

3.3. Basic Problem Formulation

The fundamental problem is to maximize the long-term utility by adapting the user association and multi-pattern resource allocation. Collecting the constraints (3.1), (3.2), and (3.5), we formulate P3.1 as:

$$\underset{\boldsymbol{r},\boldsymbol{w},\boldsymbol{y}}{\text{maximize}}\ u(\boldsymbol{r}) \tag{P3.1a}$$

$$\text{subject to } r_j = \sum_{A\subset N}\sum_{i\in A} s_{i\to j}^{A} w_{i\to j}^{A}, \qquad \forall j \in K \tag{P3.1b}$$

$$\sum_{j\in K} w_{i\to j}^{A} \le y^{A}, \qquad \forall A \subset N, \forall i \in A \tag{P3.1c}$$

$$\sum_{A\subset N} y^{A} = 1, \tag{P3.1d}$$

$$w_{i\to j}^{A} \ge 0, \quad \forall j \in K, \forall A \subset N, \forall i \in A \tag{P3.1e}$$

where $u(\boldsymbol{r})$ is the network utility function, and $\boldsymbol{y} = (y^A)_{A\subset N}$ and $\boldsymbol{w} = (w_{i\to j}^{A})_{j\in K, A\subset N, i\in A}$ represent the bandwidth allocations. The spectral efficiencies $(s_{i\to j}^{A})_{j\in K, A\subset N, i\in A}$ are known parameters. Because the rate vector $\boldsymbol{r} = [r_1, \ldots, r_k]$ is a linear transformation of the allocation vector $\boldsymbol{w}$ through (P3.1b), the utility can be expressed directly as a function of the allocations: $u(\boldsymbol{r}(\boldsymbol{w}))$.

P3.1 is a convex optimization problem as long as $u(\boldsymbol{r})$ is concave in $\boldsymbol{r}$. The sum rate, the minimum user service rate (max-min fairness), and the sum log-rate (proportional fairness) are all concave utility functions. In this chapter, we focus on the average

(negative) packet delay as the network utility function:

$$u(\boldsymbol{r}) = -\sum_{j \in K} \frac{\lambda_j}{(r_j - \lambda_j)^+}, \tag{3.6}$$

where λ_j is the homogeneous Poisson packet arrival rate of user device j, and the extended real-valued function $1/x^+ = 1/x$ if $x > 0$ and $1/x^+ = +\infty$ if $x \leq 0$. It is easy to see that $1/x^+$ is convex on $(-\infty, +\infty)$. The choice of this utility function also assumes exponential packet length and a "conservative rate" as in [**34**]. If $r_j \leq \lambda_j$, the packet delay is infinite, i.e., the system becomes unstable.

Theorem 3.1 ([**35**]). *There exists an optimal solution to P3.1 with at most k active patterns, i.e., the optimal solution satisfies:*

$$\left|\left\{A \subset N \;\middle|\; y^A > 0\right\}\right| \leq k. \tag{3.7}$$

In addition, if the coefficients $s^A_{i\to j}$ are drawn from a jointly continuous distribution, then, with probability 1, there are at most $n-1$ user devices served by multiple APs in every optimal solution to P3.1. That is, the optimal solution satisfies

$$\left|\left\{j \in K \;\middle|\; \text{there exist } A_1, A_2 \subset N, i_1 \in A_1, i_2 \in A_2 \text{ s.t. } i_1 \neq i_2 \text{ and } w^{A_1}_{i_1\to j}, w^{A_2}_{i_2\to j} > 0\right\}\right| \leq n-1. \tag{3.8}$$

Theorem 3.1 was proved in [**35**]. It guarantees that although the total number of patterns grows exponentially with the number of APs in the network, using a small number of patterns achieves the optimal performance. Furthermore, it states that although we

allow each user device to be served by all APs, most user devices will be associated with only one AP in the optimal solution.

Proposition 3.1. *If the utility function $u(\boldsymbol{r}(\boldsymbol{w}))$ is affine in $\boldsymbol{w}$, then the maximum utility in P3.1 can be attained by a single active pattern, where each AP serves only one user device.*

A simple example for an affine utility function $u(\boldsymbol{r}(\boldsymbol{w}))$ is the weighted sum rate function. Proposition 3.1 admits a simple intuition: When the utility is equal to a weighted sum of the bandwidths allocated to all links over all patterns, shifting all resources to a dominant pattern does not reduce the utility. We prove Proposition 3.1 in Appendix A.5.

3.4. A Scalable Model and Algorithm

There are $kn2^{n-1}+2^n+k$ variables in P3.1. P3.1 can be solved using a standard convex optimization solver for networks with a small number of APs. For a metropolitan area network consisting of hundreds or even thousands of APs, the space and time complexities of P3.1 become prohibitive. By first dividing the network into many small clusters, one may solve for allocation in each cluster separately by assuming away the uncertainties about inter-cluster interference. However, because interference from outside a cluster can penetrate deeply into a cluster, such divide-and-conquer solutions suffer significant loss. We also note that any distributed solution is necessarily myopic and hence suffers similar loss. In this section, we treat the network in its entirety and develop a scalable, equivalent reformulation. We then provide an efficient near-optimal method for solving the new optimization problem.

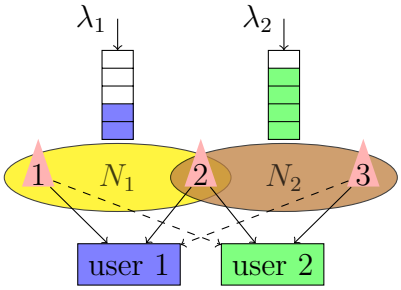


Figure 3.1. Neighborhoods in a network of 3 APs and 2 user devices.

3.4.1. Local Patterns and Allocations

In a large network with many APs, a user device can in general only be served by a small subset of nearby APs due to path loss. For every $j \in K$, let N_j denote the set of APs whose received signal-to-noise ratios at user device j are above a certain threshold ξ, i.e.,

$$N_j \triangleq \left\{ i \in N \;\middle|\; \frac{p_i g_{i \to j}}{n_0} > \xi \right\}. \tag{3.9}$$

We define N_j as the *neighborhood* of user device j. user device j treats all APs outside N_j as stationary noise sources. This can be arbitrarily precise as N_j may include all APs except those received by user device j at well below the noise level. It is fair to assume the size of all neighborhoods are upper bounded by a constant c_0, i.e., $|N_j| \leq c_0$, $\forall j \in K$. Fig. 3.1 depicts a toy network example with 3 APs and 2 user devices. Here the neighborhood of user device 1 is $N_1 = \{1, 2\}$ since AP 3 is far away from user device 1, making the received power from AP 3 below ξ. Similarly, the neighborhood of user device 2 is $N_2 = \{2, 3\}$. All APs in neighborhood N_i collectively can be thought of as a server of user device i's traffic.

We next redefine the spectral efficiencies $s^A_{i\to j}$ to facilitate problem formulation using only local patterns and variables. Since a user device can only be served by APs in its neighborhood, we set

$$s^A_{i\to j} = 0, \quad \forall j \in K, \forall i \in N, \forall A \subset N : i \notin A \cap N_j \tag{3.10}$$

without loss of optimality. Moreover, since all APs outside a user device's neighborhood are treated as noise sources, we set[1]

$$s^A_{i\to j} = s^{A\cup(N\setminus N_j)}_{i\to j}, \quad \forall j \in K, \forall i \in N, \forall A \subset N : i \in A \cap N_j. \tag{3.11}$$

A close examination of (3.10) and (3.11) reveals that the redefined spectral efficiency of link $i \to j$ depends only on the activities of APs in the neighborhood of user device j. In particular,

$$s^A_{i\to j} = s^{A\cap N_j}_{i\to j}, \qquad \forall j \in K, \forall i \in N, \forall A \subset N. \tag{3.12}$$

In the remainder of the chapter we assume (3.12) holds.

For every $j \in K$, all subsets of N_j constitute the set of *local patterns* of user device j. We adopt a new set of allocation variables $(x^B_{i\to j})$, where for every $j \in K$, $x^B_{i\to j}$ is only defined for $B \subset N_j$ and $i \in B$. Here, $x^B_{i\to j}$ denotes the bandwidth allocated to link $i \to j$ under the local pattern B, which can be obtained as:

$$x^B_{i\to j} = \sum_{A\subset N: A\cap N_j = B} w^A_{i\to j}, \qquad \forall j \in K, \forall B \subset N_j, \forall i \in B. \tag{3.13}$$

[1]If $N \setminus N_j \subset A$, the two sides of (3.11) refer to the same variable; otherwise, (3.11) resets the spectral efficiency on the left hand side to that of the right hand side (which is no higher).

That is, it is the sum bandwidth over all global patterns that match B in the neighborhood of user device j. The number of x variables is

$$\sum_{j\in K} |N_j| 2^{|N_j|-1} \leq k c_0 2^{c_0-1}. \tag{3.14}$$

From the viewpoint of user device j, $(x^B_{i\to j})_{B\subset N_j, i\in B}$ describes the bandwidths allocated to all its associated links over all its local patterns. Using (3.12) and (3.13), the summation in (P3.1b) can be written as:

$$r_j = \sum_{A\subset N} \sum_{i\in A} s^A_{i\to j} w^A_{i\to j} \tag{3.15}$$

$$= \sum_{A\subset N} \sum_{i\in A\cap N_j} s^{A\cap N_j}_{i\to j} w^A_{i\to j} \tag{3.16}$$

$$= \sum_{B\subset N_j} \sum_{i\in B} s^B_{i\to j} \sum_{A\subset N: A\cap N_j = B} w^A_{i\to j} \tag{3.17}$$

$$= \sum_{B\subset N_j} \sum_{i\in B} s^B_{i\to j} x^B_{i\to j}. \tag{3.18}$$

In (3.18), only local spectrum allocations $x^B_{i\to j}$ with $i \in B \subset N_j$ are used. Therefore, as substitutes of $kn2^{n-1}$ (global) $\boldsymbol{w}$ variables, at most $c_0 2^{c_0-1} = O(1)$ (local) $\boldsymbol{x}$ variables are involved in (3.18) for given j. This is sufficient as the sum over $B \subset N_j$ exhausts all patterns of APs that may serve user device j.

3.4.2. A Highly Scalable Reformulation

Recall from Theorem 3.1 that there exists an optimal solution to P3.1 that activates at most k patterns. Therefore, the local allocation variables $\boldsymbol{x}$ should fit into k segments, where each segment represents the allocation of one pattern. Denote the set of all segment

indexes by $L = \{1, \cdots, k\}$. By introducing replicas of the $\boldsymbol{x}$ variables in the form of $(x_{i\to j}^{A,l})_{j\in K, l\in L, i\in A\subset N_j}$, we have obtained an equivalent reformulation of P3.1, referred to as P3.2:

$$\underset{\boldsymbol{r},\boldsymbol{x},\boldsymbol{d},\boldsymbol{h}}{\text{maximize}}\ u(\boldsymbol{r}) \tag{P3.2a}$$

$$\text{subject to } r_j = \sum_{A\subset N_j}\sum_{i\in A} s_{i\to j}^{A} \sum_{l\in L} x_{i\to j}^{A,l}, \quad \forall j \in K \tag{P3.2b}$$

$$x_{i\to j}^{A,l} \le d_j^{A,l}, \ \ \forall j \in K, \forall l \in L, \forall A \subset N_j, \forall i \in A \tag{P3.2c}$$

$$d_j^{A,l} + \sum_{\substack{B\subset N_m:\\ B\cap N_j \neq A\cap N_m}} d_m^{B,l} \le 1, \quad \forall l \in L, \forall j \in K,$$

$$\forall A \subset N_j, \forall m \in K : N_m \cap N_j \neq \emptyset, \tag{P3.2d}$$

$$\sum_{j\in K: i\in N_j} \sum_{A\subset N_j: i\in A} x_{i\to j}^{A,l} \le h^l, \quad \forall i \in N, \forall l \in L \tag{P3.2e}$$

$$\sum_{l\in L} h^l \le 1, \tag{P3.2f}$$

$$d_j^{A,l} \in \{0,1\}, \forall l \in L, \ \forall j \in K, \forall A \subset N_j \tag{P3.2g}$$

$$x_{i\to j}^{A,l} \ge 0, \quad \forall l \in L, \forall j \in K, \forall A \subset N_j, \forall i \in A. \tag{P3.2h}$$

In P3.2, the spectrum is divided to k segments with bandwidths $h^1, \ldots, h^k$, each corresponding to a global pattern. For the l-th segment, the variables $(x_{i\to j}^{A,l})_{j\in K, i\in A\subset N_j}$ and $(d_j^{A,l})_{j\in K, A\subset N_j}$ represent the allocation of this segment from all user devices' viewpoints. Here (P3.2b) corresponds to (3.18). (P3.2c) implies that $d_j^{A,l}$ is the indicator of local pattern A from user device j's viewpoint over the l-th segment, i.e., $d_j^{A,l} = 1$ if there exists $i \in N_j$ such that $x_{i\to j}^{A,l} > 0$; otherwise $d_j^{A,l} = 0$. (P3.2d) constrains the consistency

of allocation over each segment among all user devices. That is, (P3.2d) enforces the allocation of no more than one pattern over segment l from every user device's viewpoint. Compared with P0 which has $O(kn2^n)$ variables, the number of variables in P3.2 is

$$k\sum_{j\in K}(|N_j|+2)2^{|N_j|-1}+2k=O(k^2). \tag{3.19}$$

Theorem 3.2. *P3.1 and P3.2 are equivalent in the sense that they achieve the same utility with identical rate vector(s). Moreover, given the optimal solution to P3.2, the patterns and bandwidths of the optimal solution to P3.1 can be obtained as:*

$$A_l=\bigcup_{j\in K}\bigcup_{B\subset N_j:d_j^{B,l}>0}B, \qquad \forall l\in L \tag{3.20}$$

$$w_{i\to j}^{A_l}=x_{i\to j}^{A\cap N_j}, \qquad \forall l\in L,\forall j\in K,\forall i\in A_l. \tag{3.21}$$

Theorem 3.2 is proved in Appendix A.6. The following result is a useful building block for an efficient algorithm for solving P3.2 to arbitrary precision.

Proposition 3.2. *Suppose $u(\boldsymbol{r}(\boldsymbol{x}))$ is an affine function of $\boldsymbol{x}$. In terms of the maximum utility and the set of feasible $(\boldsymbol{x},\boldsymbol{d})$, P3.2 is equivalent to P3.3:*

$$\underset{\boldsymbol{x},\boldsymbol{d}}{\text{maximize}}\ u(\boldsymbol{r}(\boldsymbol{x})) \tag{P3.3a}$$

$$\text{subject to } x_{i\to j}^{A}\le d_j^A, \quad \forall j\in K,\forall A\subset N_j,\forall i\in A \tag{P3.3b}$$

$$d_j^A+\sum_{\substack{B\subset N_m:\\ B\cap N_j\neq A\cap N_m}}d_m^B\le 1, \tag{P3.3c}$$

$$\forall j\in K,\forall A\subset N_j,\forall m\in K:N_m\cap N_j\neq\emptyset \tag{P3.3d}$$

$$\sum_{j\in K: i\in N_j} \sum_{A\subset N_j} x_{i\to j}^{A} \leq 1, \qquad \forall i \in N \tag{P3.3e}$$

$$d_j^A \in \{0,1\}, \qquad \forall j \in K, \forall A \subset N_j \tag{P3.3f}$$

$$x_{i\to j}^{A} \in \{0,1\}, \qquad \forall j \in K, \forall A \subset N_j, \forall i \in A. \tag{P3.3g}$$

Proposition 3.2 follows the proof of Proposition 3.1 and Theorem 3.2. The key point is that when the utility function $u(\boldsymbol{r}(\boldsymbol{x}))$ is affine in $\boldsymbol{x}$, the optimal solution to P3.2 activates only one pattern and each AP serves one user device who benefits the most, which yields a simplified formulation P3.3 (there is no need for the replica index l). More importantly, P3.3 is a binary lineary program (BLP) with only $O(k)$ variables.

3.4.3. An Efficient Algorithm with Guarantee

Although the mixed integer programming P3.2 has significantly fewer variables than the original problem P3.1 for a large network, it is NP-hard in general. It is at least as hard to compute the performance gap between an approximation of P3.2 and the global optimal. The gap can, however, be upper bounded by optimizing an upper bound of the utility function. A promising technique is then to iteratively optimize local linear expansions of the concave utility function. In fact, because the expansion in each step must be an affine upper bound, each step becomes a linear program.

For ease of notation, denote Λ as the feasible region in terms of $\boldsymbol{x}$ defined by (P3.3b)-(P3.3g) and let $v(\boldsymbol{x}) = u(\boldsymbol{r}(\boldsymbol{x}))$ denote the concave utility function. We use $\nabla v(\boldsymbol{x})$ to

denote the "gradient" of $v(\cdot)$. Specifically, if $v(\cdot)$ is differentiable at $\boldsymbol{x}$,

$$[\nabla v(\boldsymbol{x})]^A_{i\to j} = \frac{\partial v(\boldsymbol{x})}{\partial x^A_{i\to j}}, \quad \forall j \in K, A \subset N_j, i \in A. \tag{3.22}$$

If $v(\cdot)$ is not differentiable at $\boldsymbol{x}$, $\nabla v(\boldsymbol{x})$ is minus the subgradient of the convex function $-v(\cdot)$. For every $\boldsymbol{q}, \boldsymbol{x} \in \Lambda$, we denote

$$f_{\boldsymbol{q}}(\boldsymbol{x}) = v(\boldsymbol{q}) + \langle \nabla v(\boldsymbol{q}), \boldsymbol{x} - \boldsymbol{q} \rangle \tag{3.23}$$

where the inner product is defined in general as

$$\langle \boldsymbol{x}, \boldsymbol{z} \rangle = \sum_{j\in K} \sum_{A \subset N_j} \sum_{i \in A} x^A_{i\to j} z^A_{i\to j}. \tag{3.24}$$

Due to its concavity, $v(\cdot)$ must be upper bounded by its linear expansion:

$$v(\boldsymbol{x}) \leq f_{\boldsymbol{q}}(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \Lambda. \tag{3.25}$$

Given a fixed feasible point $\boldsymbol{q}$, we obtain an upper bound of the global maximum if we replace $u(\boldsymbol{r})$ in (P3.2a) by $f_{\boldsymbol{q}}(\boldsymbol{x})$. Since $f_{\boldsymbol{q}}(\cdot)$ is affine, Proposition 3.1 implies that a unique pattern can be identified to maximize the affine approximation. Based on this observation, we propose Algorithm 3.1, an iterative pattern-pursuit algorithm for finding a solution within any given $\epsilon > 0$ from the global optimum.

Algorithm 3.1 can be interpreted as the Frank-Wolfe type Algorithm (also known as the conditional gradient algorithm) [**73**]. The main difference than the conventional algorithm is that instead of doing line search, Algorithm 3.1 finds one "good" pattern in each iteration and re-optimize P3.2 using the set of "good" patterns identified so far.

Algorithm 3.1 Iterative algorithm for pattern pursuit

Input: $\epsilon > 0$.
Output: $\boldsymbol{x}^{(t)}$.
Initialization: $t \leftarrow 0$; $P \leftarrow \emptyset$; pick an arbitrary pattern $\boldsymbol{x}^{(0)}$.
repeat
Step 1. Compute $\boldsymbol{x}$ and $\boldsymbol{d}$ which maximize $\langle \nabla v(\boldsymbol{x}^{(t)}), \boldsymbol{x} \rangle$ subject to constraints (P3.3b)–(P3.3g).
Step 2. $t \leftarrow t + 1$. If $\boldsymbol{d} \notin P$, $P \leftarrow P \cup \{\boldsymbol{d}\}$; otherwise, add an arbitrary new pattern that is not in P.
Step 3. Solve P3.2 by restricting to patterns in P and obtain the optimal allocation solution $\boldsymbol{x}^{(t)}$.
until $\text{maximize}_{\boldsymbol{x} \in \Lambda} \langle \nabla v(\boldsymbol{x}^{(t)}), \boldsymbol{x} - \boldsymbol{x}^{(t)} \rangle < \epsilon$.

The "good" pattern set P grows after each iteration because either one "good" pattern is found or a random new pattern is added. In the worst case, Algorithm 3.1 takes no more than 2^n steps to terminate, because when the number of patterns in P reaches 2^n, $\boldsymbol{x}^{(t)}$ must be globally optimal, so that the condition to exit the loop must be met. Although Algorithm 3.1 does not guaranteed that an optimal solution can be obtained in polynomial time, it provides a trade-off between the running time and the optimality gap. The performance will be demonstrated in Sec. 3.5 through numerical examples.

Algorithm 3.1 has several important features:

(1) *"Good" pattern pursuit*: Algorithm 3.1 starts with the full-spectrum-reuse pattern in which all APs occupy the entire spectrum. In each iteration, it identifies one "best" pattern as the maximizer of this linear function (taken over the same domain). Due to Theorem 3.1, it usually takes no more than k iterations to find the global optimum.

(2) *Efficiency*: By Proposition 3.2, Step 1 solves a BLP in the form of P3.3. Although BLP is NP-complete in the worst case, many BLPs with sparse structure

can be solved efficiently [**39**]. As observed from numerical results, Step 1 takes a fairly small amount of time. In particular, if a branch and bound/cut method is used, the BLP step can be terminated as soon as a sufficiently tight upper bound is reached for the purpose of Algorithm 3.1.

(3) *Optimality guarantee*: Algorithm 3.1 has optimality guarantee as stated in Theorem 3.3.

Theorem 3.3. *Suppose* (3.12) *holds. For every* $\epsilon > 0$*, there exists a positive integer* k *such that* $v(\boldsymbol{x}^{(k)})$ *is at most* ϵ *away from the global optimum of P3.1.*

Proof. P3.2 is always re-optimized using more patterns than previous iterations, which results in non-decreasing series of the utility $(v(\boldsymbol{x}^{(t)}))_{t=0,1,\cdots}$. This series must converge due to boundedness of the utility function. Let $\boldsymbol{x}^*$ denote the global optimum. By (3.23) and (3.25),

$$v(\boldsymbol{x}^*) - v(\boldsymbol{x}^{(t)}) \leq \langle \nabla v(\boldsymbol{x}^{(t)}), \boldsymbol{x}^* - \boldsymbol{x}^{(t)} \rangle. \tag{3.26}$$

Therefore, when the condition for terminating the loop in Algorithm 3.1 is satisfied, the optimality gap $v(\boldsymbol{x}^*) - v(\boldsymbol{x}^{(k)})$ is guaranteed to be less than ϵ. □

3.5. Numerical Results

We use parameters compliant with the LTE standard [**66**] as given in Table 3.1. The maximum number of potential associations of a user is set as $c_0 = 3$. The results for the actual packet delay are obtained using a packet-level simulator, which adapts the transmission time[2] of each packet to the instantaneous active APs that are transmitting.

[2]The delay of a packet includes the transmission time and its waiting time in the queue.

Table 3.1. Parameter Configurations.

Parameters	Value/Function
AP transmit power	23 dBm
Total bandwidth	10 MHz
Average packet length	0.5 Mbits
AP to user device pathloss (LOS)	$30.18 + 26.7 \log_{10}(R)$
AP to user device pathloss (NLOS)	$34.53 + 36 \log_{10}(R)$
Lognormal shadowing standard deviation (LOS)	4 dB
Lognormal shadowing standard deviation (NLOS)	10 dB

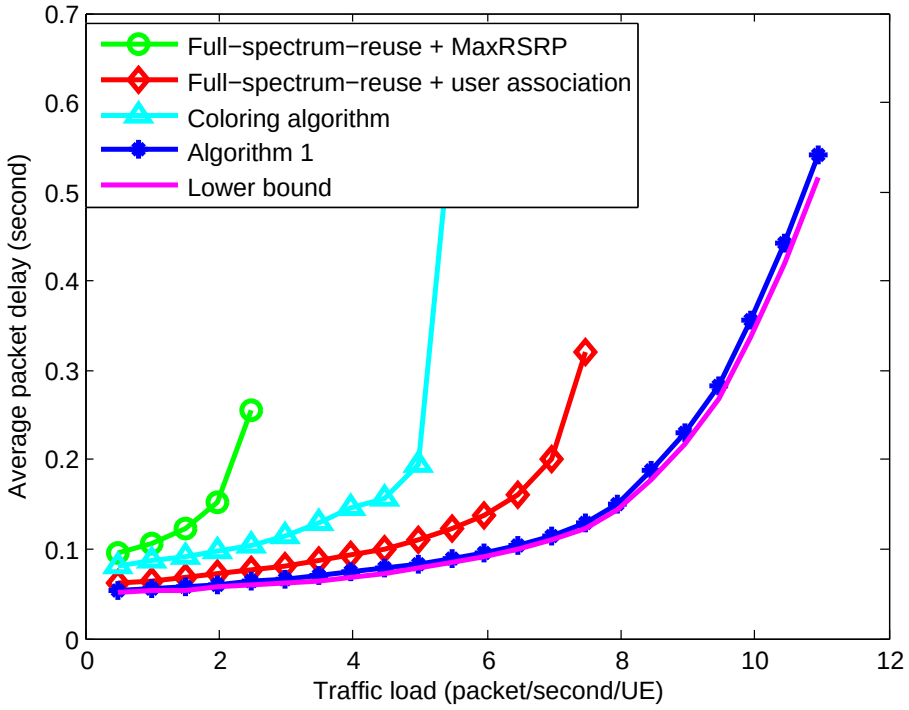


Figure 3.2. Comparison with the baseline schemes.

We investigate the performance gain of the proposed allocation schemes by comparing them with the following baseline schemes:

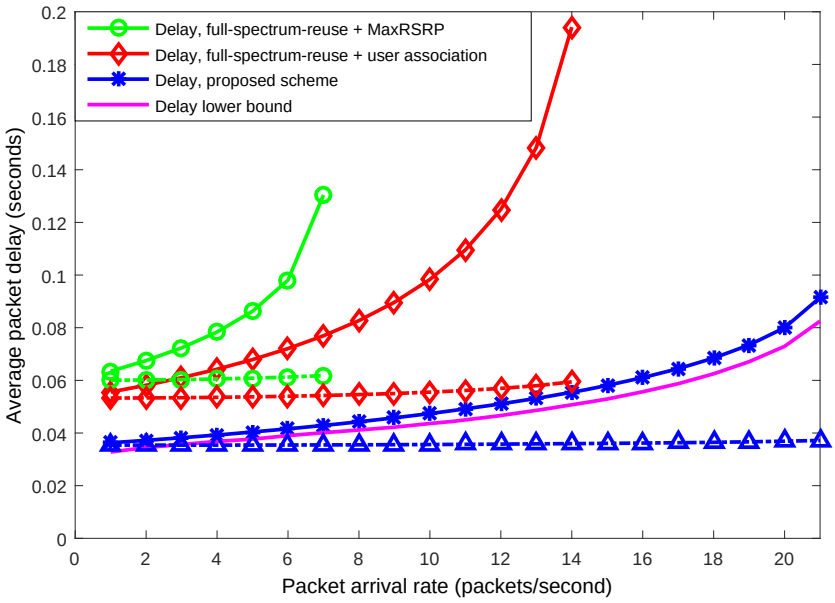


Figure 3.3. Comparison to the baseline schemes. Each dotted curve represents the average transmission time of the corresponding delay curve with identical marker and color.

(1) Full-spectrum-reuse + maximum reference signal receive power (MaxRSRP): Every AP reuses all available spectrum and every user device is associated to the strongest AP in terms of the received power.

(2) Full-spectrum-reuse + optimal user association: Every AP reuses all available spectrum and user association is optimized for the utility.

(3) The coloring algorithm proposed in [**36**].

(4) Optimal lower bound: The optimal lower bound of P3.1 that obtained through Algorithm 3.1.

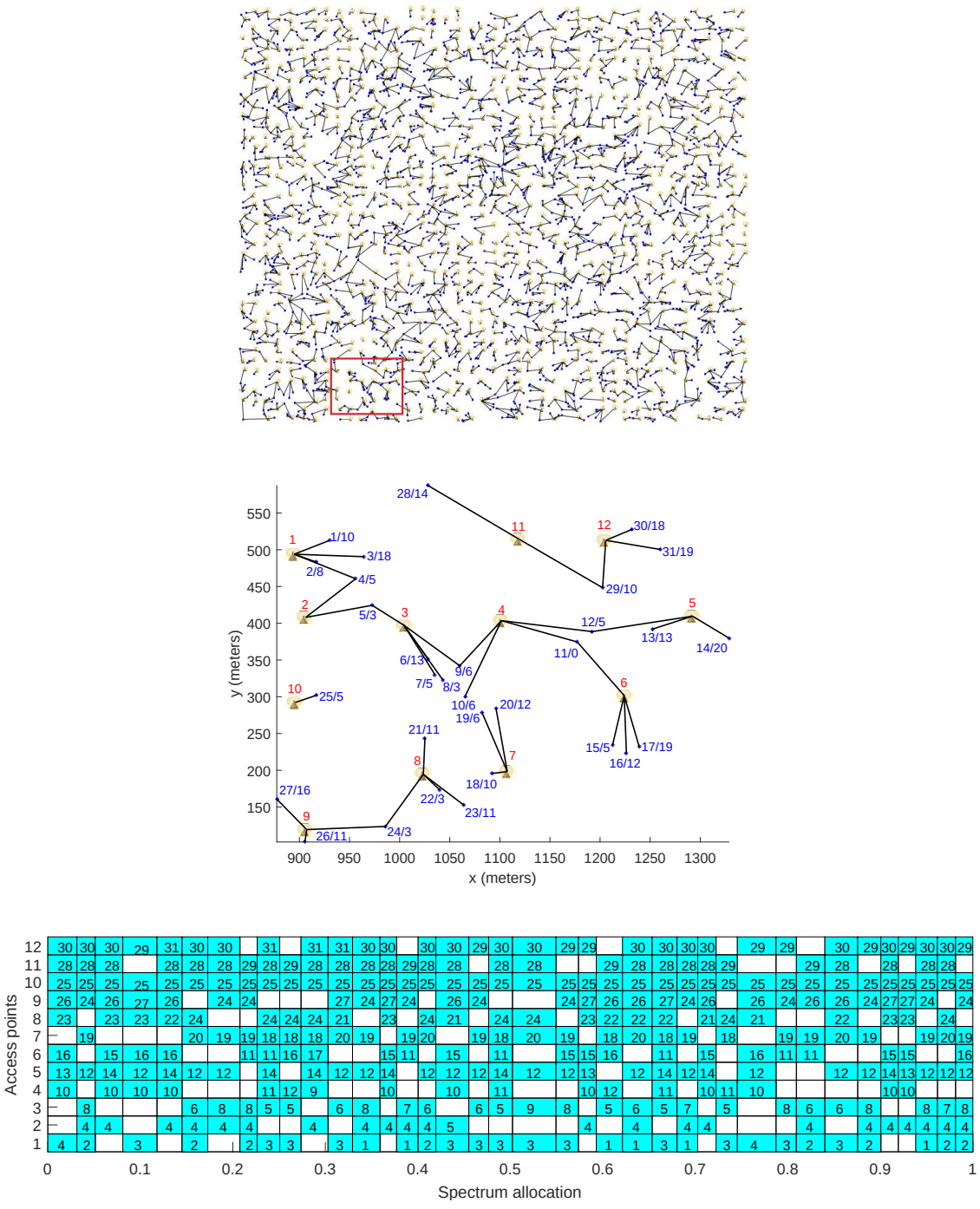


Figure 3.4. Spectrum allocation and user association in very large networks. (a) Deployment and user association for the large network. (b) Topology graph for the marked area in Fig. 3.4(a). (c) Allocation graph for the marked area in Fig. 3.4(a).

3.5.1. Performance in Medium Scale Networks

We first compare the performance of the proposed scheme in a network of medium size. We randomly drop 100 APs and 200 user devices over a 1,100 $\times$ 1,100 m^2 area. The average packet delay versus traffic intensity curves are shown in Fig. 3.2. As the average user device traffic increases to above 7.5 packets/second, all three baseline schemes fail to support all the user devices. While the proposed solution has significantly larger throughput (above 11 packets/second) than the other schemes. The proposed solution also significantly reduces the delay especially in the high traffic regime. The reason is that the proposed solution adapts to the traffic conditions, i.e., spectrum is reused more aggressively in the low traffic regime, whereas spectrum use is more often orthogonal to avoid mutual interference. Furthermore, the curve of the lower bound of the optimum is quite close to the curve of the proposed scheme. This means the proposed solution is close to the global optimum of P3.1.

3.5.2. Performance in Very Large Networks

In this section, the proposed allocation scheme is used to compute the near-optimal allocation for a network consisting of 1,000 APs and 2,500 user devices over a 4,200 $\times$ 4,200 m^2 area. Since the coloring algorithm can not afford the computation in such large scale network, we compare the proposed scheme with the first two baseline schemes.

The average packet delay versus traffic intensity curves are shown in Fig. 3.3. The proposed solution has significantly larger throughput (above 21 packets/second) than full-spectrum reuse with maxRSRP association (7 packets/second) and full-spectrum reuse with optimal user association allocation (14 packets/second). The proposed solution also

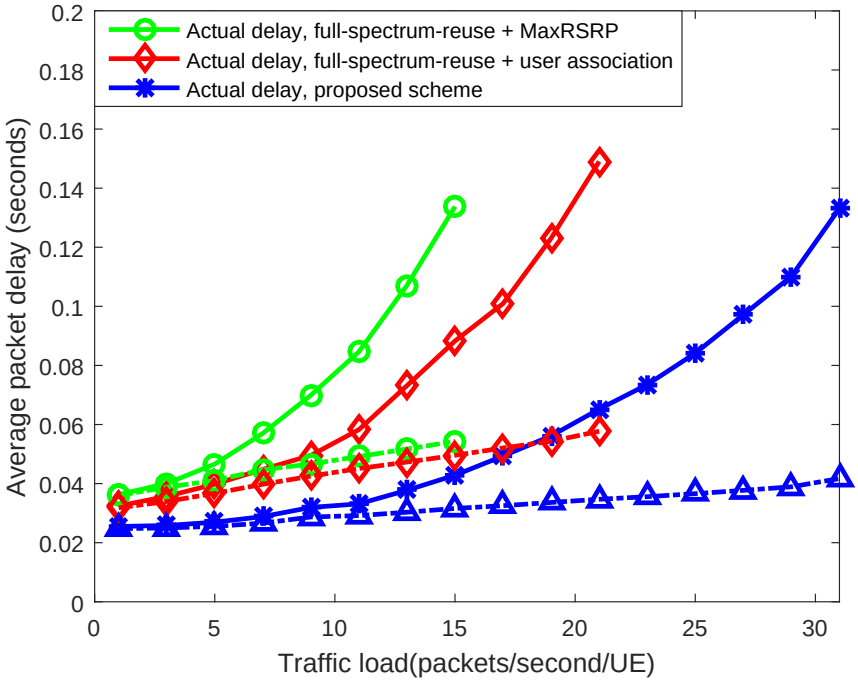


Figure 3.5. The actual packet delays of the proposed scheme and baseline schemes. Each dotted curve represents the average transmission time of the corresponding delay curve with identical marker and color.

outperforms other schemes in delay especially in the high traffic regime. Furthermore, the proposed solution is near optimal with less than 7% gap. Besides, compared with delay, the transmission time increases much more slowly with traffic load, indicating that the spectrum is efficiently allocated to mitigate interference among APs.

The obtained spectrum allocation and user association at average per user device packet arrival rate of 20 packets/second is shown in Fig. 3.4. As shown in Fig. 3.4(a), the lines connecting each AP-user pair indicate an association. To clearly present spectrum allocation and user association, the local cluster as marked in the red rectangle region is shown in enlarged display in Fig. 3.4(b) and Fig. 3.4(c). Fig. 3.4(b) shows the user association for the marked area. The numbers above each user device represent the user

device index and its traffic load, respectively. The number above each AP represents the AP index. The spectrum allocation for the marked area is shown in Fig. 3.4(c). The widths of the rectangles represent fractions of the entire spectrum of the active patterns. The solid ones in each row are the spectrum segments that are used by the corresponding AP to serve the user device whose index is marked on that spectrum segment. The algorithm achieves topology aware frequency reuse for interference management, as well as an efficient traffic aware spectrum allocation. Specifically, strongly interfering links (e.g., link $2 \rightarrow 4$ and link $3 \rightarrow 5$) are assigned different spectrum segments, and the same spectrum segments are reused by two links that are far apart (e.g., link $10 \rightarrow 25$ and link $11 \rightarrow 28$). Moreover, user devices with light traffic loads or user devices on the transmission edge of two APs (e.g., user device 5) are assigned less spectrum, and vice versa.

To compare the theoretical delay with the actual delay, a packet-level simulator is used. The actual transmission rate of a resource reserved for a pattern depends on the actual set of busy APs, which is a subset of the pattern. Fig. 3.5 compares the actual average packet delay of the proposed allocation scheme with the baseline schemes. Compared with the theoretical results in Fig. 3.3, all three schemes achieve larger throughput regions. That is because the service rate model (3.5) is "conservative", i.e., an AP's transmission rate over any spectrum segment is the worst-case rate under the corresponding pattern, which is the achievable rate when all APs in the pattern are transmitting. In addition, the proposed scheme achieves a quite larger throughput (31 packets/second/user device) than the other schemes. Moreover, the delay is also reduced by more than 50% in the high traffic regime.

3.6. Summary

We have studied slow-timescale joint user association and spectrum allocation problem in large networks. We have developed a highly scalable reformulation of the network utility maximization problem. A pattern pursuit algorithm is proposed which obtains near-optimal solution with optimality guarantee. Numerical results show substantial gains compared to all the other baseline schemes for networks with up to 1,000 APs.

CHAPTER 4

Joint Spectrum Allocation, User Association, and Power Control for Large-Scale Networks

This work studies centralized radio resource management in metropolitan area networks with a very large number of APs and user devices. A central controller collects time-averaged traffic and channel conditions from all APs and coordinates spectrum allocation, user association, and power control throughout the network on an appropriate timescale. The timescale is conceived to be seconds in today's networks, and it is likely to become faster in the future. The coordination problem in each time epoch is formulated as a network utility maximization problem, where any subset of APs may use any parts of the spectrum to serve any subsets of devices. It is proved that the network utility can be maximized by an extremely sparse spectrum allocation. By exploiting this sparsity, an efficient iterative algorithm with guaranteed convergence is developed, each iteration of which is performed in closed form. The proposed centralized optimization framework can incorporate a broad class of utility functions that account for weighted sum rates, average packet delay, and/or energy consumption, along with very general constraints on transmission powers. Numerical results demonstrate the feasibility of the algorithm for networks with up to 1,000 APs and several thousand devices. Moreover, the proposed scheme yields significantly improved throughput region and average packet delay comparing with several well-known competing schemes.

4.1. Introduction

Wireless systems have emerged as a ubiquitous part of modern data communication networks. In order to meet the ever increasing demand for wireless data services, a large number of APs of different form factors and capabilities are being deployed to improve coverage and capacity for homes, businesses, and public spaces. These APs may be densely deployed and may utilize a wide range of spectrum resources, including millimeter wave bands and unlicensed spectrum resources.

Efficient resource allocation (e.g., spectrum allocation, power control, link scheduling, routing, and congestion control) is crucial to achieving high performance and providing satisfactory QoS. Conventional spectrum allocation schemes include full spectrum reuse, partial frequency reuse, and fractional frequency reuse [**28**, **30**, **31**]. These schemes have very limited spectrum agility and are generally not adaptive to traffic conditions. In today's cellular and WiFi networks, devices typically associate with the base station from which they receive the strongest signal. Power control is another way of mitigating inter-cell interference as well as saving energy. Instead of letting each AP always transmit at full power, some inter-cell interference cancellation techniques, such as the ABS control, have been proposed in LTE and LTE-A. However, the performance of such kind of distributed algorithms is far from optimal especially for large networks with irregular AP placements.

Resource allocation in wireless networks has been extensively studied over the last few decades. Spectrum allocation schemes are basically designed from both economic perspectives (e.g., [**8**, **9**]) and technical perspectives (e.g., [**11**–**17**, **19**–**22**, **24**, **25**, **27**, **29**, **32**–**35**, **38**, **42**]). However, it remains elusive to achieve the maximum spectrum flexibility and utilization. As for the user association problem, schemes based on game theory or

optimization have also been well explored (e.g., [**20**, **21**]). It is typically hard to obtain a local optimum since the formulated problems often include integer variables and nonconvex constraints. Similar to the user association problem, power control is also known to be a hard non-convex optimization problem due to the complicated interference coupling between links. Numerous distributed power control algorithms (e.g., [**55**–**57**]) have been proposed to make proper power assignments for large-scale networks with substantially degraded system performance.

In this paper, we address the radio resource management problem by proposing a centralized optimization-based framework. A distinguishing feature here is that we propose to use a *central controller* or *cloud* to coordinate a large network with many thousand APs. This architecture can fully harness the power of cloud computing, big data, and large-scale optimization methods. Specifically, APs measure/estimate traffic and channel conditions and send the cloud regular updates. The cloud periodically solves a large-scale optimization problem for the whole network and returns the resulting allocation plan or recommendations to the APs. The cloud may also exchange information with peer network operators in case of shared spectrum.

The timescale of an allocation period is currently conceived to be on the order of seconds to allow the cloud sufficient time to solve a large-scale allocation problem using today's technologies. This timescale is fast enough to track demand shifts and large-scale fading. On this timescale, resource allocation is likely to be in a relatively coarse granularity (e.g., blocks of subcarriers). The cloud's communication overhead is small. For example, if the cloud collects one million 8-bit parameters from one thousand cells every second, the overhead is about 1 KB/s per cell or 1 MB/s in total. As technologies

improve over time, the timescale of centralized resource allocation is expected to become faster and the allocation is expected to be in finer granularities. Ultimately, the timescale is limited by the latency for collecting information from all APs via backhaul links, which maybe as short as the duration of a frame.

The proposed scheme jointly solves the problems of spectrum allocation, user association, and power control for very large networks. An efficient algorithm with low complexity is proposed to obtain a locally-optimal allocation solution in a timely manner. The proposed algorithm complements our previous work [**32**] which considered spectrum allocation and user association but did not allow power control. One key distinction of the problem formulation in this paper is that, instead of assuming all APs always transmit using full power, we allow each AP to apply an arbitrary power spectral density over the entire band.

The main contributions of this paper are summarized as follows:

- To the best of our knowledge, we are the first to propose to carry out centralized joint spectrum allocation, user association, and power allocation in metropolitan-scale networks with thousands of APs and user devices.
- We first formulate a problem of joint continuous spectrum allocation, power control, and user association. We then reduce this infinite size problem to an equivalent finite-dimensional optimization problem which is amenable to a very efficient algorithm.
- We propose an efficient algorithm with guaranteed convergence to solve the problem numerically. Each iteration is performed in closed form, which makes centralize resource allocation practical for very large networks.

- Through simulations, we demonstrate that the proposed resource allocation scheme yields significantly improved throughput region and average packet delay comparing with several well-known competing schemes.
- We generalize the problem formulation and algorithm to accommodate a much broader set of utility functions, spectral efficiencies, and power constraints than in most other studies.
- We develop a rigorous proof of the sufficiency of using piecewise constant power allocation to achieve the global optimum of the resource allocation problem under the most general set of conditions in the literature.

The remainder of this chapter is organized as follows. The system model is introduced in Sec. 4.2. The problem formulation is given in Sec. 4.3. An efficient algorithm is proposed in Sec. 4.4. A significantly generalized resource management problem is studied in Sec. 4.5. Simulation results are presented in Sec. 4.6. Concluding remarks are given in Sec. 4.7. Most technical proofs are relegated to the appendices.

4.2. System Model

We consider the downlink of a single-input single-output (SISO) network of n APs, k user devices,[1] and a central controller. The APs and the controller are connected through a wired network including backhauls. The controller continuously collects all traffic and channel/interference information from the APs. At the disposal of the controller is a radio frequency band of W Hertz. The task of the controller is to allocate the radio spectrum and power resources to all AP-device links in order to maximize a certain long-term network utility.

[1]Referred to as user equipment in LTE standards.

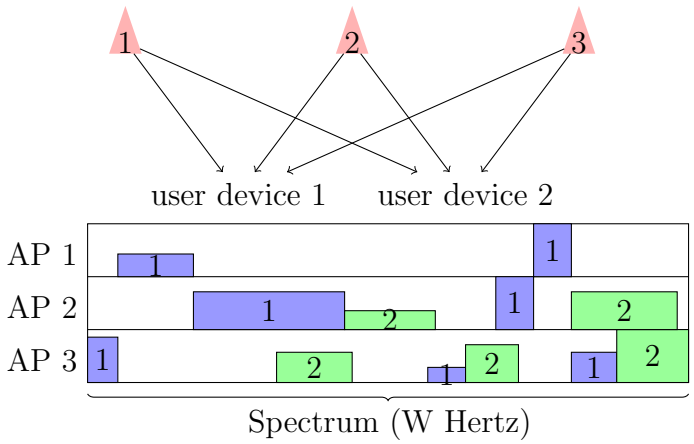


Figure 4.1. Illustration of an example of resource allocation in a 3-AP 2-device network.

The central controller performs resource allocation on a moderate timescale, which is conceived to be once every several seconds in today's networks. This allows ample time for information exchange (including channel state information feedback) and to carry out a large-scale optimization at the controller. In addition, since this timescale is much longer than the channel coherence time, the average channel conditions can be represented using path loss and the statistics of small scale fading. On this timescale, it is fair to assume that large blocks of the spectrum are homogeneous in the sense that all hertz are equally valuable and interchangeable. This spectral homogeneity assumption can be relaxed as the proposed formulation and solution can be easily generalized to treat multiple bands of different characteristics [**33**].

Denote the set of AP indexes by $N = \{1, \ldots, n\}$ and the set of device indexes by $K = \{1, \ldots, k\}$. We consider highly flexible resource allocation, in which an AP may serve any device at parts of the frequency band using any power (subject to power constraint), and a device may be served by any sets of APs.[2] Fig. 4.1 gives an example of resource

[2]In Sec. 4.3.3, we introduce a device centric model which only includes viable links in the computation.

allocation for a small network with three APs and two devices. In Fig. 4.1, the rectangles in each row are the spectrum segments used by the corresponding AP to serve the device whose index is marked on that spectrum segment. For each rectangle, its width represents the occupied bandwidth and its height represents the transmit power level. Fig. 4.1 also determines user association as AP 1 only serves device 1 while AP 2 and AP 3 serve both devices. Moreover, some links are assigned spectrum exclusively (the link from AP 1 to device 1), resulting in relatively high spectral efficiency. Some other links suffer interference from other APs operating on the same part of the spectrum, treating these interference as noise leads to relatively low spectral efficiency.[3]

The allocation problem is to determine the power spectral density functions (PSDs) of all AP-device links over the entire spectrum in order to maximize the network utility. Denote the PSD of the link from AP i to device j as a function $p_{i\to j}(f), f \in [0, W]$ and denote the PSDs of all links by $\boldsymbol{p}(f) = (p_{i\to j}(f))_{i\in N, j\in K}$. Clearly, $(\boldsymbol{p}(f),\ f \in [0, W])$ completely characterizes power spectrum allocation as well as user association in the network. Choosing a subset of AP-device links to transmit over a certain part of the spectrum is equivalent to allocating positive transmit power to these AP-device links and zero power to others over that part of the spectrum. For instance, the allocation illustrated by Fig. 4.1 corresponds to an array of piecewise constant PSDs.

For now we assume the following peak power constraint due to physical and regulatory reasons:

$$0 \le p_{i\to j}(f) \le P_{\max}, \quad \forall i \in N, j \in K, f \in [0, W]. \tag{4.1}$$

[3]While treating interference as noise is the most practical way, cooperative transmissions have recently been incorporated in a similar model [**70**].

We also assume that if an AP serves multiple devices, it uses orthogonal spectrum to do so. Mathematically,

$$\sum_{j\in K} |p_{i\to j}(f)|_0 \leq 1, \quad \forall i \in N, f \in [0, W], \tag{4.2}$$

where $|x|_0 = 1$ if $x \neq 0$, and $|x|_0 = 0$ if $x = 0$.

Constraints (4.1) and (4.2) are intrinsic to most radio system designs. We adopt them for simplicity and practicality. In Sec. 4.5 we replace those constraints to much more general forms, including total power constraint and broadcasting to multiple devices over the same slice of spectrum.

Define $g_{i\to j}$ as the average channel gain of link $i \to j$ which captures the effects of path loss and shadowing. These gains can be measured or calculated using the path loss model specified in the LTE standard [**66**]. Let $s_{i\to j}(\cdot) : \mathbb{R}^{n\times k} \to \mathbb{R}$ represent a suitable spectral efficiency function which can either be calculated based on pathloss and other impairments or be measured over time. One popular choice is to use Shannon's formula to define:

$$s_{i\to j}(\boldsymbol{p}) = \log\left(1 + \gamma_{i\to j}(\boldsymbol{p})\right), \quad \forall i \in N, j \in K, \tag{4.3}$$

where the signal-to-interference-and-noise ratio of link $i \to j$ is

$$\gamma_{i\to j}(\boldsymbol{p}) = \frac{p_{i\to j} g_{i\to j}}{n_0 + \sum\limits_{(l,q)\in(N\times K)\backslash(i,j)} g_{l\to j} p_{l\to q}}, \tag{4.4}$$

where $\boldsymbol{p} = (p_{i\to j})_{i\in N, j\in K}$ is a double array of numbers representing link powers, n_0 denotes the PSD of white Gaussian noise. Note that $(N \times K)\backslash(i, j)$ denotes the set of AP-device

links except for the link $i \to j$. Hence, $\sum_{(l,q)\in(N\times K)\setminus(i,j)} g_{l\to j}p_{l\to q}$ is the PSD of the total interference from all the APs.

Given $\boldsymbol{p}(f)$, the service rate to device j contributed by AP i can be calculated by integrating the spectral efficiency of link $i \to j$ over the entire spectrum:

$$r_{i\to j} = \int_0^W s_{i\to j}(\boldsymbol{p}(f))\,\mathrm{d}f, \quad \forall i \in N, j \in K. \tag{4.5}$$

Throughout this paper, all integrals are defined in Lebesgue's sense.[4]

The total service rate of device j denoted as r_j can be calculated by summing the service rates contributed from all APs:

$$r_j = \sum_{i\in N} r_{i\to j}. \tag{4.6}$$

4.3. Problem Formulation

Let the long-term utility be given by a function $u(\boldsymbol{r})$ of the rate tuple $\boldsymbol{r} = (r_1, \cdots, r_k)$. The goal is to maximize the long-term utility by adapting the PSDs of all AP-device links. Collecting the constraints (4.1)–(4.6), we formulate the problem as P4.1:

$$\underset{(r_j),(p_{i\to j}(f))}{\text{maximize}} \quad u(\boldsymbol{r}) \tag{P4.1a}$$

$$\text{subject to} \quad r_j = \sum_{i\in N}\int_0^W s_{i\to j}(\boldsymbol{p}(f))\,\mathrm{d}f, \quad \forall j \in K \tag{P4.1b}$$

$$\sum_{j\in K} |p_{i\to j}(f)|_0 \le 1, \quad \forall i \in N, f \in [0, W] \tag{P4.1c}$$

[4]Specifically, let μ denote the Lebesgue measure on $E = [0, W]$. Then the Lebesgue integral of a measurable function h on E is $\int_E h\,\mathrm{d}\mu$, which is also written as $\int_0^W h(f)\,\mathrm{d}f$ in this paper for convenience.

$$0 \le p_{i\to j}(f) \le P_{\max}, \quad \forall i \in N, j \in K, f \in [0, W]. \tag{P4.1d}$$

The variables are (r_j) and $(p_{i\to j}(f))$, also denoted as $\boldsymbol{r}$ and $\boldsymbol{p}(f)$. Equation (P4.1b) requires that $\boldsymbol{p}(\cdot)$ is such that $s_{i\to j}(p(f))$ is Lebesgue integrable. Also, we will justify the use of "maximize" in (P4.1a), i.e., that the maximum is actually achievable in Appendix A.7.

Most work on resource management uses the sum rate, the minimum device service rate (max-min fairness), and the sum log-rate (proportional fairness) as the utility function. The techniques developed here are applicable to all those and many other utility functions of interest. For concreteness, we consider the average packet delay [**34**] as the utility function, that is

$$u(\boldsymbol{r}) = -\sum_{j\in K} \frac{\lambda_j}{(r_j - \lambda_j)^+}, \tag{4.7}$$

where λ_j represents the amount of traffic intended for device j. The choice of this utility function is justified as follows. Suppose all downlink traffic arrive at all APs with Poisson packet arrival rate and exponential packet length. Then the packets for all the devices form k independent M/M/1 queues and (4.7) is minus the derived average packet delay. The utility (4.7) can be generalized to accommodate various kinds of distributions [**33**]. In (4.7), the extended real-valued function $1/x^+ = 1/x$ if $x > 0$ and $1/x^+ = +\infty$ if $x \le 0$. It is easy to see that $1/x^+$ is convex on $(-\infty, +\infty)$. If $r_j \le \lambda_j$, the packet delay is infinite, i.e., the system becomes unstable. We remark that, although the sum rate is often used as the utility, delay is generally a more relevant system-level figure of merit on a moderate timescale, which allows resource allocation to adapt to traffic conditions. In particular, it does not starve links with poor channel conditions.

P4.1 is computationally infeasible due to two factors: (i) the functional variable $p_{i\to j}(f)$ on $[0, W]$ implies an infinite problem size; (ii) constraints (P4.1b) and (P4.1c) are nonconvex. In the next subsections, we will first reformulate P0 to a problem with a finite number of variables, and then propose an efficient algorithm to overcome the aforementioned difficulties.

4.3.1. Finite-Size Problem Reformulation

We first reveal important characteristics of an optimal solution to P4.1, which leads to a dramatically simplified reformulation of the resource allocation problem.

Definition 4.1. *: A power allocation $(p_{i\to j}(f))_{i\in N, j\in K}$ is said to be* m-piecewise constant *if the interval $[0, W]$ can be partitioned into m sub-intervals such that, for every $i \in N$, $j \in K$, $p_{i\to j}(\cdot)$ is constant on everyone of those sub-intervals.*

Theorem 4.1. *: The optimal utility of P4.1 can be attained by a $(k+1)$-piecewise constant power allocation. If $u(\boldsymbol{r})$ is increasing in every dimension of $\boldsymbol{r}$, then a k-piecewise constant allocation suffices.*

Theorem 4.1 is proved in Appendix A.8. Theorem 4.1 guarantees that any optimal rate vector $\boldsymbol{r}$ can be achieved by dividing the entire spectrum into k intervals with bandwidths $\beta^1, \cdots, \beta^k$ on which all link PSDs are flat (assuming $u(\boldsymbol{r})$ is increasing in $\boldsymbol{r}$). The key to the proof is that all hertz are interchangeable and the frequency resource is additive. Let $L = \{1, \cdots, k\}$ denote the set of indexes of these k intervals.[5] Power allocation over

[5]Although the set L is defined to be identical to the set K here, we use a different notation to allow a flexible choice of L subsequently.

the m-th interval is represented by a vector $\boldsymbol{p}^m = (p^m_{1\to 1}, \cdots, p^m_{k\to n})$. Therefore, P4.1 can reformulated as P4.2:

$$\underset{(p^m_{i\to j}),(\beta^m)}{\text{maximize}} \quad u(\boldsymbol{r}) \tag{P4.2a}$$

$$\text{subject to} \quad r_j = \sum_{m\in L} \beta^m \sum_{i\in N} \log(1 + \frac{p^m_{i\to j} g_{i\to j}}{n_0 + \sum\limits_{(l,q)\in(N\times K)\setminus(i,j)} g_{l\to j} p^m_{l\to q}}), \quad \forall j \in K \tag{P4.2b}$$

$$\sum_{j\in K} |p^m_{i\to j}|_0 \le 1, \quad \forall i \in N, m \in L \tag{P4.2c}$$

$$\sum_{m\in L} \beta^m = W, \tag{P4.2d}$$

$$0 \le p^m_{i\to j} \le P_{\max}, \quad \forall i \in N, j \in K, m \in L \tag{P4.2e}$$

$$\beta^m \ge 0, \quad \forall m \in L. \tag{P4.2f}$$

In contrast to P4.1 with infinite dimensions, P4.2 has only $k+k^2n = O(k^2n)$ variables. For a large network consisting of thousands of APs and devices, the number of variables is still prohibitively large. In addition, the set of constraints (P4.2b) and (P4.2c) are nonconvex. Hence P4.2 is still difficult to solve. In the next subsections, we will reveal some important facts in order to solve a simple case of P4.2 where the utility function $u(\boldsymbol{r})$ is an affine function of $\boldsymbol{r}$. The technique motivates the development of the highly scalable algorithm in Sec. 4.4.

4.3.2. Affine Utility Function

Proposition 4.1. *: Suppose the utility function $u(\boldsymbol{r})$ is affine, i.e., it can be written as*

$$u(\boldsymbol{r}) = d + \sum_{j \in K} c_j r_j, \tag{4.8}$$

for some constants d and $(c_j)_{j \in K}$. Then the maximum utility in P4.2 can be attained by letting each AP apply a flat PSD over the entire spectrum to serve a single user device (or no user device if the PSD is all zero).

Proof. Let $\boldsymbol{p}^m$ represent the array $(p^m_{i \to j})_{i \in N, j \in K}$. Then P4.2 can be rewritten in the following equivalent form, referred to as P4.3:

$$\underset{(p^m_{i\to j}),(\beta^m)}{\text{maximize}} \quad d + \sum_{m \in L} \beta^m \sum_{j \in K} c_j \sum_{i \in N} \log \left(1 + \frac{p^m_{i \to j} g_{i \to j}}{n_0 + \sum\limits_{l \in N : l \neq i} g_{l \to j} \sum\limits_{q \in K} p^m_{l \to q}} \right) \tag{P4.3a}$$

$$\text{subject to} \quad (P4.2c), (P4.2d), (P4.2e), (P4.2f). \tag{P4.3b}$$

Let $(\boldsymbol{p}, \boldsymbol{\beta})$ be a feasible solution to P4.3. Define $m^* \in L$ as a maximizer of $\sum\limits_{j \in K} c_j \sum\limits_{i \in N} \log(1 + \frac{p^m_{i \to j} g_{i \to j}}{n_0 + \sum\limits_{l \in N : l \neq i} g_{l \to j} \sum\limits_{q \in K} p^m_{l \to q}})$. It is easy to see that for any fixed power allocation $(p^m_{i \to j})$, the utility is always maximized by letting $\beta^{m^*} = 1$ and $\beta^m = 0$ for all $m \neq m^*$, where the constraints (P4.2c)-(P4.2f) remain to hold. Hence letting each AP apply a flat power spectral density over the entire spectrum to serve no more than one device is optimal. □

From now on we refer to one assignment of link powers $(p_{i \to j})_{i \in N, j \in K}$ as a (power) profile. Proposition 4.1 implies that when the utility function is affine (which can be

regarded as the weighted sum rate), it is optimal to apply a single power profile over the entire spectrum, where each AP serves a single device. Therefore, for the affine utility function (4.8), P4.2 (and P4.1) is reduced to the following P4.4:

$$\underset{(p_{i\to j})}{\text{maximize}} \quad \sum_{j\in K} c_j \sum_{i\in N} \log\left(1+\frac{p_{i\to j}g_{i\to j}}{n_0+\sum\limits_{l\in N:l\neq i} g_{l\to j}\sum\limits_{q\in K} p_{l\to q}}\right) \tag{P4.4a}$$

$$\text{subject to} \quad \sum_{j\in K} |p_{i\to j}|_0 \leq 1, \quad \forall i\in N \tag{P4.4b}$$

$$0\leq p_{i\to j}\leq P_{\max}, \quad \forall i\in N, j\in K. \tag{P4.4c}$$

4.3.3. A Fast Algorithm

In the following, we first reformulate P4.4 to remove the ℓ_0 norm constraint (P4.4b) and further reduce the number of variables, and then solve it using the quadratic transform technique proposed by Shen and Yu [**71**, **72**].

In a large network with many APs, a device can in general only be served by a small subset of nearby APs due to path loss. For every $j\in K$, let N_j denote the set of APs whose received signal-to-noise ratios at device j are above a certain threshold ξ, i.e.,

$$N_j \triangleq \left\{ i\in N \;\middle|\; \frac{p_i g_{i\to j}}{n_0} > \xi \right\}, \forall j\in K. \tag{4.9}$$

We define N_j as the *neighborhood* of device j. Device j treats all APs outside N_j as stationary noise sources whose total PSD (including noise) is denoted as n_j. This can be arbitrarily precise as N_j may include all APs except those received by device j at

well below the noise level. Then for each AP i, define the set of devices that AP i can potentially serve as K_i, specifically,

$$K_i \triangleq \left\{ j \in K \;\middle|\; i \in N_j \right\}, \forall i \in N. \tag{4.10}$$

It is fair to assume the size of all N_js and K_is are upper bounded by a constant α, i.e., $|K_i|, |N_j| \leq \alpha$.

Since Proposition 4.1 guarantees that each AP serves no more than one device, let variable z_i denote the device served by AP i, where $z_i = 0$ if AP i serves no device. For convenience, we introduce additional parameters $c_0 = 0$ and $g_{i\to 0} = 0$ for all $i \in N$.Then based on Proposition 4.1, P4.4 can be reformulated as P4.5:

$$\underset{\boldsymbol{z},\boldsymbol{p}}{\text{maximize}} \quad \sum_{i\in N} c_{z_i} \log(1 + \frac{p_i g_{i\to z_i}}{n_{z_i} + \sum\limits_{l\in N_{z_i}: l\neq i} p_l g_{l\to z_i}}) \tag{P4.5a}$$

$$\text{subject to} \quad 0 \leq p_i \leq P_{\max}, \quad \forall i \in N \tag{P4.5b}$$

$$z_i \in K_i \cup \{0\}, \quad \forall i \in N. \tag{P4.5c}$$

In P4.5, $\boldsymbol{z} = (z_i)_{i\in N}$ are the device decisions for APs. It is possible for multiple APs to serve the same device, i.e., $z_i = z_{i'}$ for $i \neq i'$. Here, $\boldsymbol{p} = (p_1, \cdots, p_n)$ is the vector consisting of the constant PSDs of all APs over the entire spectrum. P4 has $2n$ variables in total.

P4.5 is a fractional programming problem studied by Shen and Yu [**71**, **72**]. Next, we are going to apply a similar technique as in [**72**]. First, apply the Lagrangian dual

transform to reformulate P4.5 to P4.6 by introducing non-negative auxiliary variable $\boldsymbol{\gamma}$:

$$\underset{\boldsymbol{z},\boldsymbol{p},\boldsymbol{\gamma}}{\text{maximize}} \quad \sum_{i\in N} c_{z_i}\log(1+\gamma_i) - \sum_{i\in N} c_{z_i}\gamma_i + \sum_{i\in N} \frac{c_{z_i}(1+\gamma_i)p_i g_{i\to z_i}}{n_{z_i} + \sum\limits_{l\in N_{z_i}} p_l g_{l\to z_i}} \tag{P4.6a}$$

$$\text{subject to} \quad 0 \le p_i \le P_{\max}, \quad \forall i \in N \tag{P4.6b}$$

$$z_i \in K_i \cup \{0\}, \quad \forall i \in N. \tag{P4.6c}$$

When $\boldsymbol{z}$ and $\boldsymbol{p}$ are fixed, we can obtain the optimal $\boldsymbol{\gamma}$ by calculating the partial derivative of (P4.6a) with respect to $\boldsymbol{\gamma}$:

$$\gamma_i^* = \frac{p_i g_{i\to z_i}}{n_{z_i} + \sum\limits_{l\in N_{z_i}: l\neq i} p_l g_{l\to z_i}}. \tag{4.11}$$

P4.5 and P4.6 are equivalent because the objective (P4.6a) is concave in $\boldsymbol{\gamma}$ and (P4.5a) is equal to (P4.6a) with $\boldsymbol{\gamma}^*$ in it.

Next, by applying the quadratic transform technique in [**71**], P4.6 is rewritten in an equivalent form to P4.7:

$$\begin{aligned}\underset{\boldsymbol{z},\boldsymbol{p},\boldsymbol{\gamma},\boldsymbol{y}}{\text{maximize}} \quad & \sum_{i\in N} c_{z_i}\log(1+\gamma_i) - \sum_{i\in N} c_{z_i}\gamma_i + \sum_{i\in N}\Bigg(2y_i\sqrt{c_{z_i}(1+\gamma_i)p_i g_{i\to z_i}} \\ & - {y_i}^2\Bigg(n_{z_i} + \sum_{l\in N_{z_i}} p_l g_{l\to z_i}\Bigg)\Bigg)\end{aligned} \tag{P4.7a}$$

$$\text{subject to} \quad 0 \le p_i \le P_{\max}, \quad \forall i \in N \tag{P4.7b}$$

$$z_i \in K_i \cup \{0\}, \quad \forall i \in N \tag{P4.7c}$$

where $\boldsymbol{y}$ consists of an array of auxiliary variables. Since (P4.7a) is a quadratic function of each y_i, when $\boldsymbol{z}$, $\boldsymbol{p}$, and $\boldsymbol{\gamma}$ are fixed, we can derive the optimal $\boldsymbol{y}$ by:

$$y_i^* = \frac{\sqrt{c_{z_i}(1+\gamma_i)p_i g_{i\to z_i}}}{n_{z_i} + \sum\limits_{l\in N_{z_i}} p_l g_{l\to z_i}}. \tag{4.12}$$

Similarly, the optimal power allocation $\boldsymbol{p}$ for fixed $\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\gamma}$ are

$$p_i^* = \min\left\{ P_{\max}, \frac{c_{z_i}(1+\gamma_i) g_{i\to z_i} {y_i}^2}{(\sum\limits_{l\in N_{z_i}} {y_l}^2 g_{i\to s_l})^2} \right\}. \tag{4.13}$$

Finally, since the the utility function (P4.7a) can be decoupled in terms of each AP, the optimal z_i should point to the device that maximizes the utility for AP i, i.e.,

$$z_i = \begin{cases} 0, \text{if } \max\limits_{j\in K_i}\{c_j \log(1+\gamma_i) - c_j\gamma_i + 2y_i\sqrt{c_j(1+\gamma_i)p_i g_{i\to j}} - {y_i}^2(n_j + \sum\limits_{l\in N_j} p_l g_{l\to j})\} < 0 \\ \arg\max\limits_{j\in K_i}\{c_j \log(1+\gamma_i) - c_j\gamma_i + 2y_i\sqrt{c_j(1+\gamma_i)p_i g_{i\to j}} - {y_i}^2(n_j + \sum\limits_{l\in N_j} p_l g_{l\to j})\}, \\ \text{otherwise} \end{cases} \tag{4.14}$$

In the case when multiple devices provide the same amount of contribution to the utility function for an AP, we choose the j with minimal index to break the tie.

So far, we derive the optimal $\boldsymbol{\gamma}$, $\boldsymbol{y}$, $\boldsymbol{p}$, and $\boldsymbol{z}$ when other variables are fixed. By updating these variables iteratively, it can be easily verified that the utility (P4.5a) is monotonically nondecreasing and hence must converge. This gives us an efficient algorithm referred to as Algorithm 4.1 to solve P4.5.

Algorithm 4.1 Joint power control and user association for affine utility function

Initialization: Pick random $z_i \in K_i \cup \{0\}$ and $p_i \in [0, P_{\max}]$ for all $i \in N$.
repeat
Step 1. Update $\boldsymbol{\gamma}$ by (4.11);
Step 2. Update $\boldsymbol{y}$ by (4.12);
Step 3. Update $\boldsymbol{p}$ by (4.13);
Step 4. Update $\boldsymbol{z}$ by (4.14);
until Convergence

Algorithm 4.1 starts with an arbitrary initial power profile $(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\gamma})$ and searches for a single optimal power profile, where APs are shut down sequentially to yield an active subset of APs eventually. The total number of variables needed to be solved at each iteration is $O(n)$, moreover, each iteration in Algorithm 4.1 is performed in closed form using local variables. In the next section, we are going to combine all the aforementioned propositions and techniques to solve P0 for more general utility functions.

4.4. A Scalable Joint Power Allocation and User Association Algorithm

Theorem 4.1 ensures that an optimal allocation partitions the spectrum into no more than k segments. The problem is to identify this set of segments and optimize these bandwidths. Each segment corresponds to a power profile.

We next describe Algorithm 4.2, a profile pursuit algorithm which iteratively identifies segment one by one until the utility converges. To initialize, we assume one feasible segment $\boldsymbol{p}^0$ is given and is added to the set of collected segment P. One example of the initial segment can be the conventional full-spectrum-reuse scheme in which each AP reuses the entire spectrum with full transmit power. Denote the resulting rate vector by $\boldsymbol{r}^0$. With P and $\boldsymbol{r}^0$ as the input to Algorithm 4.2, during the t-th iteration, one new segment $\boldsymbol{p}^{t+1}$ is identified by optimizing the local linearization of the utility function at

the point $\boldsymbol{r}^t$ using Algorithm 4.1. The new segment then is added to the set P if it is not already included, otherwise a randomly generated new segment will be added instead. Then P4.2 is re-optimized using all the collected segments in P so far to allocate the optimal bandwidth of each segment. Since the set of collected segments grows after each iteration, the utility function increases after each iteration until it converges.

Algorithm 4.2 Iterative algorithm for profile pursuit

Initialization: Pick full spectrum reuse power allocation denoted by $\boldsymbol{p}^0$ as the initial feasible solution which gives a feasible rate vector $\boldsymbol{r}^0$; $t \leftarrow 0$; $P \leftarrow \{\boldsymbol{p}^0\}$;
repeat
Step 1. Solve P4.2 with the objective function being $\langle \nabla u(\boldsymbol{r}^{(t)}), \boldsymbol{r} \rangle$ using Algorithm 4.1, which gives us $\boldsymbol{p}^{t+1}$.
Step 2. If $\boldsymbol{p}^{t+1} \notin P$, $P \leftarrow P \cup \{\boldsymbol{p}^{t+1}\}$; otherwise, add an arbitrary new segment that is not in P, $t \leftarrow t+1$.
Step 3. Solve P4.2 given power allocations in P and obtain the rate vector $\boldsymbol{r}^{(t)}$.
until Convergence

Algorithm 4.2 is a Frank-Wolfe type algorithm (also known as the conditional gradient algorithm) [**73**]. Unique to Algorithm 4.2 is that each iteration identifies one new power profile to add to the set P so as to re-optimize P4.2.

4.5. General Resource Allocation Problem

In this section, the problem formulation of resource management is further generalized to accommodate a much broader set of utility functions, spectral efficiencies, and power constraints. The general problem can also be efficiently solved using the proposed technique in this paper.

First of all, for the spectral efficiency function $s_{i\to j}(\cdot)$ in (4.3), other than using the Shannon formula, it can be any measurable function.

Secondly, the PSD constraint can be more general. For example, the SISO PSD constraint (4.2) can be generalized to accommodate multicasting over the same spectrum, i.e.,

$$\sum_{j \in K} |p_{i \to j}(f)|_0 \leq d_i, \quad \forall i \in N, \forall f \in [0, W] \tag{4.15}$$

where $d_i > 0$ constrains the degrees of freedom of AP i. In an extreme case, $d_i = k$ for all $i \in N$. That is, an AP may serve all devices using the same slice of spectrum simultaneously. The additional degrees of freedom are particularly useful in the case of multiple-input multiple-output systems with multiple antennas. Moreover, the PSD constraint (4.1) can also be generalized to the following:

$$0 \leq p_{i \to j}(f) \leq Q, \quad \forall i \in N, j \in K, \forall f \in [0, W]. \tag{4.16}$$

$$\left\| \sum_{j \in K} p_{i \to j}(\cdot) \right\|_\alpha \leq P_{\max}, \quad \forall i \in N, \tag{4.17}$$

where $\|h\|_\alpha = \left(\int_0^W (h(x))^\alpha \, \mathrm{d}x \right)^{\frac{1}{\alpha}}$ and $\alpha \geq 1$. The left hand side of (4.17) is the α-norm of the PSDs allocated to AP i. If $\alpha = 1$, (4.17) constrains the total power of every AP over the entire available spectrum. On the other hand, when $\alpha \to \infty$, (4.17) becomes the previous peak PSD constraint.

Furthermore, in addition to optimizing throughput and packet delay, we can also incorporate energy efficiency into the utility function. Specifically, APs' transmit powers

can be expressed as

$$P_i = \sum_{j \in K} \int_0^W p_{i \to j}(f) \, \mathrm{d}f, \quad i \in N. \tag{4.18}$$

In addition, if AP i can be switched off to save its maintenance power (denoted as C_i) when it is not allocated power (i.e., $P_i = 0$), then the utility function regarding energy efficiency can be expressed as minus the total energy cost:

$$u_e(\boldsymbol{P}) = -p \sum_{i \in N} (P_i + \mathbb{1}_{\{P_i > 0\}} C_i), \tag{4.19}$$

where p is the "price" of power and $\boldsymbol{P} = (P_i)_{i \in N}$. Denote the utility function regarding spectral efficiency as $u_s(\boldsymbol{r})$, then the utility function that incorporates both spectrum and energy efficiencies is expressed as

$$U(\boldsymbol{r}, \boldsymbol{P}) = u_s(\boldsymbol{r}) + u_e(\boldsymbol{P}). \tag{4.20}$$

Theorem 4.2. *: Let $s_{i \to j}(\cdot) : \mathbb{R}^{n \times k} \to \mathbb{R}$ be a bounded measurable function. A $(k + n + 1)$-sparse piecewise constant power allocation optimally solves the following problem:*

$$\underset{(r_j),(p_{i \to j}(f)),(P_i)}{\textit{maximize}} \quad U(\boldsymbol{r}, \boldsymbol{P}) \tag{P4.8a}$$

$$\textit{subject to} \quad r_j = \sum_{i \in N} \int_0^W s_{i \to j}(\boldsymbol{p}(f)) \, \mathrm{d}f, \quad \forall j \in K \tag{P4.8b}$$

$$P_i = \sum_{j \in K} \int_0^W p_{i \to j}(f) \, \mathrm{d}f, \quad \forall i \in N \tag{P4.8c}$$

$$\sum_{j \in K} |p_{i \to j}(f)|_0 \leq d_i, \quad \forall i \in N, f \in [0, W] \tag{P4.8d}$$

$$\left(\int_0^W \left[\sum_{j\in K} p_{i\to j}(f)\right]^{\alpha} \mathrm{d}f\right)^{\frac{1}{\alpha}} \leq P_{max}, \quad \forall i \in N \tag{P4.8e}$$

$$0 \leq p_{i\to j}(f) \leq Q, \quad \forall i \in N, j \in K, f \in [0, W]. \tag{P4.8f}$$

The proof of Theorem 4.2 is given in Appendix A.7. P4.8 describes a very general resource management problem, which essentially encompasses most existing optimization-oriented formulations of physical resource allocation (including spectrum and time slot allocation, user association, and power control) in the literature. The significance here is that the sparsity guaranteed by Theorem 4.2 allows us to reduce the continuous PSD function allocation problem to that of finding a discrete set of $(k + n + 1)$ power profiles.

If the spectral efficiency function $s_{i\to j}(\cdot)$ can be written as a concave function of the SINR given by (4.4), that is

$$s_{i\to j}(\boldsymbol{p}) = A(\gamma_{i\to j}(\boldsymbol{p})), \tag{4.21}$$

and $A(\cdot) : \mathbb{R} \to \mathbb{R}$ is a concave function, then P4.8 can be solved using similar technique proposed in Section 4.4.

4.6. Numerical Results

Table 4.1. Parameter Configurations.

Parameters	Value/Function
AP transmit power	23 dBm
Total bandwidth	10 MHz
Average packet length	0.5 Mbits
AP to user device pathloss (LOS)	$30.18 + 26.7 \log_{10}(R)$
AP to user device pathloss (NLOS)	$34.53 + 36 \log_{10}(R)$
Shadowing standard deviation (LOS)	4 dB
Shadowing standard deviation (NLOS)	10 dB

In this section, we investigate the performance gain of the proposed allocation scheme by comparing with the following baseline schemes:

(1) Full-spectrum-reuse + maximum reference signal receive power (MaxRSRP): Every AP reuses all available spectrum with full transmit power and every device is associated to the strongest AP in terms of the received power.

(2) Full-spectrum-reuse + optimal user association: Every AP reuses all available spectrum with full transmit power and device association is optimized for the utility.

(3) The pattern-pursuit algorithm proposed in [**32**] which optimizes both spectrum allocation and user association assuming full transmit power at each AP.

The performance metric is the average packet delay which is also the objective in P4.1. We use parameters compliant with the LTE standard [**66**] as given in Table 4.1. To make fair

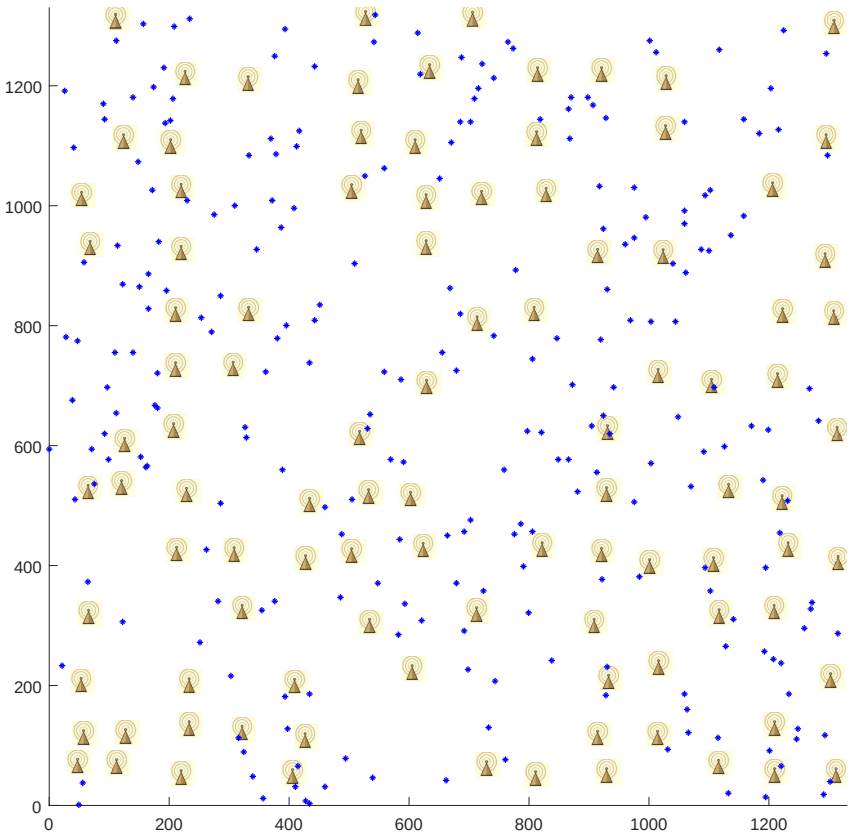


Figure 4.2. Deployment of the medium scale network.

comparisons, we use exactly the same network (same topology, same channel conditions, and same traffic loads) as in our previous paper [**32**], in which all active transmissions use the maximum allowed PSD with binary (0 or peak) power control.

4.6.1. Performance in Medium Scale Networks

We first compare the performance of the proposed scheme in a network of medium size. We randomly drop 100 APs and 250 user devices over a 1,330 $\times$ 1,330 m^2 area as shown

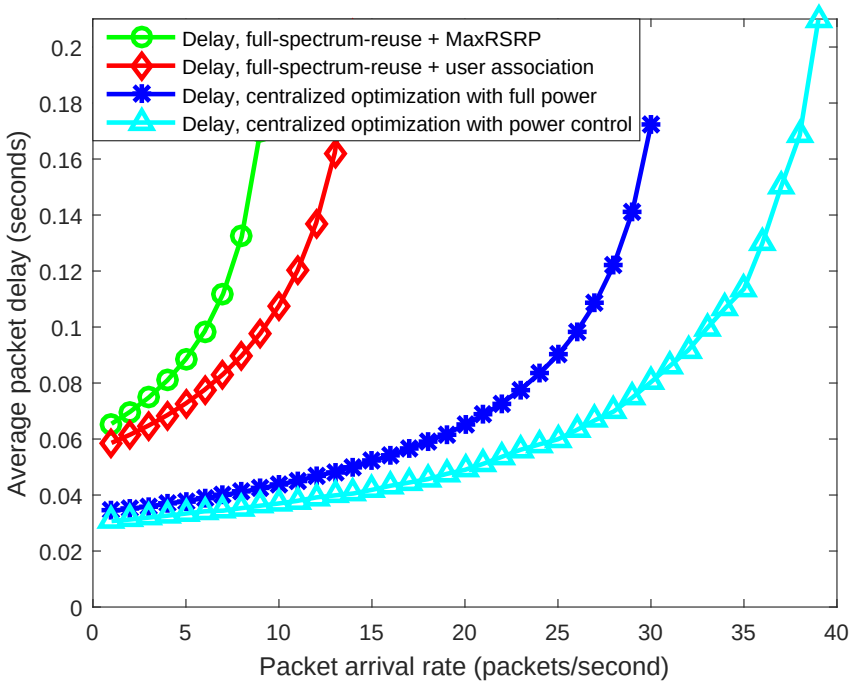


Figure 4.3. The theoretical packet delays (the objective function) of the proposed scheme and baseline schemes.

in Fig. 4.2. The objective (average packet delay) versus traffic intensity curves are shown in Fig. 4.3.

From Fig. 4.3, it can be observed that as the average device traffic increases to above 8, 13, and 30 packets/second, respectively, all three baseline schemes fail to support all the devices. While the proposed solution has significantly larger throughput (above 39 packets/second) than the other schemes. The proposed solution also significantly reduces the delay especially in the high traffic regime.

To validate the obtained allocation solution, we have built an event-driven packet-level simulator. In the simulator, the instantaneous transmission rate of a link depends on the current set of busy APs occupying the same part of the spectrum. Fig. 4.4 compares

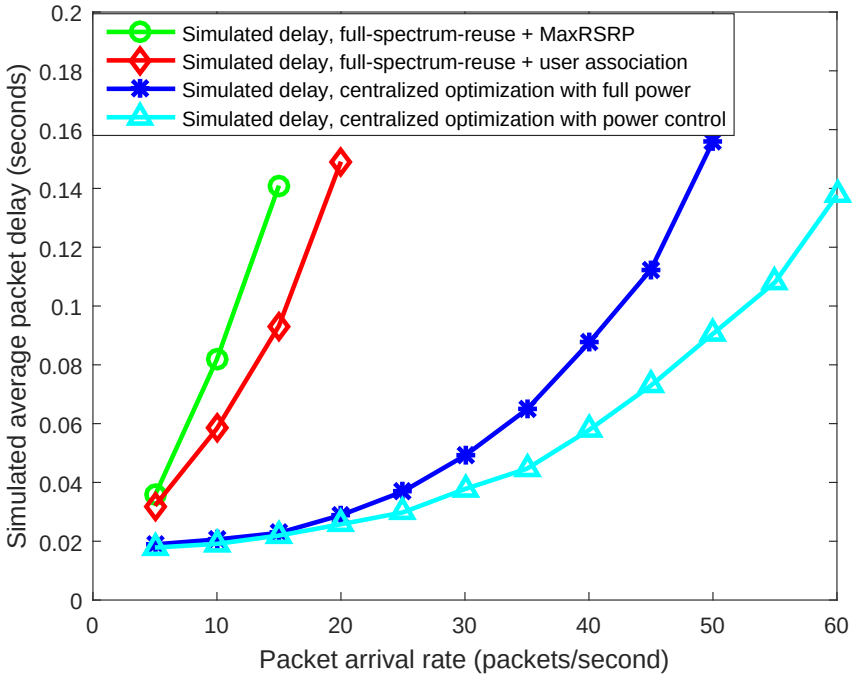


Figure 4.4. The simulated packet delays of the proposed scheme and baseline schemes.

the simulated average packet delay of the proposed allocation scheme with the baseline schemes. Compared with the theoretical results in Fig. 4.3, all the schemes achieve larger throughput regions. That is because the service rate model (4.3) is "conservative", i.e., an AP's transmission rate over any spectrum segment is the worst-case rate, which is the achievable rate when all APs in this segment are transmitting. In addition, the proposed scheme achieves a quite larger throughput (60 packets/second/user device) than the other schemes. Moreover, the delay is also reduced by more than 50% in the high traffic regime.

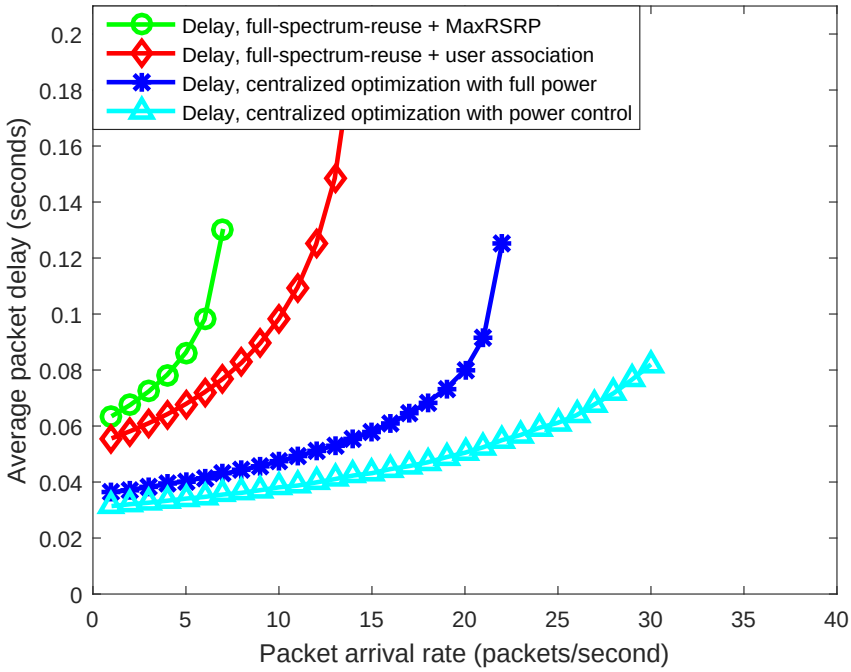


Figure 4.5. Comparison with the baseline schemes.

4.6.2. Performance in Large Scale Networks

In this subsection, the proposed allocation scheme is used to compute the allocation solution for a large network of 1,000 APs and 2,500 devices over a 4,200 m $\times$ 4,200 m area. The AP density remains the same as that of the medium-scale network. The objective (average packet delay) versus traffic intensity curves are shown in Fig. 4.5.

From Fig. 4.5, it is clear that the proposed solution has significantly larger throughput (above 30 packets/second) than full-spectrum reuse with maxRSRP association (7 packets/second), full-spectrum reuse with optimal user association allocation (14 packets/second) and centralized optimization with full power (22 packets/second). The proposed solution also outperforms other schemes in delay especially in the high traffic regime.

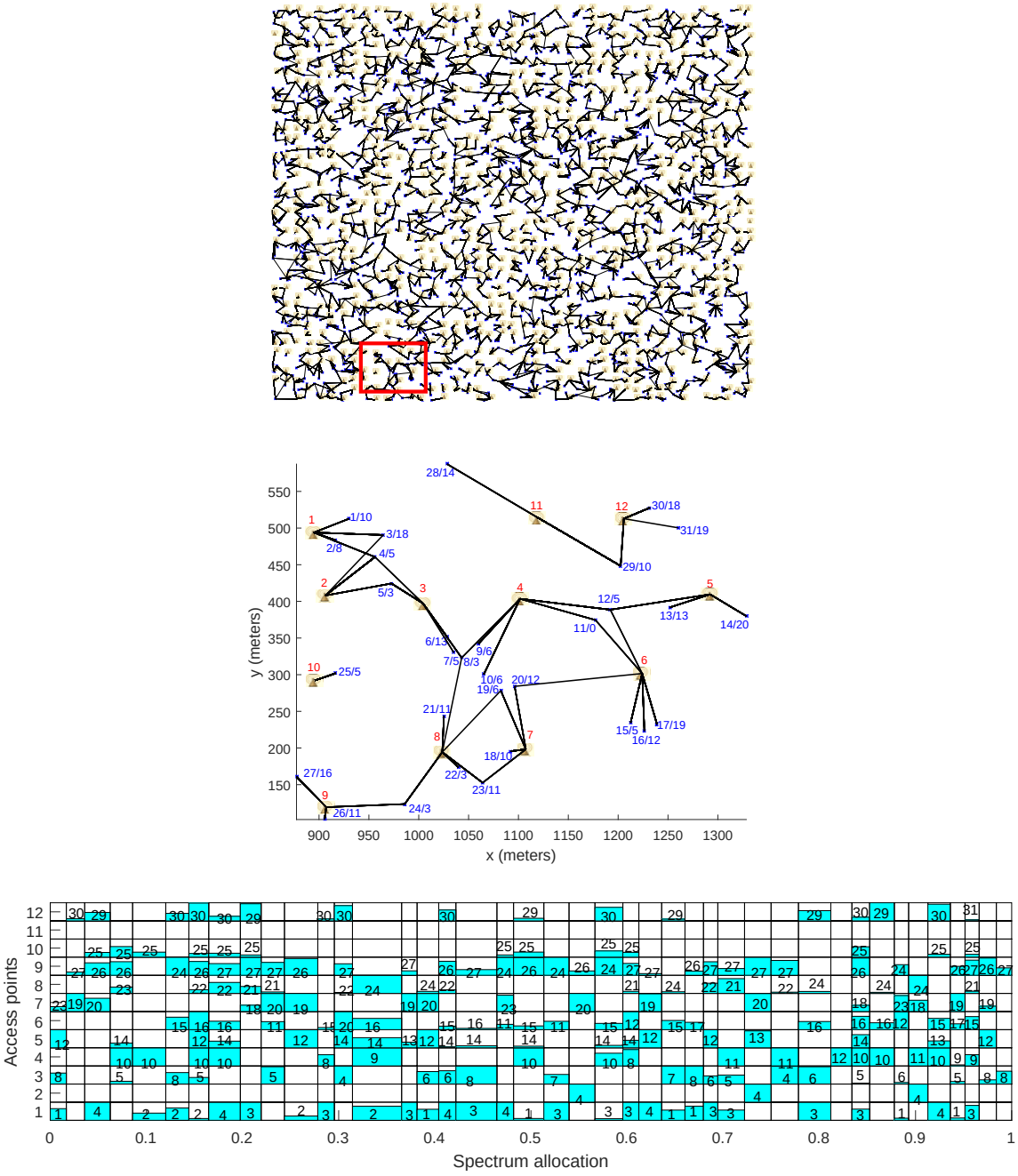


Figure 4.6. Resource management in very large networks. (a) Deployment and user association for the large network. (b) Topology graph for the marked area in Fig. 4.6(a). (c) Allocation graph for the marked area in Fig. 4.6(a).

The obtained resource allocation at arrival rate of 25 packets/second/device is shown in Fig. 4.6. As shown in Fig. 4.6(a), the lines connecting each AP-user pair indicate an association. To clearly present spectrum allocation, user association, and power allocation, the local cluster as marked in the red rectangle region is shown in enlarged display in Fig. 4.6(b) and Fig. 4.6(c). Fig. 4.6(b) shows the user association for the marked area. The numbers above each user device represent the user device index and its traffic load, respectively. The number above each AP represents the AP index. The spectrum and power allocation for the marked area is shown in Fig. 4.6(c). The widths of the rectangles represent fractions of the entire spectrum of the active segments. The solid ones in each row are the spectrum segments that are used by the corresponding AP to serve the user device whose index is marked on that spectrum segment. Therefore, each row can be viewed as the PSD of the corresponding AP. The algorithm achieves topology aware frequency reuse for interference management, as well as an efficient traffic aware spectrum allocation. Specifically, strongly interfering links (e.g., link $2 \rightarrow 4$ and link $3 \rightarrow 5$) are assigned different spectrum segments, and the same spectrum segments are reused by two links that are far apart (e.g., link $10 \rightarrow 25$ and link $9 \rightarrow 26$). Moreover, user devices with light traffic loads or user devices on the transmission edge of two APs (e.g., user device 5) are assigned less spectrum and less power, and vice versa. These numerical results demonstrate that the algorithm achieves topology aware frequency reuse for interference management, as well as an efficient traffic aware spectrum allocation.

Table 4.2. Computational cost for Algorithm 4.1 and Algorithm 4.2

Networks scales	10 APs 25 devices	100 APs 250 devices	1,000 APs 2,500 devices
# of iterations in Alg. 4.1	85.0	92.5	122.6
# of iterations in Alg. 4.2	12.8	23.8	27.2
# of active segments	10.6	23.1	27.2
Average runtime (seconds)	1.8	9.5	169.9

4.6.3. Complexity discussion

Table 4.2 shows the computational cost for the proposed method under different network scales. The density of APs and devices remains the same for all networks. One observation is that the number of iterations for Algorithm 4.1 to converge varies slightly even when the network size is enlarged by hundred-fold. The algorithms are run on an Intel Core Xeon 3.6 GHz 6-core computer. It can be observed that the total computation only takes 1.8 seconds for small-scale networks and 170 seconds for metropolitan-scale networks because of the closed-form updates in every iteration.[6] In addition, the number of active segments is very close to the number of candidate segments for all networks, meaning that at each iteration of Algorithm 4.2, a useful segment is almost always obtained to contribute to the final solution.

The proposed algorithm is able to solve the resource allocation problem for super large networks using small amount of time due to its small number of iterations and low per-iteration cost. As only convergence to a local optimum is guaranteed in all cases, the converged values may differ depending on the starting point.

[6]The runtime is expected to be reduced significantly to fit in a moderate timescale using a powerful cluster/cloud.

4.7. Summary

We have studied a challenging joint resource allocation problem in metropolitan-scale networks. We have developed a scalable formulation by exploiting the hidden sparse structure of an optimal allocation and proposed an efficient algorithm to solve the reformulated problem with guaranteed convergence. Moreover, each iteration is performed in closed form, which makes centralized resource allocation practically feasible even for very large networks. From numerical results, significant gains are observed compared with conventional resource management schemes.

One possible future work is to incorporate the slow-timescale solution with fast-timescale link scheduling [**42**]. That is, the slow-timescale solution may guide each AP to better schedule transmission links locally on a fast-timescale. Another extension is to consider allocating discrete subcarriers rather than continuous spectrum resources, which may involve more complex problem formulations. This will allow the proposed solution to be adopted in current and emerging radio access network (RAN) standards.

CHAPTER 5

Conclusions and Future Work

As for wireless networks, resource allocation is the process of deciding how to implement spectrum management, user association as well as power control. In this thesis, we address resource allocation and performance optimization for different kinds of wireless networks. in particular, we incorporate spectrum allocation and user association as well as the modeling of licensed and unlicensed bands into problem formulation to allocate different kinds of resources. Subsequently, we introduced the user-centric model to reformulate the original problem. By applying linear and nonlinear optimization techniques, a highly efficient iterative algorithm is designed with guaranteed optimality gap. Finally, power control in added into the joint optimization framework. Results are presented mainly by comparing the performance of conventional approaches with the proposed algorithms through simulations. We summarize our results by each chapter below.

For allocating resources over multiple radio access technologies, in Chapter 2, we have proposed two optimization-based spectrum allocation schemes along with efficient algorithms for computing the allocations. The proposed solutions take into account traffic loads, network topology, as well as external interference levels in the unlicensed bands. The performance is then compared with the conventional schemes (e.g., full frequency reuse). Packet-level simulation results show that the proposed schemes significantly outperform orthogonal and full-frequency-reuse allocations under all traffic conditions.

In Chapter 3, we have studied the scalability of the resource allocation schemes for large-scale networks, in the effort to make centralized optimization practically realizable for large-scale networks. Comparing with the above framework which formulates the resource allocation problem as a convex optimization problem with an exponential number of variables in the network size. A scalable reformulation is obtained by exploiting the geometric graph nature of the network and provable sparsity of the optimal solution. Moreover, we designe a highly efficient algorithm which iteratively finds patterns until a certain optimality gap is reached. Numerical results show substantial gains compared to all the other baseline schemes for networks with up to 1000 access points.

In Chapter 4, we have extended the above framework to further improve the system performance by jointly optimizing spectrum allocation, user association and power control. The proposed algorithm complements our previous work [**32**] that has only considered spectrum allocation and UE association. One key distinction of the new problem formulation is that, instead of assuming all APs always transmit using full power, we allow each AP to apply a continuous power spectral density over the entire band, i.e., the size of the problem of power allocation at each AP has already grown to infinity. Moreover, we also generalize the problem formulation to accommodate various classes of network optimization problems.

The thesis mainly covers slow-timescale interference and resource management. We propose a way to jointly optimize different kinds of resources for large-scale networks. The proposed algorithms are applicable to most conventional utility functions. Such framework can be potentially extended to take into account the features in HetNets, such as backhaul resource allocation (with relay), CoMP transmission and beamforming. One

future research direction is to incorporate the existing slow-timescale solution with fast-timescale link scheduling. For example, given the solution in Chapter 4, how to make use of such slow-timescale power control solution on the fast-timescale link scheduling problem.

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?," IEEE *J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065-1082, 2014.

[2] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in HetNets: old myths and open problems," IEEE *Wireless Commun.*, vol. 21, no. 2, pp. 18-25, 2014.

[3] J. Liu, W. Xiao, and A. C. K. Soong, "Design and Deployment of Small Cell Networks", Cambridge University Press, 2014.

[4] S. Y. Lien, S. L. Shieh, Y. Huang, B. Su, Y. L. Hsu, and H. Y. Wei, "5g new radio: Waveform, frame structure, multiple access, and initial access," IEEE *Commun. Mag.*, vol. 55, no. 6, pp. 64-71, 2017.

[5] B. Soret, A. D. Domenico, S. Bazzi, N. H. Mahmood, and K. I. Pedersen, "Interference coordination for 5g new radio," IEEE *Wireless Commun.*, 2018.

[6] Z. Zhou, D. Guo and M. L. Honig, "Allocation of licensed and unlicensed spectrum in heterogeneous networks", in *Proc. IEEE Australian Communication Theory Workshop*, Melbourne, VIC, Australia, 2016.

[7] A. Khandekar, N. Bhushan, T. Ji, and V. Vanghi, "LTE-Advanced: Heterogeneous networks," in *European Wireless Conference*, pp. 978-982, 2010.

[8] C. Chen, R. Berry, M. L. Honig and V. Subramanian, "The impact of unlicensed access on small-cell resource allocation," *IEEE INFOCOM*, San Francisco, CA, Apr. 2016.

[9] M. Rebato, F. Boccardi, M. Mezzavilla, S. Rangan and M. Zorzi, "Hybrid spectrum sharing in mmWave cellular networks," *IEEE Trans. on Cognitive Communications and Networking*, vol. 3, no. 2, pp. 155-168, 2017.

[10] L. Militano, D. Niyato, M. Condoluci, G. Araniti, A. Iera and G. M. Bisci, "Radio resource management for group-oriented services in LTE-Advanced," *IEEE Trans. Veh. Technol.*, vol. 64, no. 8, pp. 3725-3739, Aug. 2015.

[11] Q. Kuang and W. Utschick, "Energy management in heterogeneous networks with cell activation, user association and interference coordination," *IEEE Trans. Wireless Commun.*, vol. 15, pp. 3868-3879, 2016.

[12] Q. Kuang, W. Utschick, and A. Dotzler, "Optimal joint user association and multi-pattern resource allocation in heterogeneous networks," *IEEE Trans. Sig. Proc.*, vol. 64, pp. 3388-3401, 2016.

[13] A. L. Stolyar and H. Viswanathan, "Self-organizing dynamic fractional frequency reuse in OFDMA systems," *IEEE INFOCOM*, Phoenix, USA, Apr. 2008.

[14] R. Chang, Z. Tao, J. Zhang, and C.-C. Kuo, "Multicell OFDMA downlink resource allocation using a graphic framework," *IEEE Trans. Veh. Technol.*, vol. 58, pp. 3494-3507, Sep. 2009.

[15] S. H. Ali and V. C. M. Leung, "Dynamic frequency allocation in fractional frequency reused OFDMA networks," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 4286-4295, Aug. 2009.

[16] 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. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1479-1489, 2010.

[17] W.-C. Liao, M. Hong, Y.-F. Liu, and Z.-Q. Luo, "Base station activation and linear transceiver design for optimal resource management in heterogeneous networks," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3939-3952, 2014.

[18] A. Liu, V. Lau, L. Ruan, J. Chen, and D. Xiao, "Hierarchical radio resource optimization for heterogeneous networks with enhanced intercell interference coordination (eICIC)," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1684-1693, Apr. 2014.

[19] R. Etkin, A. Parekh and D. Tse, "Spectrum sharing for unlicensed bands," in *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 517-528, Apr. 2007.

[20] J. Huang, V. G. Subramanian, R. A. Agrawal, and R. Berry, "Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks," *IEEE J. Sel. Areas Commun.*, vol. 27, pp. 226-234, Feb. 2009.

[21] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 248-257, 2013.

[22] G. Lim, C. Xiong, L. J. Cimini Jr., and G. Y. Li, "Energy-efficient resource allocation for OFDMA-based multi-RAT networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2696-2705, 2014.

[23] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[24] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1100-1113, 2014.

[25] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets," *IEEE/ACM Trans. Netw.*, vol. 22, no. 1, pp. 137-150, 2014.

[26] L. P. Qian and Y. J. Zhang, "S-MAPEL: Monotonic optimization for non-convex joint power control and scheduling problems," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 1708–1719, May 2010.

[27] W. C. Ao and K. Psounis, "An efficient approximation algorithm for online multi-tier multi-cell user association," in *Proc. ACM MobiHoc*, Paderborn, Germany, 2016.

[28] H. Zhang, C. Jiang, N. C. Beaulieu, X. Chu, X. Wen, and M. Tao, "Resource allocation in spectrum-sharing OFDMA femtocells with heterogeneous services," *IEEE Trans. Commun.*, vol. 62, no. 7, pp. 2366-2377, 2014.

[29] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706-2716, 2013.

[30] P. Xue, P. Gong, J. H. Park, D. Park, and D. K. Kim, "Radio resource management with proportional rate constraint in the heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1066-1075, 2012.

[31] N. Saquib, E. Hossain, and D. I. Kim, "Fractional frequency reuse for interference management in LTE-advanced HetNets," *IEEE Wireless Commun.*, vol. 20, no. 2, pp. 113-122, 2013.

[32] Z. Zhou and D. Guo, "1000-Cell Global Spectrum Management," *ACM MobiHoc 2017*, Chennai, India, 2017.

[33] Z. Zhou, D. Guo, and M. L. Honig, "Licensed and unlicensed spectrum allocation in heterogeneous networks," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1815-1827, 2017.

[34] B. Zhuang, D. Guo, and M. L. Honig, "Traffic-driven spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun. Special Issue on Recent Advances in Heterogeneous Cellular Networks*, vol. 33, pp. 2027-2038, 2015.

[35] B. Zhuang, D. Guo, and M. L. Honig, "Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun. Special Issue on Energy-Efficient Techniques for 5G Wireless Communication Systems*, vol. 34, pp. 823-831, 2016.

[36] B. Zhuang, D. Guo, E. Wei, and M. L. Honig, "Scalable spectrum allocation and user association in networks with many small cells," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 2931-2942, 2017.

[37] B. Zhuang, D. Guo, E. Wei, and M. L. Honig, "Scalable spectrum allocation for large networks based on sparse optimization," in *Proc. IEEE ISIT*, Aachen, Germany, 2017.

[38] B. Zhuang, D. Guo, E. Wei, and M. L. Honig, "Scalable user association and resource allocation in dense heterogeneous networks using cardinality constrained optimization," to appear in *IEEE Trans. Signal Proc.*, 2018.

[39] H. Crowder, E. L. Johnson and M. Padberg, "Solving large-scale zero-one linear programming problems," *Operations Research*, vol. 31, no. 5, pp. 803–834, 1983.

[40] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *http://arxiv.org/abs/1602.08232*, 2016.

[41] F. Teng and D. Guo, "Resource management in 5G: A tale of two timescales," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2015.

[42] F. Teng and D. Guo, "Dual-timescale spectrum management in wireless heterogeneous networks," *http://arxiv.org/abs/1604.02781*, 2016.

[43] F. Liu, E. Bala, E. Erkip, M. C. Beluri and R. Yang, "Small-cell traffic balancing over licensed and unlicensed bands," in *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5850-5865, Dec. 2015.

[44] A. R. Elsherif, W. P. Chen, A. Ito, and Z. Ding, "Resource allocation and inter-cell interference management for dual-access small cells," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1082-1096, Jun. 2015.

[45] Q. Chen, G. Yu, R. Yin, A. Maaref, G. Y. Li, and A. Huang, "Energy efficiency optimization in licensed-assisted access," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, Apr. 2016.

[46] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201-220, 2005.

[47] D. Grandblaise, D. Bourse, K. Moessner, and P. Leaves, "Dynamic spectrum allocation and reconfigurability," in *Proc. Software-Defined Radio (SDR) Forum*, 2002.

[48] S.-S. Byun and J.-M. Gil, "Fair dynamic spectrum allocation using modified game theory for resource-constrained cognitive wireless sensor networks," *Symmetry*, vol. 9, no. 5, pp. 73, 2017.

[49] S. Wang, P. Xu, X. Xu, S. Tang, X. Li, and X. Liu, "Toda: Truthful online double auction for spectrum allocation in wireless networks," in *IEEE Symposium on New Frontiers in Dynamic Spectrum*, pp. 1–10, 2010.

[50] Z. Ji and K. R. Liu, "Cognitive radios for dynamic spectrum access-dynamic spectrum sharing: A game theoretical overview," *IEEE Communications Magazine*, vol. 45, no. 5, 2007.

[51] O. Ileri, D. Samardzija, and N. B. Mandayam, "Demand responsive pricing and competitive spectrum allocation via a spectrum server," in *IEEE Symposium on New Frontiers in Dynamic Spectrum*, pp. 194-202, 2005.

[52] M. Ru, S. Yin, and Z. Qu, "Power and spectrum allocation in D2D networks based on coloring and chaos genetic algorithm," *Procedia Computer Science*, vol. 107, pp. 183-189, 2017.

[53] L. Cao and H. Zheng, "Distributed spectrum allocation via local bargaining," in *SECON*, pp. 475-486, 2005.

[54] T. ElBatt and A. Ephremides, "Joint scheduling and power control for wireless ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 74-85, 2004.

[55] M. Kubisch, H. Karl, A. Wolisz, L. C. Zhong, and J. Rabaey, "Distributed algorithms for transmission power control in wireless sensor networks," in *WCNC*, vol. 1, pp. 558-563, 2003.

[56] T. ElBatt and A. Ephremides, "Joint scheduling and power control for wireless ad hoc networks," *IEEE Trans. on Wireless Commun.*, vol. 3, no. 1, pp. 74-85, 2004.

[57] C. U. Saraydar, N. B. Mandayam, and D. J. Goodman, "Efficient power control via pricing in wireless data networks," *IEEE Trans. Commun.*, vol. 50, no. 2, pp. 291-303, 2002.

[58] C. Godsil and G. Royle, *Algebraic Graph Theory*. New York: Springer-Verlag, 2001.

[59] N. Jindal and D. Breslin, "LTE and Wi-Fi in unlicensed spectrum: a coexistence study," *White paper*, 2015.

[60] C. Chen, R. Ratusuk and A. Ghosh,"Downlink performance analysis of LTE and WiFi coexistence in unlicensed bands with a simple listen-before-talk scheme," *IEEE Vehicular Conference (VTC Spring)*, Glasgow, UK, Apr. 2015.

[61] R. Ratasuk, M. A. Uusitalo, N. Mangalvedhe, A. Sorri, S. Iraji, C. Wijting, and A. Ghosh, "License-exempt LTE deployment in heterogeneous network," in *Proc. 9th International Symposium on Wireless Communication Systems*, Paris, France, Aug. 2012.

[62] F. S. Chaves, E. P. L. Almeida, R. D. Vieira, A. M. Cavalcante, F. M. Abinader Jr., S. Choudhury, and K. Doppler, "LTE UL power control for the improvement of LTE/Wi-Fi coexistence." *IEEE Vehicular Technology Conference*, pp. 1-6, 2013.

[63] E. P. L. Almeida, A. M. Cavalcante, R. C. D. Paiva, F. S. Chaves, F. M. Abinader Jr., R. D. Vieira, S. Choudhury, E. Tuomaala, and K. Doppler, "Enabling LTE/WiFi

coexistence by LTE blank subframe allocation," *IEEE International Communications Conference*, pp. 1-6, 2013.

[64] T. Nihtila, V. Tykhomyrov, O. Alanen, M. Uusitalo, A. Sorri, M. Moisio, S. Iraji, R. Ratasuk, and N. Mangalvedhe, "System performance of LTE and IEEE 802.11 coexisting on a shared frequency band," in *Proc. IEEE Wireless Communications and Networking Conf.*, Shanghai, China, pp. 1056-1061, Apr. 2013.

[65] N. Tian and Z. G. Zhang, *Vacation queueing models: theory and applications*. Springer, 2006.

[66] 3GPP TR 36.814, "Further advancements for E-UTRA physical layer aspects (Release 9)," V9.0.0, March 2010.

[67] CVX Research, "CVX: Matlab software for disciplined convex programming, version 2.0." http://cvxr.com/cvx, Aug. 2012.

[68] H. G. Eggleston, *Convexity*. No. 47, Cambridge University Press Archive, 1958.

[69] V. Gajic, J. Huang, and B. Rimoldi, "Competition of wireless providers for atomic users: Equilibrium and social optimality," in *Proc. Allerton Conf. Commun., Control, & Computing*, pp. 1203-1210, Monticello, IL, USA, 2009.

[70] J. Li and D. Guo, "Cloud-based resource allocation and cooperative transmission in large cellular networks," in *Proc. Allerton Conference*, 2017.

[71] K. Shen and W. Yu, "Fractional Programming for Communication Systems – Part I: Theory and Continuous Problems", *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616-2630, 2018.

[72] K. Shen and W. Yu, "Fractional Programming for Communication Systems – Part II: Discrete Problems", *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2631-2644, 2018.

[73] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Research Logistics Quarterly*, vol. 3, no. 1–2, pp. 95–110, 1956.

[74] J. M. Franks, "A (terse) Introduction to Lebesgue Integration." *American Mathematical Society*, 2009.

APPENDIX A

Proof

A.1. Proof of Proposition 2.1

To analyze the queueing model with vacation and nonexhaustive service, we follow a similar technique as in [**65**]. Denote the residual service time at time t by $R(t)$, and the waiting and service times for the i-th packet as W_i and X_i, respectively. We have,

$$E[X] = \frac{1}{r}. \tag{A.1}$$

Assuming the queue is stable, by the Pollaczek-Khinchine formula,

$$E[W] = \frac{E[R]}{1-\rho}, \tag{A.2}$$

and the average packet delay is given by:

$$E[T] = E[W] + E[X] = \frac{E[R]}{1-\rho} + \frac{1}{r}, \tag{A.3}$$

where $\rho = \frac{\lambda}{r}$. Denote $M(t)$ as the number of packets served up until time t. To calculate $E[R]$, we can see from Fig. A.1,

$$E[R] = \lim_{t\to\infty} \frac{1}{t} \int_0^t R(\tau) d\tau \tag{A.4}$$

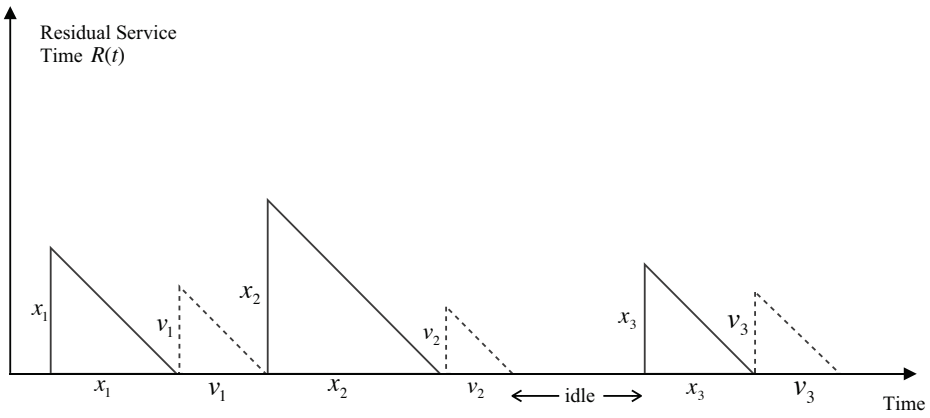


Figure A.1. Residual service time *R(t)*.

$$= \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{M(t)} \left(\frac{(X_i)^2}{2} + \frac{(V_i)^2}{2} \right) \tag{A.5}$$

$$= \lim_{t\to\infty} \frac{M(t)}{2t} \sum_{i=1}^{M(t)} \frac{(X_i)^2 + (V_i)^2}{M(t)} \tag{A.6}$$

$$= \frac{\lambda}{2} \left(E[X^2] + E[V^2] \right) \tag{A.7}$$

$$= \frac{\lambda}{r^2} + \frac{\lambda\nu}{2}. \tag{A.8}$$

The equalities hold in the almost sure sense. Substituting (A.8) into (A.3), we obtain (2.8).

To prove the convexity of (2.8), note that (2.8) can be written as the addition of two parts,

$$t = \frac{1}{(r-\lambda)^+} + \frac{\nu}{2} \frac{r\lambda}{(r-\lambda)^+}. \tag{A.9}$$

Since both parts are convex in the pair (r, λ) on $\mathbb{R}^2$, the sum is also convex in the pair (r, λ) on $\mathbb{R}^2$.

A.2. Proof of Theorem 2.1

We first reformulate P2.1 by a change of variables. The proof then follows a similar geometric argument as in [**34**].

Consider a reformulation of P2.1 by replacing constraints (P2.1c), (P2.1e), and (P2.1g) with the following three constraints, introducing a new collection of variables $(z_{i\to j}^{A,l})$:

$$r_j^{(l)} = \sum_{A\subset N}\left(\sum_{i\in N} s_{i\to j}^{A,l} z_{i\to j}^{A,l}\right) y^{A,l},\ j\in K, l\in\{1,2\} \tag{A.10}$$

$$\sum_{j\in K} z_{i\to j}^{A,l} \le 1, \qquad l\in\{1,2\}, i\in N, A\subset N \tag{A.11}$$

$$z_{i\to j}^{A,l}, y^{A,l} \ge 0, \qquad j\in K, l\in\{1,2\}, i\in N, A\subset N. \tag{A.12}$$

The new problem is equivalent to P2.1. This is because the feasible set for the rate tuple $\boldsymbol{r}$ remains the same, which is easy to see by regarding $z_{i\to j}^{A,l}$ as the fraction of spectrum under pattern A that AP i allocates to user device group j over RAT l. If one solves the new problem, the actual spectrum allocations can be recovered as $x_{i\to j}^{A,l} = y^{A,l} z_{i\to j}^{A,l}$.

We show that if $\boldsymbol{r}^*$ is an optimal rate tuple of P2.1, then we can attain each sub-tuple $\boldsymbol{r}^{*(l)}$ over RAT l with a k-sparse $\boldsymbol{y}^{(l)}$. For each RAT l, let us begin with a generally nonsparse optimal solution $\boldsymbol{y}^{(l)}$, whose support is $S^{(l)}$ ($y^{A,l}=0$ if $A\notin S^{(l)}$). Let us freeze the optimal $z_{i\to j}^{A,l}$ variables and define a k-vector $\boldsymbol{q}^{A,l}$ for every $A\in S^{(l)}$ with its elements determined by $q_j^{A,l} = \sum_{i\in N} s_{i\to j}^{A,l} z_{i\to j}^{A,l}$. According to (A.10), a convex combination of the vectors $(\boldsymbol{q}^{A,l})_{A\in S^{(l)}}$ with $(y^{A,l})_{A\in S_{(l)}}$ as coefficients form the optimal rate tuple $\boldsymbol{r}^{(l)}$ over RAT l. By Carathèodory's Theorem [**68**], $\boldsymbol{r}^{(l)}$ can be represented as a convex combination of at most $k+1$ of those vectors. Moreover, $\boldsymbol{r}^{(l)}$ must be on the boundary, not in the

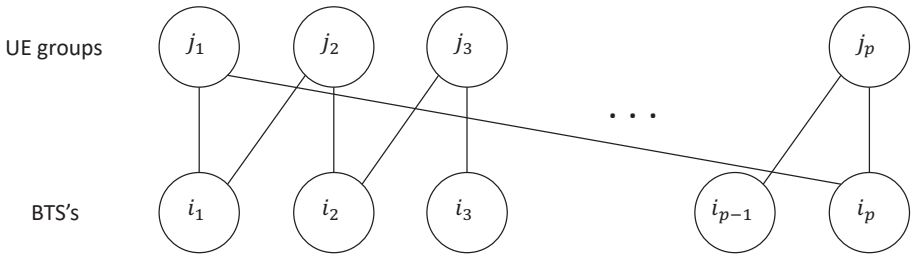


Figure A.2. A loop in the BGR for the user device groups served by multiple APs over the same RAT.

interior of the convex hull of $(\boldsymbol{q}^{A,l})$, because otherwise the rate tuple can be increased in all dimensions, contradicting the optimality assumption. Thus, for each RAT l, we can identify a vector $\boldsymbol{y}^{*(l)}$ whose support is a subset of $S^{(l)}$ with k or fewer elements, such that the optimal $\boldsymbol{r}^{(l)}$ is a convex combination of $(\boldsymbol{q}^{A,l})$ with $y^{*A,l}$ as coefficients.

A.3. Proof of Theorem 2.2

The proof follows an analogous proof in [**12**]. The KKT conditions for P2.1 for nonzero elements of $\boldsymbol{x}$ and $\boldsymbol{y}$ are :

$$\frac{\partial U}{\partial r_j^{(l)}} s_{i\to j}^{A,l} - \mu_i^{A,l} = 0, \tag{A.13}$$

$$\sum_{i\in N} \mu_i^{A,l} = \xi^{(l)}, \tag{A.14}$$

where $\mu_i^{A,l}$ and $\xi^{(l)}$ are the Lagrange multipliers for constraints (P2.1e) and (P2.1f), respectively. Assuming user device group j is served by two APs i_1 and i_2 over RAT l, define $s_{i\to j}^{(l)} = \sum_{A\subset N} s_{i\to j}^{A,l}$ and $\mu_i^{(l)} = \sum_{A\subset N} \mu_i^{A,l}$. According to (A.13) and (A.14), we have

$$\frac{s_{i_1\to j}^{(l)}}{s_{i_2\to j}^{(l)}} = \frac{\mu_{i_1}^{(l)}}{\mu_{i_2}^{(l)}}. \tag{A.15}$$

Following the argument in [**12**], a bipartite graph representation is used as in [**69**]. For each RAT, denote the user device groups served by multiple APs and the corresponding APs as nodes. An edge between a user device group and an AP represents an association. It remains to show that the graph contains no loop. Suppose there is a loop in the bipartite graph as shown in Fig. A.2. Then there exists a sequence of nodes of users $j_1, \ldots, j_p$ and a sequence of APs $i_1, \ldots, i_p$, where user j_q is connected with AP i_q for $q = 1, \ldots, p$, user j_{q+1} is connected with AP i_q for $q = 1, \ldots, p-1$, and user 1 is connected with AP p. The nodes are distinct otherwise we can find a smaller loop with this property. According to (A.15), the loop implies:

$$\frac{s^{(l)}_{i_p \to j_1}}{s^{(l)}_{i_1 \to j_1}} \frac{s^{(l)}_{i_1 \to j_2}}{s^{(l)}_{i_2 \to j_2}} \cdots \frac{s^{(l)}_{i_{p-1} \to j_p}}{s^{(l)}_{i_p \to j_p}} = \frac{\mu^{(l)}_{i_p}}{\mu^{(l)}_{i_1}} \frac{\mu^{(l)}_{i_1}}{\mu^{(l)}_{i_2}} \cdots \frac{\mu^{(l)}_{i_{p-1}}}{\mu^{(l)}_{i_p}} = 1.$$

Since $s^{(l)}_{i \to j}$ is the sum of spectral efficiencies over all patterns over RAT l, it is a random variable based on random topology. Therefore, $\frac{s^{(l)}_{i_p \to j_1}}{s^{(l)}_{i_1 \to j_1}} \frac{s^{(l)}_{i_1 \to j_2}}{s^{(l)}_{i_2 \to j_2}} \cdots \frac{s^{(l)}_{i_{p-1} \to j_p}}{s^{(l)}_{i_p \to j_p}} = 1$ is a zero probability event, which shows that w.p.1 there is no loop in the graph. Since there are n APs, the largest possible BGR without a loop has $n-1$ user nodes, which proves Theorem 2.2.

A.4. Proof of Proposition 2.2

Taking the second derivatives of $\lambda^{(1)}_j \hat{t}^{(1)}_j$ with respect to $\lambda^{(1)}_j$ and $r^{(1)}_j$, we have:

$$\frac{\partial^2 \lambda^{(1)}_j \hat{t}^{(1)}_j}{\partial {\lambda^{(1)}_j}^2} = \frac{\eta r^{(1)}_j}{(r^{(1)}_j - \beta \lambda^{(1)}_j)^3}, \tag{A.16}$$

$$\frac{\partial^2 \lambda^{(1)}_j \hat{t}^{(1)}_j}{\partial {r^{(1)}_j}^2} = \frac{\eta \lambda^{(1)}_j}{\beta} \left(\frac{1}{(r^{(1)}_j - \beta \lambda^{(1)}_j)^3} - \frac{1}{\left(r^{(1)}_j\right)^3} \right) + \frac{2\beta \lambda^{(1)}_j}{\left(r^{(1)}_j\right)^3}. \tag{A.17}$$

When $\frac{1}{\lambda_j^{(1)}} > \frac{\beta}{r_j^{(1)}}$, in other words, when the queue is stable, the derivatives are positive, which means $\lambda_j^{(1)}\hat{t}_j^{(1)}$ is bi-convex in $\lambda_j^{(1)}$ and $r_j^{(1)}$. In addition, since $\lambda_j^{(2)}\hat{t}_j^{(2)}$ is similar to $\lambda_j^{(1)}\hat{t}_j^{(1)}$ except for an additional term $\frac{\nu_j(\lambda_j^{(2)})^2 r_j^{(2)}}{2(r_j^{(2)}-\beta\lambda_j^{(2)})^+}$, which is also bi-convex in $\lambda_j^{(2)}$ and $r_j^{(2)}$, $\lambda_j^{(2)}\hat{t}_j^{(2)}$ is also bi-convex in $\lambda_j^{(2)}$ and $r_j^{(2)}$. Since $\hat{U}$ is a linear combination of bi-convex functions, we conclude that (2.20) is bi-convex in $\boldsymbol{\lambda}$ and $\boldsymbol{r}$.

A.5. Proof of Proposition 3.1

The affine utility function can be written as

$$u(\boldsymbol{r}(\boldsymbol{w})) = d + \sum_{j\in K}\sum_{A\subset N}\sum_{i\in A} c_{i\to j}^A w_{i\to j}^A \tag{A.18}$$

for some constants d and $(c_{i\to j}^A)$.

Then P3.1 can be rewritten as:

$$\begin{array}{l} \underset{y^A\geq 0,\ \sum_{A\subset N} y^A = 1}{\text{maximize}} \quad \underset{w_{i\to j}^A \geq 0,\ \sum_{l\in K} w_{i\to l}^A \leq y^A \ \forall j\in K, \forall A\subset N, \forall i\in A}{\text{maximize}} \\ \sum_{j\in K}\sum_{A\subset N}\sum_{i\in A} c_{i\to j}^A w_{i\to j}^A + d. \end{array} \tag{A.19}$$

Define $j^*(i,A)\in K$ as a maximizer of $c_{i\to j}^A$. It is easy to see that the solution to the inner problem in (A.19) is to let each AP serve the single user device with the largest weight for each pattern, i.e.,

$$w_{i\to j}^A = y^A \mathbb{1}(j = j^*(i,A)) \tag{A.20}$$

for all $A \subset N$ and $i \in A$, where $\mathbb{1}(\cdot)$ is the general indicator function. Then (A.19) can be written as:

$$\underset{y^A \geq 0, \sum_{A \subset N} y^A = 1}{\text{maximize}} \sum_{A \subset N} \sum_{i \in A} c^A_{i \to j^*(i,A)} y^A. \tag{A.21}$$

Again, (A.21) can be solved by allocating all the resources to one pattern that has the largest weight, i.e., letting $y^{A^*} = 1$ where A^* maximizes $\sum_{i \in A} c^A_{i \to j^*(i,A)}$.

A.6. Proof of Theorem 3.2

To prove Theorem 3.2, we shall introduce two additional equivalent optimization problems as bridges between P3.1 and P3.2.

Lemma A.1. *P3.1 is equivalent to PA.1:*

$$\underset{\boldsymbol{r},\boldsymbol{w},\boldsymbol{y},\boldsymbol{h}}{\text{maximize}}\ u(\boldsymbol{r}) \tag{PA.1a}$$

$$\text{subject to } r_j = \sum_{A \subset N} \sum_{i \in A} s^A_{i \to j} \sum_{l \in L} w^{A,l}_{i \to j}, \quad \forall j \in K \tag{PA.1b}$$

$$\sum_{j \in K} w^{A,l}_{i \to j} \leq y^{A,l}, \qquad \forall A \subset N, \forall i \in A, \forall l \in L \tag{PA.1c}$$

$$\sum_{A \subset N} y^{A,l} \leq h^l, \qquad \forall l \in L \tag{PA.1d}$$

$$\sum_{A \subset N} |y^{A,l}|_0 \leq 1, \qquad \forall l \in L \tag{PA.1e}$$

$$\sum_{l \in L} h^l \leq 1, \tag{PA.1f}$$

$$w^{A,l}_{i \to j} \geq 0, \quad \forall l \in L, \forall j \in K, \forall A \subset N, \forall i \in A. \tag{PA.1g}$$

Proof. We first show that P3.1 is equivalent to PA.1 with constraint (PA.1e) removed. To see this, we recognize that the latter problem basically splits the variables in the former into k constituents in identical form. The equivalence is then due to the concavity of the utility function. To be precise, without (PA.1e), if all variables with subscript l is set to 0 except for $l = 1$, PA.1 reduces to P3.1. Thus PA.1 is a relaxation to P3.1. On the other hand, from any solution to PA.1 without constraint (PA.1e), we can combine the variables of $l = 1, \cdots, k$ to one feasible solution of P3.1. Hence the equivalence.

It remains to show that the additional l_0 constraint (PA.1e) does not change the optimal solution. Recall that P3.1 has an optimal solution that activates at most k patterns by Theorem 3.1. If we let the k active patterns each correspond to a distinct subscript l in PA.1, we obtain a feasible solution to PA.1 that yields the same utility. Specifically, suppose the k active patterns found for P3.1 are $A^1, \cdots, A^k \subset N$, and the optimal $\boldsymbol{w}$ and $\boldsymbol{y}$ variables are $(w^A_{i\to j})_{j\in K, A\subset N, i\in A}$ and $(y^A)_{A\subset N}$. Then the variables of PA.1 are constructed as follows:

$$h^l = y^{A^l} \tag{A.22}$$

$$y^{A,l} = y^{A^l} \mathbb{1}(A = A^l) \tag{A.23}$$

$$w^{A,l}_{i\to j} = w^{A^l}_{i\to j} \mathbb{1}(A = A^l) \tag{A.24}$$

for $l = 1, \cdots, k$. Then it is easy to see that all constraints in PA.1 are satisfied and the same optimal utility is achieved. □

Lemma A.2. *PA.1 is equivalent to PA.2:*

$$\underset{\boldsymbol{r},\boldsymbol{w},\boldsymbol{z},\boldsymbol{h}}{\text{maximize}}\ u(\boldsymbol{r}) \tag{PA.2a}$$

$$\text{subject to } r_j = \sum_{A\subset N}\sum_{i\in A} s_{i\to j}^{A}\sum_{l\in L} w_{i\to j}^{A,l}, \quad \forall j\in K \tag{PA.2b}$$

$$w_{i\to j}^{A,l} \le z_j^{A,l} \quad \forall A\subset N, \forall i\in A, \forall l\in L, \forall j\in K \tag{PA.2c}$$

$$z_j^{A,l} + \sum_{B\subset N: B\neq A} z_m^{B,l} \le 1, \tag{PA.2d}$$

$$\forall A\subset N, \forall l\in L, \forall j,m\in K \tag{PA.2e}$$

$$\sum_{j\in K}\sum_{A\subset N} w_{i\to j}^{A,l} \le h^l, \quad \forall i\in N, \forall l\in L \tag{PA.2f}$$

$$\sum_{l\in L} h^l \le 1, \tag{PA.2g}$$

$$z_j^{A,l}\in\{0,1\}, \quad \forall l\in L, \forall j\in K, \forall A\subset N \tag{PA.2h}$$

$$w_{i\to j}^{A,l}\ge 0, \quad \forall l\in L, \forall j\in K, \forall A\subset N, \forall i\in A. \tag{PA.2i}$$

Proof. We first note that the utility functions of PA.1 and PA.2 are identical. Also, constraints (PA.1b), (PA.1f), (PA.1g) are identical to constraints (PA.2b), (PA.2g), (PA.2i). Next, we will prove that every maximum of PA.1 is also a maximum of PA.2.

Suppose $(\boldsymbol{r}^*, \boldsymbol{w}^*, \boldsymbol{y}^*, \boldsymbol{h}^*)$ is a maximum of PA.1. We seek $\boldsymbol{z}^*$ such that $(\boldsymbol{r}^*, \boldsymbol{w}^*, \boldsymbol{z}^*, \boldsymbol{h}^*)$ is feasible for PA.2. Fix $l\in L$. Constraint (PA.1e) dictates that there is at most one active global pattern for every $l\in L$. Namely, we can identify one $A_l^*\subset N$, such that

${y^*}^{B,l} = 0$ for every $B \neq A_l^*$. From constraints (PA.1c), (PA.1d), we have

$$\sum_{j \in K} {w^*}_{i \to j}^{A_l^*, l} \leq {h^*}^l, \tag{A.25}$$

$${w^*}_{i \to j}^{B,l} = 0, \ \forall B \neq A_l^*. \tag{A.26}$$

For every i, j, l, and A, let ${z^*}_j^{A,l} = 0$ if ${w^*}_{i \to j}^{A,l} = 0$ and ${z^*}_j^{A,l} = 1$ otherwise. Then by (A.25) and (A.26), we have

$${z^*}_j^{A_l^*, l} \leq 1, \tag{A.27}$$

$${z^*}_j^{B,l} = 0, \ \forall B \neq A_l^*. \tag{A.28}$$

Then it is obvious that these variables satisfy constraints (PA.2c), (PA.2d), and (PA.2h). For the remaining (PA.2f), we have

$$\sum_{j \in K} \sum_{A \subset N} {w^*}_{i \to j}^{A,l} = \sum_{j \in K} {w^*}_{i \to j}^{A_l^*, l} \leq {h^*}^l. \tag{A.29}$$

Therefore, $(\boldsymbol{r}^*, \boldsymbol{w}^*, \boldsymbol{z}^*, \boldsymbol{h}^*)$ is feasible for PA.2.

To show the converse, we show that if $(\boldsymbol{r}^*, \boldsymbol{w}^*, \boldsymbol{z}^*, \boldsymbol{h}^*)$ is a maximum of PA.2, then there exists $\boldsymbol{y}^*$ such that $(\boldsymbol{r}^*, \boldsymbol{w}^*, \boldsymbol{y}^*, \boldsymbol{h}^*)$ is feasible for PA.1. Fix $l \in L$. Constraints (PA.2d) and (PA.2h) dictate that there is at most one active global pattern for all $j \in K$. Namely, we can identify one $A_l^* \subset N$, such that ${z^*}^{B,l} = 0$ for every $B \neq A_l^*$. From constraints (PA.2c), (PA.2f) and (PA.2i), we have

$${w^*}_{i \to j}^{B,l} = 0, \ \forall B \neq A_l^* \tag{A.30}$$

$$\sum_{j\in K}\sum_{A\subset N} {w^*}^{A,l}_{i\to j} = \sum_{j\in K} {w^*}^{A^*_l,l}_{i\to j} \leq {h^*}^l. \tag{A.31}$$

We let $\boldsymbol{y}^*$ be defined as ${y^*}^{A,l} = \sum_{j\in K} {w^*}^{A,l}_{i\to j}$, then we have

$${y^*}^{B,l} = 0,\ \forall B \neq A^*_l \tag{A.32}$$

$${y^*}^{A^*_l,l} = \sum_{j\in K} {w^*}^{A^*_l,l}_{i\to j}. \tag{A.33}$$

By (A.32) and (A.33), it is obvious that

$$\sum_{j\in K} {w^*}^{A,l}_{i\to j} \leq {y^*}^{A,l}. \tag{A.34}$$

In addition, we have

$$\sum_{A\subset N} {y^*}^{A,l} = {y^*}^{A^*_l,l} \leq {h^*}^l, \tag{A.35}$$

and

$$\sum_{A\subset N} |{y^*}^{A,l}|_0 = |{y^*}^{A^*_l,l}|_0 \leq 1. \tag{A.36}$$

Therefore, these variables satisfy constraints (PA.1c), (PA.1d), and (PA.1e). Hence $(\boldsymbol{r}^*, \boldsymbol{w}^*, \boldsymbol{y}^*, \boldsymbol{h}^*)$ is also feasible for PA.1. We conclude that every maximum of PA.1 corresponds to a maximum of PA.2, and vice versa. Hence the equivalence of PA.1 and PA.2. □

Lemma A.3. *PA.2 is equivalent to P3.2.*

Proof. The difference between P3.2 and PA.2 are entirely in the $(\boldsymbol{w}, \boldsymbol{z})$ variables associated with global patterns and $(\boldsymbol{x}, \boldsymbol{d})$ variables associated with local patterns. The utilities (PA.2a) and (P3.2a) are identical. Constraints (PA.2g) and (P3.2f) are identical. We next relate the global variables $(\boldsymbol{w}, \boldsymbol{z})$ to the local variables $(\boldsymbol{x}, \boldsymbol{d})$, so that feasibility of P3.2 and feasibility of PA.2 imply each other.

We first show that if $(\boldsymbol{r}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{h})$ satisfy all constraints of PA.2, then there exist $(\boldsymbol{x}, \boldsymbol{d})$ such that $(\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{d}, \boldsymbol{h})$ satisfy all constraints of P3.2. Let $\boldsymbol{x}$ and $\boldsymbol{d}$ variables be obtained as

$$x_{i\to j}^{A,l} = \sum_{C\subset N: C\cap N_j = A} w_{i\to j}^{C,l}, \qquad \forall j \in K, l \in L, A \in N_j, i \in A \tag{A.37}$$

$$d_j^{A,l} = \sum_{C\subset N: C\cap N_j = A} z_j^{C,l}, \qquad \forall j \in K, l \in L, A \in N_j. \tag{A.38}$$

By (PA.2c), (PA.2h) and (PA.2i), it is obvious that (P3.2c), (P3.2g) and (P3.2h) hold. For every user device $j \in K$, every local pattern $A \in N_j$, and every global pattern $C \subset N$ that satisfies $C \cap N_j = A$, we have $s_{i\to j}^{C} = s_{i\to j}^{A}$ by (3.12). From (PA.2b), (3.12), and (A.37) for every $j \in K$,

$$r_j = \sum_{C\subset N} \sum_{i\in C} s_{i\to j}^{C} \sum_{l\in L} w_{i\to j}^{C,l} \tag{A.39}$$

$$= \sum_{A\subset N_j} \sum_{i\in A} s_{i\to j}^{A} \sum_{l\in L} \left(\sum_{C\subset N: C\cap N_j = A} w_{i\to j}^{C,l} \right) \tag{A.40}$$

$$= \sum_{A\subset N_j} \sum_{i\in A} s_{i\to j}^{A} \sum_{l\in L} x_{i\to j}^{A,l} \tag{A.41}$$

which is (P3.2b). Moreover, fix $l \in L$. Constraints (PA.2d) and (PA.2h) dictates that there is at most one active global pattern for all $j \in L$. Namely, we can identify one $A_l \subset N$, such that $z_j^{B,l} = 0$ for every $j \in K$ and $B \neq A_l$.

We next examine inequality (P3.2d) where $\boldsymbol{d}$ is defined by (A.38). For every j, l, m, A, if $A_l \cap N_j = A$, then

$$\begin{aligned} & d_j^{A,l} + \sum_{B \subset N_m : B \cap N_j \neq A \cap N_m} d_m^{B,l} \\ & = z_j^{A_l,l} + \sum_{D \subset N : D \cap N_m \cap N_j \neq A_l \cap N_m \cap N_j} z_m^{D,l} \qquad \text{(A.42)} \\ & \leq z_j^{A_l,l} + \sum_{D \subset N : D \neq A_l} z_m^{D,l} \leq 1 \qquad \text{(A.43)} \end{aligned}$$

where (A.43) is due to (PA.2d).

If $A_l \cap N_j \neq A$, then

$$\begin{aligned} & d_j^{A,l} + \sum_{B \subset N_m : B \cap N_j \neq A \cap N_m} d_m^{B,l} \\ & = \sum_{C \subset N : C \cap N_j = A} z_j^{C,l} + \sum_{B \subset N_m : B \cap N_j \neq A \cap N_m} \sum_{D \subset N : D \cap N_m = B} z_m^{D,l} \qquad \text{(A.44)} \\ & \leq 0 + \sum_{D \subset N} z_m^{D,l} \leq 1 \qquad \text{(A.45)} \end{aligned}$$

where (A.45) is due to the special case of (PA.2d) with $j = m$. Therefore (P3.2d) is established. It remains to show (P3.2e). By definition (A.37),

$$\sum_{j \in K} \sum_{A \subset N_j} x_{i \to j}^{A,l} = \sum_{j \in K} \sum_{A \subset N_j} \sum_{C \subset N : C \cap N_j = A} w_{i \to j}^{C,l} \qquad \text{(A.46)}$$

$$\leq \sum_{j \in K} \sum_{C \subset N} w_{i \to j}^{C,l} \leq h^l \tag{A.47}$$

where (A.47) is due to (PA.2f).

Thus $(\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{d}, \boldsymbol{h})$ satisfy all constraints (P3.2b)–(P3.2h) as long as $(\boldsymbol{r}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{h})$ satisfy constraints (PA.2b)–(PA.2i).

We next show that if $(\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{d}, \boldsymbol{h})$ satisfy all constraints of P3.2, then there exists $(\boldsymbol{w}, \boldsymbol{z})$ such that $(\boldsymbol{r}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{h})$ satisfy all constraints of PA.2. The key is to reconstruct global variables $(\boldsymbol{w}, \boldsymbol{z})$ from local variables $(\boldsymbol{x}, \boldsymbol{d})$. Fix $l \in L$. Constraints (P3.2d) and (P3.2g) dictate that there is at most one active local pattern in every neighborhood. Namely, for every $j \in K$, we can identify one $B_j^l \subset N$, such that $d_j^{B,l} = 0$ for every $B \neq B_j^l$. Let us define a global pattern:

$$A_l = \bigcup_{j \in K} B_j^l. \tag{A.48}$$

Due to (P3.2d), we have $A_l \cap N_j = B_j^l$. Define global variables:

$$w_{i \to j}^{C,l} = x_{i \to j}^{B_j^l, l} \mathbb{1}(C = A_l) \tag{A.49}$$

$$z_j^{C,l} = d_j^{B_j^l, l} \mathbb{1}(C = A_l). \tag{A.50}$$

Then (PA.2h) and (PA.2i) are trivial. Moreover,

$$r_j = \sum_{A \subset N_j} \sum_{i \in A} s_{i \to j}^A \sum_{l \in L} x_{i \to j}^{A,l} \tag{A.51}$$

$$= \sum_{i \in N_j} \sum_{l \in L} s_{i \to j}^{B_j^l} x_{i \to j}^{B_j^l, l} \tag{A.52}$$

$$= \sum_{i \in N} \sum_{l \in L} s_{i \to j}^{A_l} w_{i \to j}^{A_l, l} \tag{A.53}$$

$$= \sum_{A \subset N} \sum_{i \in A} s_{i \to j}^{A} \sum_{l \in L} w_{i \to j}^{A, l}, \tag{A.54}$$

where (A.53) is due to (A.49). Therefore, (PA.2b) is established. (PA.2c) is established from (P3.2c), (A.49) and (A.50). In addition, (PA.2d) is established due to (P3.2d) and (A.50). Finally, for (PA.2f), we have

$$\sum_{j \in K} \sum_{A \subset N} w_{i \to j}^{A, l} = \sum_{j \in K} x_{i \to j}^{B_j^l, l} \tag{A.55}$$

$$\leq \sum_{j \in K} \sum_{A \subset N_j} x_{i \to j}^{A, l} \leq h^l \tag{A.56}$$

where (A.56) is due to (P3.2e).

In all, the utility and constraints of PA.2 are equivalent to those of P3.2. Hence the equivalence of the two optimization problems. □

From Lemma A.1, Lemma A.2 and Lemma A.3, we can conclude that P3.1 and P3.2 are equivalent. Hence Theorem 3.2 is proved.

A.7. Proof of Theorem 4.2

The PSDs in P4.8 are arbitrary (Lebesgue measurable) functions in an infinite dimensional function space. They determine the utility function through the rate and power vectors ($\boldsymbol{r}$ and $\boldsymbol{P}$) of finite dimensions. The key to the proof is to invoke the Carathéodory's theorem [**68**] to assert the existence of a $(k+n+1)$-dimensional allocation that achieves the desired $(\boldsymbol{r}, \boldsymbol{P})$.

Let $\mathcal{L}$ denote the set of Lebesgue integrable functions on $[0, W]$. Let us define:

$$
\begin{aligned}
R = \Bigg\{ & (\boldsymbol{r}, \boldsymbol{P}, \boldsymbol{t}) \in \mathbb{R}^k \times \mathbb{R}^n \times \mathbb{R}^n : \exists\ \boldsymbol{p}(\cdot) = (p_{i\to j}(\cdot)) \in \mathcal{L}^{nk} \\
& \text{s.t. } r_j = \sum_{i\in N} \int_0^W s_{i\to j}(\boldsymbol{p}(f))\,\mathrm{d}f,\ \forall j \in K, \\
& P_i = \sum_{j\in K} \int_0^W p_{i\to j}(f)\,\mathrm{d}f,\ \forall i \in N, \\
& t_i = \int_0^W \left[\sum_{j\in K} p_{i\to j}(f)\right]^{\alpha} \mathrm{d}f,\ \forall i \in N, \\
& t_i \le (P_{\max})^{\alpha},\ \forall i \in N, \\
& \sum_{j\in K} |p_{i\to j}(f)|_0 \le d_i,\ \forall i \in N, f \in [0, W], \\
& 0 \le p_{i\to j}(f) \le Q,\ \forall i \in N, j \in K, f \in [0, W] \Bigg\}.
\end{aligned}
\tag{A.57}
$$

We adopt the Lebesgue integral throughout as a convenient justification of using "maximize" in P4.1 and P4.8, i.e., that the maximum is actually achieved. While it is perhaps possible to adopt the less technical Riemann integral to the same outcome, it is not as easy to work with. In particular, the limit of a sequence of Riemann integrable functions may not be Riemann integrable at all.

In the special case P4.1, as long as $s_{i\to j}(\boldsymbol{p}(\cdot))$ is measurable, $\boldsymbol{p}$ does not have to be measurable. Nonetheless, we can restrict to measurable ones without loss of generality because any feasible rate vector can be constructed by a measurable $\boldsymbol{p}$.

We can think of R as a manifold in the $(k+2n)$-dimensional Euclidean space induced by (continuous) power allocations $(p_{i\to j}(f))_{i\in N, j\in K, f\in[0,W]}$ that satisfy some power constraints.

R is nonempty because it includes the vector induced by the all-zero power allocation. For every $(\boldsymbol{r}, \boldsymbol{P}, \boldsymbol{t}) \in R$, $(\boldsymbol{r}, \boldsymbol{P})$ are a pair of feasible rate and power vectors. The reason to also include the vector $\boldsymbol{t}$ becomes clear later.

Define another set

$$
\begin{aligned}
S = \Bigg\{ & (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \in \mathbb{R}^k \times \mathbb{R}^n \times \mathbb{R}^n : \exists\ \boldsymbol{q} \in [0, Q]^{nk} \\
& \text{s.t. } u_j = W \sum_{i \in N} s_{i \to j}(\boldsymbol{q}),\ \forall j \in K, \\
& v_i = W \sum_{j \in K} q_{i \to j},\ \forall i \in N, \\
& w_i = W \left[\sum_{j \in K} q_{i \to j} \right]^{\alpha},\ \forall i \in N, \\
& \sum_{j \in K} |q_{i \to j}|_0 \leq d_i, \forall i \in N \Bigg\}.
\end{aligned}
\tag{A.58}
$$

The set S is bounded, closed, and therefore compact. S is also a manifold in the $(k+2n)$-dimensional space. Every vector in S is induced by a single power profile $\boldsymbol{q} \in [0, Q]^{nk}$, which can be achieved by the flat power allocation over the entire bandwidth defined according to $p_{i \to j}(f) = q_{i \to j},\ \forall i \in N, j \in K, f \in [0, W]$. If indeed this single power profile

is adopted over $[0, W]$, we have according to (A.58):

$$\begin{aligned} r_j &= \sum_{i\in N} \int_0^W s_{i\to j}(\boldsymbol{p}(f))\,\mathrm{d}f = W \sum_{i\in N} s_{i\to j}(\boldsymbol{q}) = u_j, \\ P_i &= \sum_{j\in K} \int_0^W p_{i\to j}(f)\,\mathrm{d}f = W \sum_{j\in K} q_{i\to j} = v_i, \\ &\text{and} \\ t_i &= \int_0^W \left[\sum_{j\in K} p_{i\to j}(f)\right]^\alpha \mathrm{d}f = W \left[\sum_{j\in K} q_{i\to j}\right]^\alpha = w_i \end{aligned} \tag{A.59}$$

for all $i \in N$ and $j \in K$.

The remainder of the proof includes three technical steps: We first show that R is a subset of the convex hull of S, i.e., $R \subset \mathrm{conv}(S)$. We then invoke the Carathéodory's theorem to show that every point in R is a convex combination of $(k + 2n + 1)$ points in S, and is hence achieved using a $(k+2n+1)$-piecewise constant power allocation. Finally, we reduce the number of sub-bands needed from $k + 2n + 1$ to $k + n + 1$.

Step 1: The goal of this step is to show $R \subset \mathrm{conv}(S)$. Because S is compact, $\mathrm{conv}(S)$ must be closed. Given $(\boldsymbol{r}, \boldsymbol{P}, \boldsymbol{t}) \in R$, if we can explicitly define a sequence $\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots \in \mathrm{conv}(S)$, which converges to $(\boldsymbol{r}, \boldsymbol{P}, \boldsymbol{t})$, then $(\boldsymbol{r}, \boldsymbol{P}, \boldsymbol{t}) \in \mathrm{conv}(S)$. This part of the proof hinges on the simple function approximation of Lebesgue measurable functions.[1]

Specifically, for every $(\boldsymbol{r}, \boldsymbol{P}, \boldsymbol{t}) \in R$, there exists a power allocation $\boldsymbol{p}(\cdot)$ that lies in the feasible region of P4.8, which results in the rate vecor $\boldsymbol{r}$ and tranmist power vector $\boldsymbol{P}$ as well as $\boldsymbol{t}$ as defined in (A.57). Since the PSD of each AP is a bounded measurable function over $[0, W]$, it can be arbitrarily closely approximated by a sequence of *simple*

[1]This approach does not apply to Riemann integrable functions in general.

functions [**74**]. Fix a positive integer l, and partition the kn-dimensional cube $[0, \sqrt{l})^{kn}$ into l^{kn} disjoint cubes $\{C^m\}_{m=1}^{l^{kn}}$ with equal side length $\frac{1}{\sqrt{l}}$ on each dimension. Each dimension of the cube is indexed by a pair (i, j) with $i \in N$ and $j \in K$. Let the edge of partition C^m in the (i, j)-th dimension be represented by $[a^m_{i\to j}, a^m_{i\to j} + \frac{1}{\sqrt{l}})$. Then $C^m = \prod_{i\in N}\prod_{j\in K}[a^m_{i\to j}, a^m_{i\to j} + \frac{1}{\sqrt{l}})$. Since $\boldsymbol{p}$ is a measurable function, we can define measurable partitions of $[0, W]$ by $A^m = \boldsymbol{p}^{-1}(C^m),\ m = 1, \cdots, l^{kn}$ and $A^0 = [0, W]\backslash \bigcup_{m=1}^{l^{kn}} A_m$. A^m describes the spectrum on which the power profiles find their values in partition C^m. We define a simple function $\boldsymbol{p}^l : [0, W] \to [0, \sqrt{l}]^{kn}$ as

$$p^l_{i\to j}(f) = \sum_{m=1}^{l^{kn}} a^m_{i\to j} \mathbb{1}_{\{f\in A^m\}}. \tag{A.60}$$

It is obvious that $p^l_{i\to j}(f) \le p_{i\to j}(f)$ for all $f \in [0, W], i \in N, j \in K$. As $l \to \infty$, this sequence of simple functions $(p^l_{i\to j})$ converges to $(p_{i\to j})$ from below almost everywhere. Moreover, let $\boldsymbol{r}^l = (r^l_j)_{j\in K}$ where $r^l_j = \sum_{m=1}^{l^{kn}} \sum_{i\in N} S_{i\to j}(a^m_{i\to j})\mu(A^m)$,

let $\boldsymbol{P}^l = (P^l_i)_{i\in N}$ where $P^l_i = \sum_{m=1}^{l^{kn}} \sum_{j\in K} a^m_{i\to j}\mu(A^m)$,

and let $\boldsymbol{t}^l = (t^l_i)_{i\in N}$ where $t^l_i = \sum_{m=1}^{l^{kn}} [\sum_{j\in K} a^m_{i\to j}]^\alpha \mu(A^m)$,

where $\mu(\cdot)$ is the Lebesgue measure in $\mathbb{R}^{kn}$. It is clear that $\boldsymbol{r}^l \le \boldsymbol{r}$ and $\boldsymbol{P}^l \le \boldsymbol{P}$. And we also have

$$t^l_i = \int_0^W \left[\sum_{j\in K} p^l_{i\to j}(f)\right]^\alpha \mathrm{d}f \tag{A.61}$$

$$\le \int_0^W \left[\sum_{j\in K} p_{i\to j}(f)\right]^\alpha \mathrm{d}f \tag{A.62}$$

$$\le (P_{\max})^\alpha, \quad \forall i \in N. \tag{A.63}$$

Therefore, $(\boldsymbol{r}^l, \boldsymbol{P}^l, \boldsymbol{t}^l) \in \text{conv}(S)$ and $(\boldsymbol{r}^l, \boldsymbol{P}^l, \boldsymbol{t}^l) \leq (\boldsymbol{r}, \boldsymbol{P}, \boldsymbol{t})$. As l goes to ∞, the number of cubes in which $[0, Q]^{kn}$ is partitioned goes to ∞, $(\boldsymbol{r}^l, \boldsymbol{P}^l, \boldsymbol{t}^l)$ converges to $(\boldsymbol{r}, \boldsymbol{P}, \boldsymbol{t})$. Therefore, $R \subset \text{conv}(S)$.

Step 2: By Carathéodory's theorem [**68**], $(\boldsymbol{r}, \boldsymbol{P}, \boldsymbol{t})$ can be constructed as a convex combination of $k + 2n + 1$ vectors in S.[2] That is, for any feasible solution $(\boldsymbol{r}, \boldsymbol{P}, \boldsymbol{t})$ of P4.8, it can be rewritten as a convex combination of $k + 2n + 1$ points in S. That is

$$(\boldsymbol{r}, \boldsymbol{P}, \boldsymbol{t}) = \sum_{m=1}^{k+2n+1} \beta^m (\boldsymbol{u}^m, \boldsymbol{v}^m, \boldsymbol{w}^m), \tag{A.64}$$

where $(\boldsymbol{u}^m, \boldsymbol{v}^m, \boldsymbol{w}^m) \in S$, $\beta^m \geq 0$, $m = 1, \cdots, k+2n+1$, and $\sum_{m=1}^{k+2n+1} \beta^m = 1$. Specifically, once the $k + 2n + 1$ power profiles and their weights are known, the power allocation $\boldsymbol{p}$ can be constructed as

$$p_{i\to j}(f) = q_{i\to j}^l, \text{ if } f \in \left[\sum_{m=1}^{l-1} \beta^m W, \sum_{m=1}^{l} \beta^m W\right] \text{ for } l = 1, \cdots, k + 2n + 1. \tag{A.65}$$

As in (A.59), it is straightforward to verify that this $(k+2n+1)$-piecewise constant power allocation is feasible and achieves the optimal utility. Therefore, any feasible solution to P4.8 can be attained with a $(k + 2n + 1)$-piecewise constant power allocation.

Step 3: We next reduce the number of sub-bands needed from $k+2n+1$ to $k+n+1$.

[2]Carathéodory's theorem has been invoked in the past to establish similar results (e.g., [**19**] and [**38**]), but this is the first analysis that rigorously examine all conditions.

For any feasible solution $(\boldsymbol{r}, \boldsymbol{P})$ of P4.8, as we just proved, we can rewrite it as a convex combination of $k+2n+1$ points in S. That is

$$(\boldsymbol{r}, \boldsymbol{P}) = \sum_{m=1}^{k+2n+1} \beta^m (\boldsymbol{u}^m, \boldsymbol{v}^m), \tag{A.66}$$

where $(\boldsymbol{u}^m, \boldsymbol{v}^m) \in S$, $m = 1, \cdots, k+2n+1$. Moreover, we have

$$\sum_{m=1}^{k+2n+1} \beta^m = 1, \tag{A.67}$$

$$\sum_{m=1}^{k+2n+1} \beta^m \boldsymbol{w}^m \leq (P_{\max})^\alpha. \tag{A.68}$$

Since $k+2n+1 > \dim((1, \boldsymbol{r}, \boldsymbol{P})) = k+n+1$, row vectors $(1, \boldsymbol{u}^1, \boldsymbol{v}^1), \cdots,$ $(1, \boldsymbol{u}^{k+2n+1}, \boldsymbol{v}^{k+2n+1})$ must be linearly dependent. Therefore, there are real scalars μ^m, $m = 1, \cdots, k+2n+1$, not all zero, such that

$$\sum_{m=1}^{k+2n+1} \mu^m = 0, \tag{A.69}$$

and

$$\sum_{m=1}^{k+2n+1} \mu^m (\boldsymbol{u}^m, \boldsymbol{v}^m) = 0. \tag{A.70}$$

Moreover, the coefficients can be chosen to satisfy

$$\sum_{m=1}^{k+2n+1} \mu^m \boldsymbol{w}^m \geq 0. \tag{A.71}$$

because if (A.71) does not hold, we can replace μ^m by $-\mu^m$ to satisfy (A.69)–(A.71). Then $(\boldsymbol{r}, \boldsymbol{P})$ can be rewritten as

$$(\boldsymbol{r}, \boldsymbol{P}) = \sum_{m=1}^{k+2n+1} (\beta^m - a\mu^m)(\boldsymbol{u}^m, \boldsymbol{v}^m), \tag{A.72}$$

for any real-valued a. Since not all of the μ^m are equal to zero. Therefore, there exists at least one $\mu^m > 0$. Define

$$\hat{a} = \min_{1 \le m \le k+2n+1} \left\{ \frac{\beta^m}{\mu^m} : \ \mu^m > 0 \right\} = \frac{\beta^q}{\mu^q}, \tag{A.73}$$

where q is an index with

$$\beta^q - \hat{a}\mu^q = 0. \tag{A.74}$$

Note that $\hat{a} > 0$, and for every m between 1 and $k + 2n + 1$,

$$\beta^m - \hat{a}\mu^m \ge 0. \tag{A.75}$$

Therefore,

$$(\boldsymbol{r}, \boldsymbol{P}) = \sum_{m=1}^{k+2n+1} (\beta^m - \hat{a}\mu^m)(\boldsymbol{u}^m, \boldsymbol{v}^m), \tag{A.76}$$

with

$$\sum_{m=1}^{k+2n+1} (\beta^m - \hat{a}\mu^m)\boldsymbol{w}^m \le \sum_{m=1}^{k+2n+1} \beta^m \boldsymbol{w}^m - 0 \le (P_{\max})^{\alpha}. \tag{A.77}$$

Every coefficient $\beta^m - \hat{a}\mu^m$ is nonnegative, their sum is one, and furthermore, the index q satisfies (A.74). In other words, $(\boldsymbol{r}, \boldsymbol{P})$ is represented as a convex combination of at most $k + 2n$ points of S. This process can be repeated until $(\boldsymbol{r}, \boldsymbol{P})$ is represented as a convex combination of at most $k + n + 1$ points in S.

A.8. Proof of Theorem 4.1

P4.1 is a special case of P4.8. Using the same technique developed in Appendix A.7 (and dropping the $\boldsymbol{P}$ dimension in defining R), we can show that the optimal utility of P4.1 can be achieved by a $(k + 1)$-piecewise constant power allocation. Moreover, an optimal solution $\boldsymbol{r}^*$ must be found on the boundary of the feasible set of rate vectors. (If one has an optimal rate vector being an interior point, one can increase all of its dimensions until reaching the boundary with no loss of utility.) Thus, k-piecewise constant power allocation achieves the optimum of P4.1.