Heterogeneous Cellular Networks: From Resource Allocation To User Association

by

Jagadish Ghimire

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2015

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Heterogeneous networking paradigm addresses the ever growing need for capacity and coverage in wireless networks by deploying numerous low power base stations overlaying the existing macro cellular coverage.

Heterogeneous cellular networks encompass many deployment scenarios, with different backhauling techniques (wired versus wireless backhauling), different transmission coordination mechanisms and resource allocation schemes, different types of links operating at different bands and air-interface technologies, and different user association schemes. Studying these deployment scenarios and configurations, and understanding the interplay between different processes is challenging. In the first part of the thesis, we present a flow-based optimization framework that allows us to obtain the throughput performance of a heterogeneous network when the network processes are optimized jointly. This is done under a given system "snapshot", where the system parameters like the channel gains and the number of users are fixed and assumed known. Our framework allows us to configure the network parameters to allocate optimal throughputs to these flows in a fair manner. This is an offline-static model and thus is intended to be used at the engineering and planning phase to compare many potential configurations and decide which ones to study further. Using the above-mentioned formulation, we have been able to study a large set of deployment scenarios and different choices of resource allocation, transmission coordination, and user association schemes. This has allowed us to provide a number of important engineering insights on the throughput performance of different scenarios and their configurations.

The second part of our thesis focuses on understanding the impact of backhaul infrastructure's capacity limitation on the radio resource management algorithms like user scheduling and user association. Most existing studies assume an ideal backhaul. This assumption, however, needs to be revisited as backhaul considerations are critical in heterogeneous networks due to the economic considerations. In this study, we formulate a global α-fair user scheduling problem under backhaul limitations, and show how this limitation has a fundamental impact on user scheduling. Using results from convex optimization, we characterize the solution of optimal backhaul-aware user scheduling and show that simple heuristics can be used to obtain good throughput performance with relatively

low complexity/overhead. We also study the related problem of user association under backhaul-limitations. This study is a departure from our "snapshot" approach. We discuss several important design considerations for an online user association scheme. We present a relatively simple backhaul-unaware user association scheme and show that it is very efficient as long as the network has fine-tuned the resource allocation.

Acknowledgements

I am deeply indebted to my PhD supervisor Prof. Catherine Rosenberg for the constant support and guidance that she provided me throughout my time as a graduate student in the University of Waterloo. This thesis would not have been complete without her supervision and encouragement. She gave me the rare opportunity to be a part of her research group and for this I am forever grateful.

I would also like to thank the members of my PhD advisory committee, Prof. Patrick Mitran, Prof. Amir K. Khandani, Prof. Samir Elhedhli, and Prof. Ekram Hossain for offering their time to review my work as well as their invaluable suggestions that improved this thesis greatly.

I would not have completed this thesis without the unconditional faith that my wife (Natasha) had on me. Her smile was what kept me sane during the lows of my graduate student life.

My parents have always valued education and learning as the most noble pursuit. This view of theirs had influenced me to undertake PhD research. Their constant love, never in short supply, coming via regular phone calls has been a key ingredient of this achievement.

My sister (Urja) has always been the light of our family. Her positivity helped me get past moments of difficulty. She is my inspiration. I am forever grateful for her unconditional love. I am also thankful to my two brothers (Thuldaju and Sandaju) for always encouraging me. I am also grateful to my mother-in-law and father-in-law for the constant encouragement.

Last but not the least, I am thankful to all my friends in Waterloo. Their company made life such a joy.

Dedication

To my mother, *Jagadambika Ghimire*!

Table of Contents

List of Figures

List of Tables

List of Abbreviations

3GPP	3rd Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
ABSF	Almost Blank Sub-frame
ADSL	Asymmetric Digital Subscriber Line
AG	Antenna Gain
AI	Air Interface
AL	Access Link
BHU	Backhaul-unlimited
BL	Backhaul Link
BS	Base Station
C-RAN	Cloud-RAN
CAPEX	Capital Expenditure
CCD	Co-channel Deployment
CDMA	Code Division Multiple Access
CIR	Carrier-to-Interference Ratio
CoMP	Coordinate Multipoint
DL	Direct Link
FBS	Femto Base Stations
FD	Full-duplex
GM	Geometric Mean
HD	Half-duplex
HetNet	Heterogeneous Network
IC	Interference Cancellation
IP	Integer Problem
ISD	Inter-site Distance
ISet	Independent Set

JP	Joint Processing
KKT	Karush-Kuhn-Tucker
LMDS	Local Multipoint Distribution Service
LOS	Line-of-sight
LTE	Long-Term Evolution
LTE-A	LTE-Advanced
MBS	Macro Base Station
MCS	Modulation and Coding Scheme
MIMO	Multiple-input and Multiple-output
mmWave	Millimeter-wave
MUD	Multi-user Diversity
NC	No Coordination
NLOS	Non-line-of-sight
NUD	Non-uniform Distribution
NUM	Network Utility Maximization
O	ON-OFF coordination
OD	Orthogonal Deployment
OFDM	Orhtogonal Frequency-Division Multiplexing
OFDMA	Orthogonal Frequency-Division Multiple Access
PBS	Pico Base Station
PF	Proportional Fairness
PSD	Partially Shared Deployment
QAM	Quadrature Amplitude Modulation
RA	Resource Allocation
RAN	Radio Access Network
rAI	receive AI
RE	Range Extension
REC	Relay-enhanced Cellular
RN	Relay Node
RR	Round-robin
RRM	Radio Resource Management

SC	Small Cell
SCF	Small-cell First
SINR	Signal to Interference plus Noise Ratio
SNR	Signal-to-noise ratio
SON	Self-organization
tAI	transmit AI
TC	Transmission Coordination
TDMA	Time-Division Multiple Access
UA	User Association
UD	Uniform Distribution
UE	User Equipment
US	User Scheduling
WLAN	Wireless Local Area Network

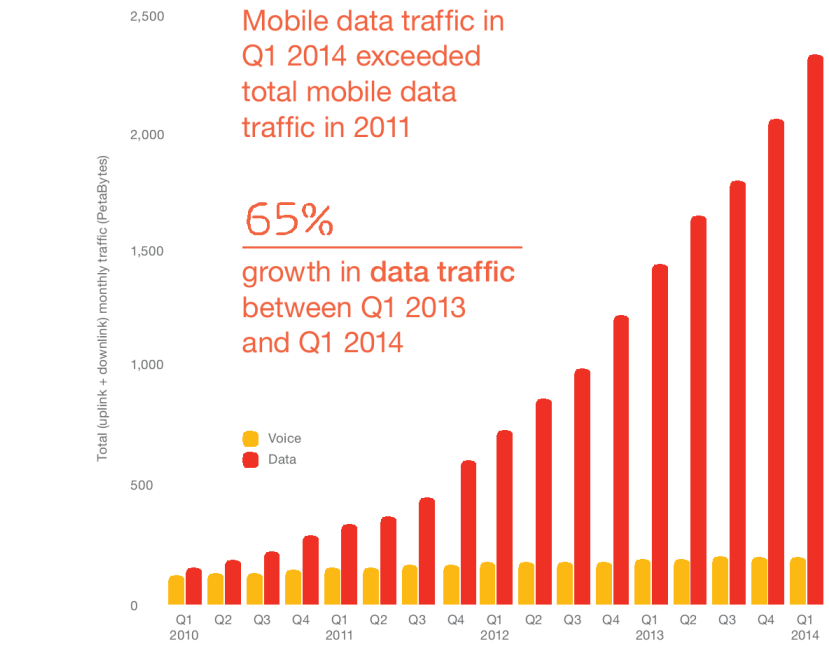
Chapter 1

Introduction

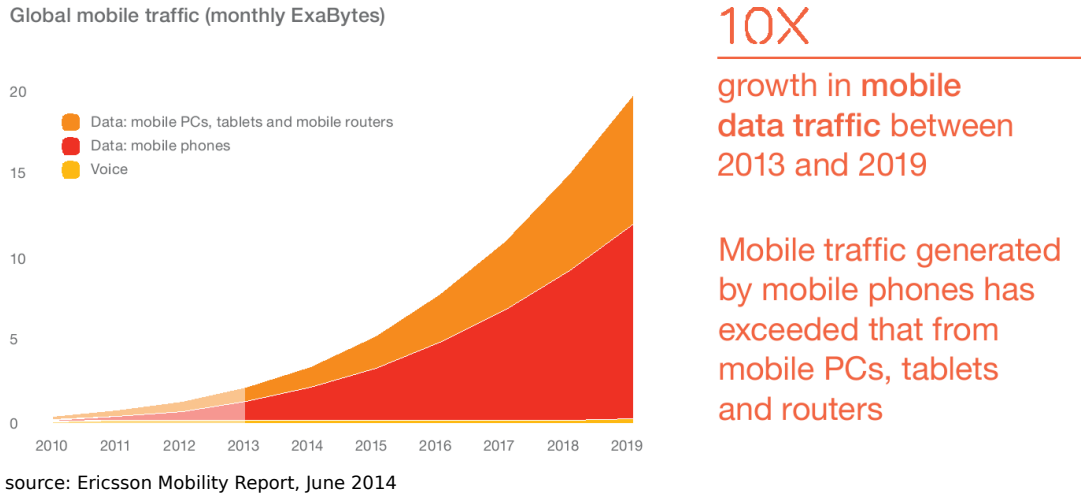
1.1 Overview

Cellular networks were initially designed for *voice* applications. However, with the introduction of data service with ubiquitous connectivity to the Internet, cellular network operators are facing an overwhelming growth of data traffic demands mainly fueled by the rapid development of high-end mobile devices including smart-phones and tablets. A large portion of this data is expected to be mobile video that has a much larger rate requirement than voice or web-browsing [34]. Fig. 1.1 summarizes the findings of a recent industry report [34] that shows a 65% increase in mobile traffic demand between the years 2013 and 2014. It also forecasts a 10 folds increase in mobile traffic between the years 2013 and 2019. Different generations of wireless cellular networks (3G, LTE, 4G etc.) have tried to keep up with this ever-increasing demand. The next generation of these technologies, often referred to as the fifth generation (5G), is expected to support even more traffic [97].

Unlike in wired networks, capacity expansion of wireless networks is not easy. "Adding more copper" approach does not work for wireless networks, mainly due to the limited availability of wireless spectrum. Improving the utilization of the spectrum by employing smart radio technologies like *cognitive radio* has been the subject of many recent studies [10] [48]. Improving the spectral efficiency of a point-to-point link has always been a major



(a) Growth of mobile traffic between 2011 and 2014



(b) Expected growth in mobile traffic

Figure 1.1: Statistics showing the growth in mobile traffic and the expected forecast (Ericsson Mobility Report, June 2014 [34])

focus in wireless research. Important innovations in physical layer technologies have made wireless broadband access to Internet feasible. However, as the spectral efficiency of a point-to-point link is quickly approaching theoretical limits [31], this line of thinking alone is not going to be sufficient. Obviously, if more spectrum are being made available for wireless network operation, that would bring many-fold increase in network capacity. Due to some recent policy-level decisions, more spectrum is indeed available especially at higher frequencies (e.g., mmWave bands). However, communication technologies at these bands are still far from mature, and are not expected to be main-stream any time soon.

In the recent years, solutions based on topological and architectural innovations have gathered a lot of interest, both in the industry [1] and the academia [30] [64] [57]. The main idea involves *network densification* in the form of *Heterogeneous Cellular Networks* [51].

1.2 Heterogeneous Networks

Heterogeneous Networks (HetNets) comprise a set of low-power base stations (BSs) overlaying the existing macro cellular system [31]. These low power BSs form small cells within the macro cellular coverage area of macro base stations (MBS). These BSs are simply referred to as small cells (SC)[1]. These SCs are often connected to the core via some backhaul infrastructure. Pico base stations (PBS) are operator-deployed small BSs connected to the core via wired backhaul links. Femto base stations (FBS) are much smaller in form-factor and coverage, and are often used for indoor coverage with inexpensive backhaul links. Relay Nodes (RNs) are small cells with wireless backhaul links to the macro cell. An example of a HetNet with a mix of PBSs and RNs is depicted in Fig. 1.2.

1.2.1 Small cells with wired backhaul links

Connecting the small cell base stations to the macro base station (MBS) via wired backhaul links is the most common scenario (*wired scenario*). In this scenario, there are three types

[1]The term base station (BS) refers to both the MBS and the SCs.

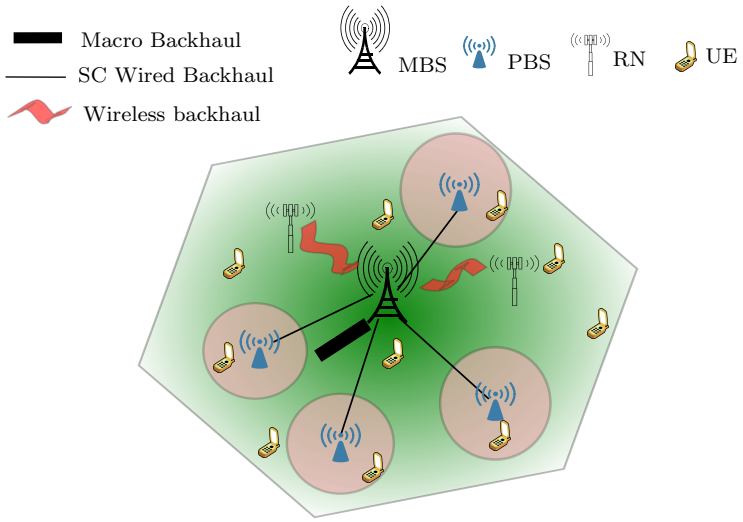


Figure 1.2: A Heterogeneous Network

of links, the *direct links* (DL) between the MBS and the User Equipments (UEs), the *access links* (AL) between the SCs and the UEs, and the wired *backhaul links* (BL) between the MBS and the SCs. The first two types are often referred to as *user links*. If the backhaul links are sufficiently provisioned, the performance of such HetNet would depend on the radio resource management (RRM) algorithms and techniques used in the user links. It is historically the case that the capacity bottlenecks are in the wireless access end, and hence backhauling is often assumed to be ideal. In the future ultra-dense HetNets, this assumption might need to be revisited, meaning that backhaul link considerations are to be incorporated.

1.2.2 Small cells with wireless backhaul links

Deploying wired backhaul links is not always feasible. In many circumstances, the flexibility offered by wireless backhaul links makes deploying relay nodes (RNs) an attractive alternative to the wired small cells. 3GPP LTE-A standards include HetNets with relay nodes (RNs) as an important enhancement for improving macro-cell capacity and coverage, forming the so-called *relay-enhanced cellular* (REC) networks. Type 1 static RN of LTE-A,

as defined in [6], enhances the communication between an MBS and a UE by *decoding and forwarding* the data packets. These RNs form cells of their own and can be viewed as small cell BSs with wireless backhaul links operating in the same set of bands as the user links. In other words, no additional band is available for operating the backhaul links. In this thesis, we refer to such a scenario as the *user-band relay scenario*. Unlike in the wired scenario, the backhaul links compete with user links for the radio resources (frequency, time, and transmit power). A number of different configurations can exist which differ depending on the node capabilities (including the number of air interfaces, directivity of each air interface, etc.), and the way in which the channel resources are allocated to different links. It is not clear how different configurations perform with respect to each other and with respect to a wired scenario.

Another wireless backhauling scenario also exists where a separate band is available for operating the backhaul links. We call such a scenario as the *dedicated-band relay scenario*. In recent years, using the mmWave bands for wireless data links is attracting a lot of attention. However, using this resource for user links is not as straightforward, mainly due to the fact that designing hand-held UEs for this spectrum is challenging. On the other hand, utilizing this spectrum for the static backhaul links can be seen as a potential solution for tackling the spectrum scarcity problem in the wireless industry, provided that a sufficient bandwidth is available and that the propagation characteristics (path-loss, shadowing etc.), available transmit power and the underlying physical layer techniques support the required capacity of the backhaul links. Note that such a backhauling (unlike the *user-band relay scenario*) can potentially be configured to approach the performance of a wired scenario.

1.3 Challenges

1.3.1 Diverse deployment scenarios

It is evident from the above discussion that HetNets comprise different deployment scenarios, with different types of backhaul links with different types of limitations, bands and

air interface technologies, as well as diverse topologies. From an engineering point of view, it is highly desirable to study these scenarios under the same framework. The ability to model them using a unified framework would enable us to perform a comparative study of the different deployment choices. This is one of the main objectives of this thesis.

1.3.2 Different network processes and their complex interplay

A heterogeneous architecture brings in a rich topology to the otherwise flat network architecture, but the deployment of different low power BSs over existing MBS coverage poses new challenges on important network processes including *resource allocation*, *user scheduling and transmission coordination*, and *user association* which are all intricately linked to interference management and throughput performance.

Resource allocation

Resource allocation is the process of allocating wireless resources to the network nodes, i.e., to the base stations and the users. In OFDM-based wireless networks, sub-channels[2] are the obvious examples of such resources. A particular resource allocation scheme determines how the available channel resources are allocated to the different nodes. In homogeneous cellular networks, resource allocation is often very simple as it involves allocation among nodes of similar coverage, and load. For example, a reuse-pattern can be used to allocate equal number of channel resources to different BSs. This approach, not only simplifies the resource allocation, but also offers a certain level of interference guarantees, as the influence of interference among the BSs is rather symmetric. In heterogeneous networks, different nodes have different coverage area and load. Moreover, the power disparity between the MBS and the SCs makes the interference asymmetric, which together make resource allocation considerations more complicated.

In the case of wired deployment of small cells, resource allocation involves assigning the channel resources to the direct links and the access links. In this thesis, we look at three types of resource allocation schemes: *Co-channel deployment* (CCD), *Orthogonal*

[2]We use the terms sub-channel and channel, interchangeably.

deployment (OD), and *Partially shared deployment* (PSD). In CCD, the available radio resource is used by all the BSs within a given macro coverage. In OD, channels are divided between the MBS and the SCs so that interference is kept at a reasonable level. In PSD, MBS is allowed to transmit on the channels available to the SCs, albeit at a lower power. Some of these resource allocation schemes have parameters that need to be tuned. An optimal choice of these parameters would yield good performance, but is not trivial. If we call these parameters the *resource allocation variables*, a good resource allocation algorithm would aim at finding good values for these variables. These resource allocation schemes determine how the channel resources are divided among the SCs and the MBS within a macro cell. There is another (higher level) resource allocation scheme that determines what channels are available in a given macro cell.

The case of relay deployment introduces an additional type of link, the backhaul link. The addition of this new type of link results in more distinct ways in which we could allocate resources and it is often very difficult to understand which ones are better than the others.

User scheduling and transmission coordination

Often in homogeneous networks, user scheduling (US) is done locally by each BS on the channels allocated by RA, to meet some throughput objective (e.g., proportional fairness (PF)). In that case, a BS schedules its users independently of the other BSs.However, this per-BS (also called local) user scheduling model needs to be revisited in the HetNet context. User scheduling is seen as an important network process that can be used to manage interference among the BSs. This is done by coordinating the transmissions of the BSs. Such a *transmission coordination* (TC), in the most general form, can be carried out by scheduling the BSs in time together with power control at each BS, which is very complex. In this thesis, we will focus on a simple type of TC called the *ON-OFF TC* where transmission coordination is carried out by scheduling BSs such that a BS can either be transmitting with the maximum available power or not transmitting. 3GPP considers such a coordination mechanism as a viable option in LTE-A networks [6]. If transmission coordination among BSs is possible, independent local user schedulings are not optimal

from the network performance point of view. In this case, user scheduling decision might need to be taken across different BSs jointly. Thus, the tightly coupled nature of TC with user scheduling across multiple BSs mandates a global (i.e., across multiple BSs) optimization approach. It is however not clear what magnitude of gains can be expected by the introduction of such sophisticated US.

Even without transmission coordination, there is a need to look at user scheduling as a global process for the scenarios where the backhaul links have capacity limitations. This includes the wired as well as the relay deployment scenarios. In these cases, it can be shown that user scheduling decision at a BS impacts another BS, and thus a global approach to user scheduling can yield the best performance (a local approach could yield infeasible solutions if not performed properly). However, a global approach to user scheduling leads to high complexity. The trade-off between local user scheduling and globally optimized user scheduling is not well understood.

User association

In homogeneous networks, cells are usually non-overlapping (except at the cell-edge) and thus a user associates to the BS that offers the best Signal to Noise-plus-Interference Ratio (SINR) value. However, in a HetNet context, such an approach does not work well. Since the MBS transmits at a higher power, the received SINR from the MBS is usually much higher than the SINR from the low power BSs, thereby making more users associate to the MBS. This in effect nullifies the vision of the HetNet deployment. It is thus important to revisit user association in the HetNet context.

Interplay of network processes

Each of the above-mentioned network processes can be fine-tuned to yield good network performances. But, it is often the case that one process impacts the other, and the interplay is usually complex. Understanding them, and taking the right deployment configuration is important to realize the *HetNet potential*.

1.4 Contributions

In this thesis, we will focus on the downlink, and will study the HetNet from a throughput performance point of view. Our contributions can be summarized in two different headings as follows. A more detailed summary of contributions are presented in the beginning of the relevant chapters.

1.4.1 Unified optimization framework

As discussed earlier, there is a need for a unified approach that allows us to study the different HetNet deployment scenarios and configurations under the same framework. Such a framework should allow us to characterize the performance of a deployment option when the network processes are optimized. Our main contributions in this context can be summarized as follows:

1. We present a flow-based[3] optimization framework (in Chapter 3) that allows us to obtain the throughput performance of HetNet deployment when the network processes are optimized jointly. This is done under a given system "snapshot", where the system parameters like the channel gains and the number of users are fixed and assumed known. We only consider the active users in the network and hence assume that there is one flow per user. Moreover, we assume that the users are greedy and hence want to maximize their individual flow-rates. Our framework allows us to configure the network parameters to allocate optimal throughputs to these flows in a fair manner. This is an offline-static model and thus is intended to be used at the engineering and planning phase to compare many potential configurations and decide which ones to study further. To make our framework tractable, we have made a key assumption of multi-path routing, which is equivalent to allowing users to associate to more than one BS. We validate that the upper-bounds provided by this assumption are tight.

2. Using the above-mentioned formulation, we provide important engineering insights on the throughput performance of different configurations under a global proportional

[3]A flow corresponds to a data stream from the network to a particular user.

fairness (PF) objective function.

- A detailed study of different wired deployment scenarios (in Chapter 4) shows the performance of different resource allocation schemes, transmission coordination mechanisms, and user association schemes. It shows how CCD requires transmission coordination, but OD/PSD can perform well without such complicated coordination. It also shows that associating a UE to more than one BS is not likely to offer significant throughput gains.
- A detailed study of different wireless deployment configurations (in Chapter 5) highlights the importance of the right configuration for a successful relay deployment. For a user-band scenario, the results show that some configurations yield negative to negligible gains, whereas some others offer gains close to the wired upper-bound. We also show that a mmWave band backhauling is a very promising solution to achieve huge capacity gains.

1.4.2 Analytical insights and simple algorithms

The afore-mentioned joint modeling approach allows us to study different network processes together, but it suffers from some limitations. The model is limited to the offline study phase, due to its "snapshot" approach. Moreover, since all network processes are optimized jointly, it cannot reflect the reality of networks where different network processes are optimized at different time-scales. In order to yield simple models that can result in useful insights, and to obtain results that can be used to design good online algorithms, we take a different approach where we study these network processes, one at a time (assuming that the other ones are fixed).

1. In Chapter 6, we study the global α-fair user scheduling problem under limited backhaul capacities. Here, we depart from the flow-based approach as we look at a system where the complexities of such an approach are not required. Our contributions in this scope are as follows:

    - We present the decomposition structure of the global α-fair scheduling problem under different scenarios of backhaul limitations.

- We present analytical solutions to the decomposed local α-fair scheduling problems and show the conditions when the decomposed problems yield global optimal solutions.
- For the general case, where decomposition does not yield optimal results, we present a very good heuristic that can be implemented as a simple online scheduling algorithm.

2. In Chapter 7, we introduce a dynamic user arrival/departure process to the model, and study optimal and sub-optimal α-fair user association schemes under backhaul limitations. Some of the key findings in this chapter can be summarized as follows:
   - For backhaul unlimited case and with $\alpha = 1$, we show how a very simple rule can be used to achieve optimal user association.
   - In the general case, the optimal algorithm can be very complex, but we show that if other network processes are optimized, a very simple user association scheme can be employed, without a huge penalty in performance.

1.5 Outline

The rest of the thesis is organized as follows. Chapter 2 presents a summary of the related work. In Chapter 3, we present the flow-based joint optimization model for a given snapshot of the network that allows us to characterize the network performance when the network variables are jointly optimized. We use this model in Chapter 4 to study different transmission coordination, resource allocation, and user association algorithms in HetNets with wired backhaul links. In Chapter 5, we use the formulated model to study different scenarios and configurations of relay node deployment. In Chapter 6, we focus on the wired deployment case and present different analytical results and algorithms on the global α-fair scheduling under backhaul limitations. In Chapter 7, we extend the model to incorporate user arrival/departure process and study user association schemes under the α-fairness framework. Chapter 8 presents a summary and some extensions of this thesis work.

Chapter 2

Literature Review

In this chapter, we provide an overview of the relevant literature on resource allocation, transmission coordination, user scheduling, and user association in HetNet. In addition to providing a context to our research, we present our view on the limitations of the existing work, and how we approach to address them in this thesis.

2.1 Resource Allocation

2.1.1 Resource allocation under wired deployment

Let us first focus on the SCs with wired backhaul links. Example of such deployments are pico base stations (PBSs) with dedicated wired backhaul to the network core. Under such deployment scenarios, two types of wireless links are relevant: *direct* and *access* links. Under downlink, a direct link is identified as the wireless link from the macro base station to a user equipment (UE). An access link is identified as the wireless link from an SC to a UE (e.g. PBS-UE link). Resource allocation schemes can be distinguished based on how the channel resources are allocated to these two types of links.

Orthogonal deployment (OD) allocates orthogonal frequency to the MBS (i.e., the direct links), and the SCs (i.e., the access links). It is an obvious solution to protect the SC users

from MBS interference. Such an approach results in a simple interference management mechanism. Under orthogonal deployment, the following research questions have been studied.

Channel splitting between tiers In an OFDM-based system, the pool of sub-channels[1] can be divided into two *disjoint sets*, one for the macro operation and the other for the low power BSs. Performing an optimal split is crucial for a better performance of the HetNet deployment. Under a given user association, Sundaresan *et al.* showed that the optimal spectrum allocation to the low power BSs is proportional to the fraction of users allocated to them [91]. They also propose an iterative algorithm that converges to the optimal split. Their result is based on the assumption of a fixed user association and a per-base-station (BS) proportional fair scheduling. This result does not hold for open-access femto cells or PBS deployments with user association as one of the problem variables. In [23], Chandrasekhar *et al.* study the optimal channel splitting problem for a macro cell overlaid with a set of randomly deployed femto cells. For a given user association, they study the optimal channel splitting parameter for maximum area-spectral efficiency under a given per-BS scheduling policy and under the settings where the channel-access by femto-cells are randomized. Via numerical results, they show that the optimal channel splitting depends on the density of hot-spots and the data-rate requirements of users. Because of the nature of random channel access and assumption on user association and per-BS scheduling, the solutions to optimal channel split problem can not be generalized to joint problems involving other network processes.

Channel allocation in the same tier In [104], Yonezawa *et al.* apply the idea of chromatic polynomial from graph theory to minimize the co-channel interference among the SCs based on the interference graph which is constructed centrally. Similarly, in [91], the authors propose a distributed resource allocation mechanism among femto cells based on the distributed hashing of the largest maximal clique size of the interference graph constructed among the femto cells. It is a distributed method and the probabilistic nature of the resource allocation occasionally results in resource collision

[1] We use the terms channel and sub-channel interchangeably.

which is corrected by rehashing. In [65], the authors study sub-channel allocation problem in an OFDM-based femto cell network. A distributed mechanism among the femto cells is divided into a *sensing phase* and a *tuning phase*. Each femto BS adjusts its sub-channel usage based on the reports that it gets from other femto BSs. An iterative algorithm like this can suffer from slow convergence behavior and can often end up in local optimal solutions.

Co-channel deployment (CCD) is an alternative approach to the orthogonal deployment. Under co-channel deployment, all of the low power and high power (macro) BSs transmit in the same set of frequency resources. The following benefits of co-channel deployment have been highlighted in the literature [31] and in industry reports [1].

- Simplicity of resource allocation: Optimal channel splitting problem is avoided.
- Easier hand-off procedures for mobile UEs as the cell-search is easier.
- Co-channel deployment does not rely on the availability of a large spectrum.

Despite these potential benefits, co-channel deployment incurs severe interference problems. In downlink, the power disparity between the MBS and the small power BSs (SCs) is the main hurdle. A great deal of research in co-channel deployment has focused in interference avoidance or mitigation. Interference in co-channel deployment can be minimized by using advanced physical layer technologies, the most famous being *Interference Cancellation (IC)*. In [13], Andrews *et al.* present a high-level description of IC and its potential use in the future cellular networks. In [80], Sahin *et al.* propose an iterative co-channel interference cancellation technique and via simulation demonstrate that an improved symbol error rate performance is possible. Such IC techniques are known to require sophisticated signal processing and synchronization capabilities. A set of system-level solutions, on the other hand, do not require sophisticated physical layer technologies and moreover are expected to be flexible to implement. A common system-level solution to combat interference involves *coordination of transmissions* at time-domain (as the pure form of downlink co-channel deployment requires all transmissions to be carried out in the same carrier frequency). In the most general form, such a coordination entails coordination

of transmit power, commonly called network-wide power control, where the BSs mutually coordinate the transmit power so as to maximize the performance (e.g., minimum SINR maximization or throughput maximization). In [28], Claussen *et al.* carry out the performance evaluation of co-channel deployment of MBS and a number of FBSs in presence of a power tuning mechanism that maintains a constant femto cell coverage. Even though such an arbitrary selection of power control rule does not guarantee optimal power control, it demonstrates that co-existence of MBS and low power BSs is feasible. It also highlights the importance of auto-configuration and public access. In [46], Guvenc *et al.* take a slightly different approach, fundamental focus being the fact that users associating to low power BS and that are close to MBS face more downlink interference from the MBS. Such power disparity problem is not severe for users of small BSs which are far away from the MBS. The authors propose to split the deployment area into an inner area and an outer area. Co-channel deployment is advocated in the outer area and a split spectrum deployment is proposed for the inner area.

In [31], Damnjanovic *et al.* discuss a type of resource allocation where one frequency resource set is used for macro coverage and the other is shared by both the macro and the low power base stations. In the shared resource, MBS transmits at a lower than nominal power for avoiding power disparity. Such an overlapped channel allocation can be thought of as an alternative to two extremes of pure channel allocation paradigms. This will allow some protection to the SC users, while maintaining macro coverage, as shown in [49].

A comparative study of orthogonal deployment, co-channel deployment, and the overlapped channel allocation is carried out in [60]. Via simulation, the authors show that orthogonal deployment outperforms the co-channel deployment in terms of control channel coverage while co-channel deployment achieves a higher system capacity. These performance comparisons however are carried out with channel splitting that is not performed optimally. As the performance of resource allocation heavily depends upon the channel splitting parameter, these results can not be considered fair. Moreover, the user associations are carried out following simple SINR based criteria, which are known to perform poorly. Some other works including [31] also suffer from these shortcomings. In order to compare different resource allocation schemes more fairly, other network processes have to be chosen optimally, or at least not arbitrarily. To the best of our knowledge, a com-

prehensive model for comparative assessment of the resource allocation schemes with the joint consideration of other important network variables (including user association and scheduling) is missing in the literature.

2.1.2 Resource allocation under relay deployment

So far we have discussed the resource allocation schemes proposed for the deployment of SCs with dedicated wired backhaul. Relay nodes (RNs) are another type of low power BSs, differing from the conventional PBSs in the sense that they do not have dedicated wired backhaul to connect to the network core. In relay deployment, in addition to the *direct* and *access* links, a third type of wireless links called *backhaul* links also need to be considered for resource allocation.

Wireless *backhaul* links are the links connecting the RNs to the MBS such that downlink flow to any user associated with an RN is routed in two hops via a backhaul and an access link. Let us first look at the case of *user-band relay scenario* where the backhaul link also operates on the same band as the user links.

User-band deployment

All three types of resource allocation discussed above are relevant in this case. Two main categories of relay operation are identified in the literature which are the immediate result of the specific choice of channel allocation [1].

In-band relays The backhaul (MBS-RN) link of an in-band relay is operated on the same set of frequency resources as that of the access (RN-UE) link. As the transmission at access link interferes with the simultaneous reception at the backhaul link, RN with half-duplex (HD) communication constraint requires to operate the backhaul link and the access links at non-overlapping times (e.g., Type 1 RN in 3GPP LTE-A [1]). However, if the backhaul receiver is protected from the interference generated by the access link transmitter (e.g. by the spatial separation of antennas), a full-duplex (FD) operation might also be possible (e.g., Type 1(b) RN in 3GPP LTE-A [1]). [22]

presents the results for the peak spectral efficiency of LTE-Advanced with in-band backhauling. It proposes an analytical model to calculate the cell spectral efficiency of such deployments.

Out-of-band relays The backhaul link of an out-of-band relay is operated on different set of frequency resources as that of the access links. This means that a simultaneous transmission at an access link while receiving at the backhaul link is possible (e.g., Type 1(a) RN in 3GPP LTE-A [1]). Orthogonal channel splitting among the MBS and RNs is an example of resource allocation that results in out-of-band relay operation. [45] presents a comparative study of these two types of deployments.

In the general form, different resource allocation schemes in user-band relay deployment can be distinguished based on the way available frequency resource is divided among the three types of wireless links. The focus in this thesis is not on the link-level benefits of relay channels, which has gathered a lot of attention. We want to understand different choices that we have in terms of resource allocation schemes and how they compare with each other, as well as with the wired SC deployments. To the best of our knowledge, there is a general lack of such studies.

Dedicated-band deployment

Millimeter-wave (mmWave) band has been available for use in telecommunication industry, which has generated a lot of interest in this new field. There have been some recent studies suggesting that this newly available spectrum can be technologically viable as user band spectrum [77] [76]. There are however challenges to operate the user links in the high frequency spectrum. What is clear at this point is using these high frequency bands is much easier for static backhaul links than the mobile user links. Hence, using mmWave band for backhauling is seen as an attractive solution. Our approach in this thesis is to look at them from a system-level perspective, as opposed to link-level perspective. To the best of our knowledge, the literature lacks studies characterizing the feasibility of mmWave backhaul in the LTE HetNet context and its comparison with other backhauling options.

2.2 User Scheduling

Single-cell network

User scheduling in a single-cell network is a well-studied problem and different user-scheduling policies have been proposed in the literature. User scheduling policy is usually based on some throughput-based performance objective. A maximum fair throughput allocation tries to maximize the throughput of the worst user, thereby dedicating more transmission times to them. This notion of egalitarian fairness however sacrifices the system aggregate throughput [98]. *Proportional fairness* is another scheduling criteria where the objective is to maximize the geometric mean of the user throughputs. It is known to provide a good trade-off between fairness and aggregate throughput [55]. Under the assumption that the channel variations of users are identical around the long-term average, an opportunistic scheduling scheme is proposed in the literature, which exploits multi-user diversity and yet maintains the proportional fairness in the long-term [7]. In such channel-aware proportional fair scheduling, at a given time-slot t, the user with the best instantaneous rate normalized with the accumulated average throughput, i.e. the user $i*$ where $i^* = \arg\max_i \frac{R_i(t)}{\overline{R}(t)}$ is scheduled [16]. $R_i(t)$ and $\overline{R}_i(t)$ respectively represent the instantaneous rate at time t and average throughput allocated by time t to user i. The benefit of such an opportunistic scheduling is the ability to exploit the multi-user diversity (MUD) gain $G(N)$ which increases with the number of users N [19] [59]. Under the scenario of completely static channel gains, opportunistic scheduling is equivalent to the RR scheduling [19]. A more general form of fairness has been introduced in [71], which is commonly referred to as the α-fairness, and has been used often in throughput allocation frameworks usually under *network utility maximization* formulations [90], [72]. By changing the α parameter, different levels of *fairness-efficiency trade-off* can be achieved, which is the main strength of this approach.

From single-cell to multi-cell

Under the assumption that the wired backhaul are of unlimited capacities and that all of the BSs are transmitting all the time, it has been shown that global optimal scheduling

coincides with independent per-BS local optimal scheduling [99]. But this result can not be generalized to the case when BSs cooperate and thus channel rates are the result of not only the channel variations, but also the coordination among the BSs [99]. In general, performing optimal scheduling independently at each base station does not coincide with the global optimal scheduling when other network variables (like user association decisions) are jointly allocated [21]. The study of user scheduling as a global process is thus important in HetNets. However, because of the intricate dependencies of the scheduling with other network processes, a global scheduling problem can be a very difficult problem to solve. An ability to decouple user scheduling processes at per-BS levels would certainly be attractive from the implementation point of view, but can potentially lead to performance degradation. To the best of our knowledge, there have not been many studies that explore these aspects.

2.3 User Association

The problem of user association arises whenever a user can get service from more than one BS (i.e., overlapping coverage). Even under homogeneous cellular networks, user association (also sometimes called *cell-site selection*) problem was studied at least as early as 1995. In [103] and [47], the authors study the user association problem jointly with power control in the context of homogeneous CDMA networks. In particular, they formulate an optimization problem of optimal power control and cell-site selection for minimization of total transmit power, subject to maintaining individual Carrier-to-Interference Ratio (CIR) targets for each mobile. These studies look at the problem from the physical layer capacity point of view and can be seen as the techniques of exploiting user assignment for interference reduction.

User association in homogeneous networks is perhaps not as critical as in heterogeneous networks. Study of user association has thus gathered much more interest recently, as it seems to be crucial for a successful HetNet deployment. Below, we survey a class of user association problems and proposals in the literature, under a heterogeneous network settings.

2.3.1 Optimal user association

Optimal user association problems are combinatorial in nature and the resulting optimization formulations are usually NP-hard. In [21], the authors formulate a user association problem under a global proportional fair throughput allocation framework. In a system with a given number of users and a given number of BSs, they show that arbitrary user association can lead to non-Pareto optimal results. In addition to highlighting the need for a globally optimal user association, they also present offline and online algorithms for user association. Among a big list of papers relating to user association problem, this particular work is special in the sense that it presents a rigorous and yet simple formulation for optimal user association under a global throughput objective. Under a related scenario of multiple access-points of a WLAN, authors in [61] study similar problem and propose a method for obtaining user association with one user per access-point restriction using rounding of fractional solution via generalized assignment problem (GAP).

2.3.2 User association rules

Due to the high complexity of computing the optimal user association, different simple user association rules have been introduced in the literature. These rules simplify the user association process at the expense of some performance degradation. Good rules are expected to be simple to implement and yet perform well with respect to the optimal association. The simplest model that was popular (and to some extent meaningful) in homogeneous network setting was the *best-SINR* rule. In this rule, a user associates to a BS that provides the highest SINR. Such a rule performs poorly in HetNet mainly because of the power disparity. As the MBS transmits with high power, a large number of users tend to associate to the macro cell. This might result in highly loaded MBS. In order to overcome this problem, a user association rule called *range extension* is introduced in [57]. Under range extension, a user associates to the BS that has the highest channel-gain. This allows more users to associate to the nearby SCs, as the power disparity is avoided to some extent by normalization. Via simulation, it was shown that range extension can improve the throughput performance as compared to the conventional best-SINR based user association [57].

Some other user association rules have also been proposed in the literature (e.g. "Based on Queue" in [75]). The main idea behind these heuristic approaches is to bias user association in favor of the small power BSs. A class of user association rules can be abstracted well with a user association rule called *Small-cell first (SCF)*, proposed by Fooladivanda *et al.* in [36]. Under a small-cell first user association rule with a given parameter δ, a user associates to an SC as long as the received SINR from the SC is greater than or equal to δ. Since δ is a parameter that can be tuned, this rule can be tuned to obtain very good performance. However, this introduces a new parameter-tuning problem, and thus increases the complexity. Rules very similar to SCF have been proposed in recent 3GPP technical reports, for the heterogeneous deployment of LTE-A [2].

Even though simple rules greatly simplify user association, it is in general difficult to establish that these rules will work over a large set of scenarios and network instances. Moreover, most of these user association rules do not incorporate the network-load information and thus might result in sub-optimal load distribution.

In a dynamic scenario, a closely related problem is the problem of *re-association*. Deciding when to trigger re-association is an equally important problem, and understandably has gathered a lot of attention. In our thesis, we have not studied this aspect in any details, and hence we have not included a survey of the literature on re-association.

2.3.3 User association and in-cell routing under relay deployment

User association problem under relay operation involves the effective routing of downlink flow, assumed to have originated at the MBS to a user either directly, or via one or more relay nodes in two hops. Some recent works in the literature extend the idea of user association from wired deployments to relay deployment. Under a local PF (proportional fair) scheduling, in [8], Ahn *et al.* formulate a user routing problem. As evident from the NP-hardness of similar problem in the wired scenario, they resort to a sub-optimal user association method which is based on the ordering of users in terms of the throughput difference between the MBS-UE link and the MBS-RN-UE link. They compare the performance of their simplified solution with the optimal solution obtained based on exhaustive

search for very small number of users (less than 17). They consider a simple scenario of one MBS and one RN and the analysis is carried out under a given relay link resource allocation.

In [67], Ma *et al.* acknowledge the limitations in [8] and present a formulation with multiple RNs and also consider the effect of backhaul link resource allocation. The complexity of the resulting in-cell routing formulation is tackled by using a set of greedy iterations. For each new user (whose association is yet to be determined), the improvement in the throughput objective is predicted for each of the available association options. The best of such options is chosen. This method can reduce the complexity associated with user association. However, such an arbitrary ordering of users is susceptible to the obvious problems a greedy sequential algorithm inherits.

2.4 Transmission coordination

Transmission coordination is considered as a tool to improve the coverage, cell-edge throughput, and system capacity, in both high and low load scenarios. Two general categories of coordinations are envisioned for 4G networks [1]:

Joint processing (JP) Under joint processing, a set of BSs form a cooperation set (also called the CoMP cooperating set in LTE-A). A user can be served jointly by any number of these BSs within the same cooperation set. Under *joint transmission* mode, downlink transmission can be done simultaneously from multiple transmission points to improve the effective SINR and/or actively cancel interference for other UEs. Under *dynamic cell selection* mode, at any given slot, only one BS can transmit the data to a UE. However, the transmission point can be changed in subsequent slots. Both of these mechanisms require each BS in the cooperation set to have a copy of the downlink data. There are a number of studies that propose techniques to form good cooperation sets (or clusters), for example, [81] [69].

Transmission coordination by power control This is another method where BSs mutually coordinate their transmission powers to result in optimal network operation.

Sophisticated power control are shown to result in intractable problems. A particularly simple form of power control is the binary power control where a BS can choose one of the two states: either transmit with the maximum power or stay idle [44]. It can be also be seen as the "coordinated scheduling" [99]. Coordinated scheduling has been proposed for LTE-A systems in [1]. In [32], Das *et al.* propose scheduling schemes in which scheduling decisions are made centrally by a central controller jointly for a cluster of BSs. Via simulation experiments, they show that gain resulting from coordinated scheduling is significant in a multi-cell CDMA network. They call it *interference gain*. Optimal coordinated scheduling is a result of trade-off between *spatial reuse* where more number of simultaneous transmissions are intended and *interference minimization* which favors small number of simultaneous transmissions. It is clear that the level of interference among co-channel BSs dictates the benefit of coordinated scheduling. In a fully orthogonal deployment, there is no benefit of coordinated scheduling whereas a complete co-channel deployment requires coordinated scheduling for acceptable performance.

Almost blank sub-frame (ABSF) proposal in LTE-A is another example of transmission coordination. The MBS mutes its data transmission in ABSF so that the SCs can get better SINR. This offers some protection of the SC users from the macro interference. There are a number of studies that present algorithms to perform optimal muting of the MBS [14] [63].

2.5 Joint Resource Allocation, User Association, Transmission Coordination, and User Scheduling

In [36], Fooladivanda *et al.* have formulated a joint resource allocation, user association, and reuse pattern optimization problem in a heterogeneous network comprising a macro cell and a set of pico base stations. Under a channel splitting resource allocation, they present a method for obtaining the optimal channel split as well as the optimal user association. They showed that a full reuse of the available channels among the PBSs results in optimal solution, thereby showing that if channel allocation and user associations are carried out

optimally, co-channel interference among the PBSs does not degrade the performance of the deployed HetNet. Moreover, they have shown that user association is very important and the current practice is far from optimal. In [68], Madan *et al.* have studied a joint user association and resource allocation problem in HetNets. They formulate a "semi-static resource allocation" problem with transmit power, user scheduling and association as problem variables. The resulting problem is a combinatorial problem of very large complexity (given as $(\emptyset(P^{RN}N^{M}))$ for P power levels, R sub-channels, N BSs and M users). Owing to this complexity, few heuristic algorithms with restrictions on either power control or on user associations are carried out. This formulation with a large set of variables as well as a global proportional fair throughput objective captures the true complexity of the throughput optimization problem, for HetNet deployment. However, the problem is intractable and thus there is a need for more tractable models.

Under relay deployment scenario, the benefits of joint resource allocation, routing, and scheduling have not been studied as well as under the wired deployments. In [96], Vishwanathan *et al.* study the throughput-optimal scheduling policy derived from the work of Tassulas and Ephremides [93] where they find the queue-aware optimal scheduling policy that maximizes the user throughput while maintaining that the system is stable under all arrival rates that can be stabilized. Via simulation, they show that relay deployment offers user throughput enhancements. In [84], Salem *et al.* also propose similar queue-aware scheduling policies and study the fairness property of such schemes. Numerous other recent works have studied some aspects of relay deployment, but often from a link level perspective. Despite this rich set of works, we believe that the following aspects are missing.

- A powerful and yet tractable framework that allows us to evaluate and benchmark different *resource allocation* schemes (including the co-channel deployment and optimal orthogonal deployments), *user association* rules (optimal user association, and other simple user association rules) and *transmission coordination mechanisms* (including the ones with and without coordinated scheduling) under a global throughput-based metric that incorporates a guaranteed fairness performance.

- Unifying models for the wired deployment and the relay deployment scenario so that

mixed model systems can be analyzed under the same framework.

2.6 Backhaul Limitations

Network operators see small cell (SC) backhauling as an immediate challenge for the successful deployment of HetNets, as discussed in [29] and [73]. The following three aspects have been identified as the reasons why SC backhaul can be limited:

1. **Economic consideration**: [73] presents some statistics showing how the ultra-dense deployment of small cells with low average number of users per BS means that the *cost of backhauling* for small cells becomes a significant part of the total Capital Expenditure (CAPEX), in some cases exceeding the cost of the small cell BS equipment. It is thus desirable that the backhauling cost for small cells is kept low. This economic consideration can often limit the capacity of the installed SC backhaul links. For example, a number of cheap solutions are being proposed, including ADSL [35], mesh networks [94], and even non-licensed microwave links [35].

2. **Need for a flexible infrastructure**: Besides economy, *flexibility* is also a key requirement as there will be numerous SCs added or moved frequently. Many industry reports like [29], [11] have acknowledged that fiber or copper infrastructures are often not flexible. This has given rise to different mobile backhauling solutions (e.g., [11], [17]).

3. **Physical constraints**: The third constraint is *physical*. A small cell might be at an inaccessible street furniture where bringing a fiber link can be infeasible. In [29], it is argued how a low capacity solution like non-line-of-sight (NLOS) wireless backhauling might be the only available option in such a case.

Macro base-station (MBS) backhaul limitations, on the other hand, are less likely to be a concern right now, since MBS backhauling is a small portion of the CAPEX [73], and thus can be well provisioned (with high capacity fiber). However, in the future, wireless networks are expected to operate with highly efficient wireless links (e.g., using massive

MIMO [18]) and on very high bandwidth spectrum (e.g., mmWave [77]). This will translate to a huge increase in traffic load on the backhaul. Moreover, many multi-cell architectures are emerging where signaling for coordination between BSs is done via the backhaul links (e.g., Joint Processing (JP) CoMP [58]), which increases the traffic load on the backhaul links as well as pose more stringent delay requirements. The deployment of cloud-RAN (C-RAN) [26] architectures is also going to put a lot of pressure on the MBS backhaul. Finally, in the future, the sharing of fiber among different operators, i.e., the virtualization of the backhaul, might result in capacity constraints. So, it is possible that MBS backhaul limitation might also become a concern for future networks.

In summary, small cell backhaul limitation has been identified as an immediate concern for the ultra-dense HetNets. MBS backhaul, on the other hand, is not as likely now to be a bottleneck, but can be a problem in the future.

Recently, the backhauling aspect of wireless networks has started to attract some attention in the research community. Its study can be broadly divided into two types: *provisioning-related* and *impact-related*.

Provisioning-related studies try to characterize the traffic load that a typical cellular deployment imposes on the backhaul network. For example, [53] looks at the LTE-Advanced HetNet deployment and characterizes the traffic load and delay requirements that it can impose in the presence of Joint Transmission based Coordinate Multipoint (CoMP) transmission.

Impact-related studies try to characterize how a limited backhauling can affect the system performance. [89] surveys the impact of limited backhaul on the link level performance due to the reduction in cooperation related capacity gains. Beyond link level performances, backhaul limitations can also impact the user scheduling process in HetNets. There are some studies in the literature that have studied the interplay between backhaul limitation and user scheduling. A number of these works including [101], [105] deal with coordination cluster formation as part of user scheduling decision and they try to make BS clusters so as to reduce the backhaul communication.

Backhaul limitation is not only relevant in multi-cell cooperative transmission. Even in HetNets without BS cooperation, limited backhauls can impact performances due to

the delay and/or the rate constraints. Even in the absence of cooperation, the total flow (user-plane traffic or data-plane traffic) to/from a BS is affected by the capacity of the backhaul network. Under such limitations, user scheduling decisions have to be made so as to maximize a given system performance by properly utilizing the constrained backhaul resource as well as the precious radio resource. A number of optimization formulations based on *network utility maximization* framework have been proposed in the literature for user scheduling in HetNets (e.g., [37], [68]) for different network-level performance metrics, in the absence of backhaul limitations. In this thesis, we build on these work and study the impact of backhaul limitations on user scheduling and user association.

Chapter 3

Flow-based Optimization Framework

Summary: In this chapter, we

- introduce the diverse set of scenarios/configurations arising in HetNets,
- present a unified view of the network, and
- formulate a flow-based framework for throughout optimization, under a given network "snapshot".

3.1 Introduction

In Chapter 1, we presented a number of important network processes, namely resource allocation, user scheduling and transmission coordination, and user association which impact the throughput performance of a HetNet greatly. We also discussed how these network processes have a complex interplay, which is not clearly understood. In this chapter, we propose a *unified framework* to study them and to enable their fair comparisons under different types of HetNet deployment scenarios. We will use the developed optimization framework in the next two chapters to study in details the performance of different options available.

Our framework is based on a flow-model, with a focus on the downlink. In that case, a flow[1] corresponds to a data stream from the network to a particular user. In the literature, optimization frameworks have been proposed for HetNets with small cell deployment in [36], and [68], which do not require a flow-based framework. However, the notion of flow helps us model the ON-OFF transmission coordination mechanism, as well as the relay deployment, which can be seen as a *two-hop* wireless network. Network-flow based modeling is a common approach taken for the study of wireless multi-hop networks [86, 66].

We formulate our optimization model for a system "snapshot". Under a given system "snapshot", the system parameters like the channel gains and the number of users are assumed to be fixed and known. We consider only the active users in the network and hence assume that there is one flow per user. Moreover, we assume that the users are greedy and hence want to maximize their individual flow-rates. Our framework allows us to configure the network parameters to allocate throughputs fairly and optimally to these flows. This is an offline-static model and thus it is intended to be used at the engineering and planning phase to compare many potential configurations and decide which ones to study further. Chapters 4, and 5 present detailed offline studies based on this framework.

3.2 System Overview

We consider a cellular network comprising a set of macro cells as shown in Fig. 3.1. Each macro cell, in addition to a centrally placed MBS, has X low-powered BSs making X small cells (SCs)[2] (see Fig. 3.1). These small cells are connected to the network core via backhaul links, either with wired backhaul links, or with wireless backhaul links. Recall that, in addition to the backhaul links, there are two other types of links, namely the *direct links (DL)* from the MBS to a UE, and the *access links (AL)* from an SC to a UE. The direct and the access links are collectively called the *user links*.

[1]This notion of flow is similar to the notion used in multi-commodity network-flow problems [56]. It is the same notion used in the existing literature on wireless networks in similar contexts (e.g., [86], [66]).

[2]Note that SCs are not always contained within a macro cell (for example, [85] considers SCs located over multiple macro cells).

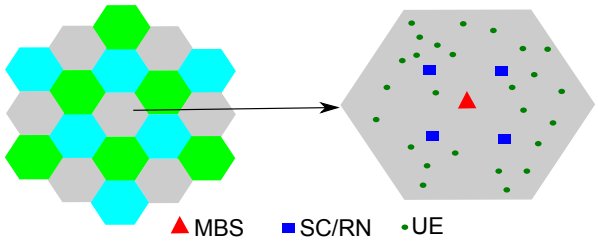


Figure 3.1: Multi-cell system and a HetNet

3.2.1 Scope

We consider each macro cellular area, with its MBS, X SCs, and N UEs as a standalone HetNet system, and we optimize a number of network processes (resource allocation, user association, user scheduling/coordination) within the scope of such a single macro cellular area only. However, a physical layer signal-to-interference-and-noise ratio formulation allows us to take into account the interference coming from nearby macro cells. Restricting our formulation to one macro cell level is justifiable since inter-cell resource allocations to different macro cells are usually carried out via planning. Also, to keep the complexity of network operation at a reasonable level, such decoupling can be desirable. However, decoupling the multi-cell system at a macro cell level can sometimes come with a penalty in throughput performance due to the inability to exploit some degrees of freedom (e.g., inter-cell coordination, load-balancing etc.).

3.2.2 Main features to be modeled

Below, we outline the features that we want to incorporate in our optimization framework.

Different types of backhaul links As already mentioned, a HetNet can comprise SCs with wired backhaul links, or SCs with wireless backhual links. We can identify the following two broad categories of deployment based on this property:

1. *Wired SC deployment*, sometimes referred to as *pico deployment*, refers to the network comprising of only SCs with wired backhaul links, and

2. *Relay deployment*, sometimes referred to as *wireless SC deployment*, refers to the network comprising of only relay nodes (with wireless backhaul links).

Multiple Bands The future HetNets are expected to be operating at diverse set of bands, with potentially different radio-access technologies. Some links (e.g., the user links) are expected to be operating in some specific bands, whereas some other links (e.g., the backhaul links) are expected to exploit new types of bands[3] (e.g., mmWave bands, in addition to LTE bands). In our study, we assume that the user links operate all on the same specific band, for example the LTE band. The backhaul links, if they are wireless, can operate either on the same band as the user links or on a different band with the same technology (say a different LTE band), or even on a completely different band and technology (say, a non-LTE band).

Different Resource Allocation Schemes The resource allocation scheme involves channel allocation and is crucial as it affects the interference between links, as well as can be used for proper resource provisioning of different links. A large number of channel allocation schemes can be envisaged with different complexity. For some deployment scenarios, resource allocation also involves the allocation of the total transmit power to different links, for example, an MBS allocating certain power to the direct links and certain power to the backhaul links if the backhaul links are wireless. We want to define our model so that it is able to incorporate many resource allocation schemes.

Different User Association Schemes A lot of user association schemes have been proposed in the literature, with different potential effects on the network performance. We want to formulate our model so that our model can incorporate the existing schemes, and also provide the optimal user association.

Transmission Coordination There has been a growing interest in coordinated transmissions in the HetNet context. We want to model ON-OFF transmission coordination between the BSs.

The complexity of a two-hop wireless network with multiple bands, channels, and potentially multiple types of radio-access technologies, and diverse choices of radio-resource

[3]Our usage of the term "band" is to refer to the band as well as the associated radio access technology.

management algorithms motivates us to formulate an optimization problem that can model these complexities and details into a unified framework.

3.3 General Optimization Model

Let 0 represent the MBS, and $\mathcal{P} = \{1, 2, \cdots, X\}$ be the set of SCs in the macro cell under study. Let $\mathcal{U}$ be the set of UEs corresponding to a random realization ω, which constitutes a snapshot of the UEs in the system. We focus on the downlink with a set of flows $\mathcal{F}$, where each flow f originates at the MBS (node 0) and terminates at one of the UEs u. The source and destination of a flow f are represented by f_s and f_d respectively. We take a full-buffer traffic model, and assume that the flows are greedy. We assume that the MBS has a fixed transmit power budget of P_M and each SC has a fixed transmit power budget of P_S. $\mathcal{N} = \{0\} \cup \mathcal{P} \cup \mathcal{U}$ represents the set of all nodes in the HetNet. Let $\mathcal{B} = \{1, 2, \cdots, S\}$ be the set of available bands[4]. Band $i \in \mathcal{B}$ is associated with its own technology (e.g., LTE) and has a number of channels $\mathcal{M}^{(i)}$, and a per-channel bandwidth b^i. We assume that at least one of them is LTE (say, Band 1) with $\mathcal{M}^{(1)}$ OFDM channels.

Remark 1. *Even though we will finally present an optimization model that can encompass both the wired and relay deployment scenarios (and even a mixed deployment),* ***we will first develop modeling concepts by restricting ourselves to the relay scenario****, where all the three types of links are wireless. Then, we will show how we can incorporate wired backhaul links into our model. So, until we explicitly discuss how we can incorporate wired links, the concepts discussed below apply to the relay scenario.*

3.3.1 Air interfaces

Each node is equipped with one or more air interfaces (AI)[5]. An air interface m is associated with one of the available bands given as $B(m) \in \mathcal{B}$. A node needs to have at least one AI for each band at which it is operating. A node can have more than one AIs operating

[4]Even though the model allows for more bands, we study $S = 1$ and 2.

[5]Note that, for a wired deployment, there is only one air interface per node.

on a given band. Having multiple AIs for the same band allows a node to have multiple simultaneous links on that band. More precisely, a node with x ($x > 1$) AIs on the same band could transmit simultaneously on up to x AIs in a given channel in that band[6].

In a given channel c in $\mathcal{M}^{(B(m))}$, at any given time, an AI m can either transmit or receive, but not simultaneously. We also assume that an AI can transmit in a set of channels of the associated band while receiving in an orthogonal set of channels of the same band. We assume that a node cannot transmit on channel c in one of its AIs while receiving on the same channel in another AI. Note that such considerations would be non-existent for the wired deployment. We also assume that a UE has only one AI in the LTE band. Each AI has an associated directivity. Let $D_m(\phi)$ be the directivity of AI m on direction ϕ. Directional AIs with very narrow beams can be used to avoid interference between AIs operating on the same set of channels. Different deployment scenarios can be identified based on the number and types of AIs. We show some examples in Table. 3.1. For example, in the case of wired SC deployment, each transmitter has 1 AI, where as for user-band relay deployment, Configuration 1 has 1 AI in the MBS where as Configuration 2 has $X + 1$ AIs in the MBS.

For the ease of exposition, we logically separate an AI into a transmit AI (tAI) and a receive AI (rAI). Note that such a distinction is not necessary in a pure wired deployment because in that case nodes are full duplex while in the wireless case, they are half-duplex. A node $n \in \mathcal{N}$ contains a set of transmit AIs (tAIs) T_n and a set of receive AIs (rAIs) R_n [7]. Let $G^c_{m,n}$ be the channel gain between AI m and n in channel c, which is determined by the realization ω. Let T and R be the set of all tAIs and rAIs in the HetNet, respectively.

Each tAI $m \in T_n$ is allocated a transmit power P_m such that $\sum_{m \in T_n} P_m \leq \overline{P}_n$ for all $n \in \mathcal{N}$, where $\overline{P}_n$ is the total power budget of node n (e.g., $\overline{P}_0 = P_M$). We focus only on the transmit power. Hence, no such power constraints exist for the rAIs.

Let $\mathcal{K}^m \subseteq \mathcal{M}^{(B(m))}$ be the set of channels allocated to AI m. Here, we discuss channel allocation in the most general form, and this will help us formulate a general model. It

[6]LTE AI capable of transmitting on multiple LTE bands (carrier aggregation) is viewed as two AIs on different bands.

[7]This distinction is merely logical and hence we have $|T_n| = |R_n|$.

Table 3.1: Different configurations based on the available air-interfaces

Wired SC Deployment

- MBS: 1 omni-directional AI in the LTE band, used for the direct links
- SC: 1 omni-directional AI in the LTE band, used for the access links
- UE: 1 omni-directional AI in the LTE band, used for both the direct and the access links

User-band Relay Deployment

- MBS:
  - Config. 1: 1 omni-directional AI in the LTE band, used for both the direct and the backhaul links
  - Config. 2: 1 omni-directional AI in the LTE band for the direct link and X directional AIs in the LTE band for the backhaul links
- SC: 1 omni-directional AI in the LTE band, used for both the backhaul and the access links
- UE: 1 omni-directional AI in the LTE band, used for both the direct and the access links

Dedicated-band Relay Deployment

- MBS:
  - Config. 1: 1 omni-directional AI in the LTE band for the direct links. and 1 omni-directional AI in a non-LTE band (e.g., LMDS) for the backhaul links
  - Config. 2: 1 omni-directional AI in the LTE band for the direct links and X directional AIs in a non-LTE band for the backhaul links
- SC:
  - Config. 1: 1 omni-directional AI in the LTE band used for the access links, and 1 omni-directional AI in a non-LTE band for the backhaul links
  - Config. 2: 1 omni-directional AI in the LTE band used for the access links, and X directional AIs in a non-LTE band for the backhaul links
- UE: 1 omni-directional AI in the LTE band, used for both the direct and the access links

should however be noted that there will be different constraints and limits on the set of feasible channel allocations, based on the exact channel allocation scheme being deployed. We will study a number of concrete channel allocation schemes later while studying different scenarios. A tAI m has to divide the transmit power P_m to its channels, allocating P_m^c to channel c, i.e.,

$$\sum_{c \in \mathcal{K}^m} P_m^c \leq P_m, \quad \forall m \in T_n, \forall n \in \mathcal{N}. \tag{3.1}$$

Power allocation of the total transmit power to individual subchannels can be seen as part of scheduling.

3.3.2 SINR, rate functions, and links

SINR $\gamma_{m,n}^c$ between tAI m and rAI n is defined as the ratio of the received signal power from m to n and the total interference and noise at node n, at channel c, i.e.,

$$\gamma_{m,n}^c = \frac{P_m^c \cdot G_{m,n}^c \cdot D_m\left(\phi_{m,n}\right) \cdot D_n\left(\phi_{n,m}\right)}{N_{B(m)} + I}$$

where I is the interference from nearby BSs transmitting on channel c, and N_i is the per-channel noise power in band i. Each band i is characterized by *rate functions* that map a per-channel SINR to communication rate. Let $\theta_{(m,n)}^{(i)}(.)$ represent the mapping from SINR γ between tAI m and rAI n in any channel $c \in \mathcal{M}^{(i)}$ to one of the supported rates $R = \theta_{(m,n)}^{(i)}(\gamma) \in \mathcal{R}_{(m,n)}^{(i)}$ where $\mathcal{R}_{(m,n)}^{(i)}$ is the set of supported rates between tAI m and rAI n in band i. The mapping function is determined by the available Modulation and Coding (MCS) schemes in the given band, between two AIs. Note that a band can have different rate functions for different pair of AIs (e.g., LTE backhaul links support up to 256 QAM whereas LTE user links support up to 64 QAM). $\mathcal{R}_{(m,n)}^{(i)}$ can be discrete and finite (in which case the mapping is called a discrete rate function) or it can be continuous (in which case the mapping is called a continuous rate function). In this case, $\mathcal{R}_{(m,n)}^{(i)}$ is an uncountable set. For a given rate R, we can define the minimum required SINR as follows: $\beta_{(m,n)}^{(i)}(R) = \min \gamma$ s.t. $\theta_{(m,n)}^{(i)}(\gamma) \geq R$. Next, we define two notions of *wireless* link: a *physical link* and a *logical link*. This distinction between a physical and a logical link

allows us to view scheduling as a process of activating a feasible set of logical links, to be defined later.

A *physical link* $\tilde{l}$ is defined as a tuple (m, n) where $m \in T$ and $n \in R$. Each HetNet is characterized by a set of adjacency indicators A. $A[j, i]$, if equal to 1 means tAI j can form a physical link with rAI i, if equal to 0 means otherwise. For example, a tAI of one RN cannot form a link with an rAI of another RN since we do not allow direct links. Also, two AIs in different bands cannot form a physical link. Adjacency indicators are a reflection of the network's topology. By introducing this notion, we have the ability to use our model for diverse topologies.

Given the adjacency indicators, the set of possible/potential physical links can be defined as follows:

$$\mathcal{L}^{Phy} = \{(m, n) : m \in T, n \in R \text{ s.t. } A[m, n] = 1\}.$$

For a given channel allocation $(\mathcal{K}^m)$ for all $m \in T \cup R$, let $\mathcal{K}(\tilde{l})$ represent the set of channels at which physical link $\tilde{l} = (m, n)$ operates, i.e., $\mathcal{K}(\tilde{l}) = \mathcal{K}^m \cap \mathcal{K}^n$.

3.3.3 Assumptions

Even though in the most general form, channel allocation as well as power allocation to AIs can be performed arbitrarily, we make the following assumptions to simplify the resulting optimization model.

A1. A physical link operates on all channels allocated to its tAI, and there is no partially overlapped channel allocation across links. i.e., if $\mathcal{K}(\tilde{l}_1) \cap \mathcal{K}(\tilde{l}_2) \neq \emptyset$ for some $\tilde{l}_1, \tilde{l}_2 \in \mathcal{L}^{Phy}$, then we have $\mathcal{K}(\tilde{l}_1) = \mathcal{K}(\tilde{l}_2)$.

A2. Transmit power allocated to a given physical link $\tilde{l} = (m, n)$, represented as $P(\tilde{l})$, equal to $P_{o(\tilde{l})}$, is equally divided among the allocated channels. So, if $p(\tilde{l})$ is the power per-channel in $\tilde{l}$, then we have $P^c_m = p(\tilde{l}) = P(\tilde{l})/|\mathcal{K}(\tilde{l})|$ for all $c \in \mathcal{K}(\tilde{l})$.

A3. Channel gains for different channels in a given physical link are equal, i.e., $G^c_{m,n} = G^{c'}_{m,n} = G_{m,n}$.

Since channels have identical channel gains, and an rAI of a physical link observes the same set of interferers with identical power for all allocated channels, these assumptions make all channels of a physical link identical in terms of SINR and supported rate.

We define a *logical link* l as a tuple $(o(l), d(l), R(l))$ where $o(l)$ is the tAI, $d(l)$ is the rAI, and $R(l)$ is its communication rate per channel. Each logical link is thus associated to a unique physical link. Let $\tilde{l} = (o(l), d(l))$ represent the physical link associated to logical link l. Given the set of all physical links, the set of all logical links, can be defined as follows.

$$\mathcal{L}^{All} = \{(o(l), d(l), R(l)) : \quad \tilde{l} = (o(l), d(l)) \in \mathcal{L}^{Phy}, \\ R(l) \in \mathcal{R}_{\tilde{l}}^{(B(o(l)))}\}. \tag{3.2}$$

User scheduling can be seen as a process to activate these logical links for a certain amount of time, as discussed next.

3.3.4 User scheduling and independent sets

In the most general form, the scheduling process in a multi-hop network with a given set of logical links $\mathcal{L}^{All}$ can be represented as the time-fraction β_s for which a given sub-set of logical links $s \subseteq \mathcal{L}^{All}$ is activated. We will call such a subset an independent set. Clearly, not every subset of logical links can be activated simultaneously. There are at least three fundamental limits:
1) Two links can be activated simultaneously on the same set of channels only if they do not share a tAI or an rAI.
2) *SINR feasibility constraints*: When a number of logical links are activated simultaneously, the SINR at each rAI should be large enough so that the signals can be decoded successfully.
3) *Half-duplex communication capability*: Depending on whether a tAI and an rAI of a given node are allocated the same set of channels, there is limit on whether a tAI can transmit while an rAI in the same node is receiving. For our cellular HetNet in downlink, RNs are the only nodes that could use both a tAI and an rAI. Thus, this limit is associated with the RNs only.

We are now ready to formally define an independent set (ISet) as follows.

Definition 1. *For a given channel allocation ($\mathcal{K}(\tilde{l})$) and a given power allocation per channel ($p(\tilde{l})$), $s \subseteq \mathcal{L}^{All}$ is an ISet if the following conditions are satisfied.*

$$\forall l = (m, n, R_l) \in s :$$

$$\frac{p(\tilde{l}) \cdot G_{m,n} \cdot D_m(\phi_{m,n}) \cdot D_n(\phi_{n,m})}{N_{B(m)} + \tilde{I}_n + I_l(s)} \geq \beta_{\tilde{l}}^{(B(m))}(R_l). \tag{3.3}$$

$$\forall l, l' \in s \ s.t. \ l \neq l' : \quad o(l) \neq o(l') \ and \ d(l) \neq d(l'). \tag{3.4}$$

$$\forall l, l' \in s \ s.t. \ l \neq l' \ and \ \mathcal{K}(\tilde{l}) = \mathcal{K}(\tilde{l'}) :$$

$$\sum_{n \in \mathcal{N}} \mathbf{1}_{\{o(l) \in T_n\}} \mathbf{1}_{\{d(l') \in R_n\}} = 0. \tag{3.5}$$

where $I_l(s)$ is given as

$$\sum_{\substack{l' \in s: \\ l' \neq l, \\ \mathcal{K}(\tilde{l}) = \mathcal{K}(\tilde{l'})}} p(\tilde{l'}) \cdot G_{o(l'),n} \cdot D_{o(l')}\left(\phi_{o(l'),n}\right) \cdot D_n\left(\phi_{n,o(l')}\right).$$

$\phi_{m,n}$ is the angle of AI n from AI m. N_i is the noise power per channel in band i. $\tilde{I}_n$ is the interference from nearby macro cells to rAI n (determined by the reuse pattern).

(3.3) guarantees that the *SINR feasibility constraints* are satisfied for each logical link. (3.5) guarantees that the *half-duplex communication constraints* of the nodes are satisfied, so that a node cannot activate a tAI if one of its rAI is receiving in the same set of channels. This constraint represents a rather important concept, that is associated with the ability to have a simultaneous transmission and reception at a relay node. LTE-A standard puts an emphasis on this distinction and introduces the notion of an *in-band* relay deployment and an *out-of-band* relay deployment. With respect to channel c, we can call RN j to be an *in-band* relay if c is allocated to both the tAI as well as the rAI of this relay. In this case, the *half-duplex constraint* affects the definition of an ISet. Our generalization, in terms of ISets, can model many more scenarios, some of which we will present later.

Let $\mathcal{I}^{All}$ be the set of all ISets $s \subseteq \mathcal{L}^{All}$. If $\mathcal{R}_{(m,n)}^{(i)}$ is continuous (i.e., there exists a continuous rate function) for some m and n in band $i = B(m)$, then $\mathcal{L}^{All}$ contains infinitely

many links and hence it is not possible to compute $\mathcal{I}^{All}$. In order to overcome this difficulty, we define the notion of dominant ISet as follows.

Definition 2. *Let $L_{Phy}(s) = \{(o(l), d(l)) : l \in s\}$ be the set of physical links in ISet s. Then, $s \in \mathcal{I}^{All}$ dominates $s' \in \mathcal{I}^{All}$ (written as $s \geq s'$) if $L_{Phy}(s) = L_{Phy}(s')$ and $R(l) \geq R(l')$ whenever $\tilde{l} = \tilde{l}'$ for all $l \in s$ and $l' \in s'$.*

It can be shown that, for a given channel and power allocation, we can find one ISet $S_{max}[v]$ such that $L_{Phy}(S_{max}[v]) = v$ and that dominates all ISets s' with the same set of physical links v.

$$S_{max}[v] = s \in \mathcal{I}^{All} \text{ s.t. } s \geq s', \forall s' \in \mathcal{I}^{All}, L_{Phy}(s') = v.$$

Then, from the point of view of throughput optimization, we can easily show that it is sufficient to consider only the set of dominant ISets $\mathcal{I} \subset \mathcal{I}^{All}$, which is defined as follows.

$$\mathcal{I} = \{S_{max}[v] : v \subseteq \mathcal{L}^{Phy}\}.$$

Note that $\mathcal{I}$ (unlike $\mathcal{I}^{All}$) is finite even if $\mathcal{R}^{(i)}_{(m,n)}$ is a continuous set. Then the set of relevant logical links can also be reduced to a finite set: $\mathcal{L} = \{l \in s : s \in \mathcal{I}\}$.

ON-OFF transmission coordination

If all of the ISets $s \in \mathcal{I}$ defined as above were allowed to be scheduled, it means that we are implicitly assuming that the MBSs and SCs perform a transmission coordination where a BS can improve the transmission rate of a physical link in another BS by occasionally pausing its own transmission. We call this the *ON-OFF transmission coordination* among the BSs. Let $\mathcal{I}^O = \mathcal{I}$ be the set of all ISets as defined in Def. 1. At any given time, only one ISet from each $\mathcal{I}^O$ can be activated. Then, scheduling problem involves finding the values of β_s that satisfy the following constraints.

$$\sum_{i \in \mathcal{I}^0} \beta_s \leq 1 \tag{3.6}$$

Remark 2. *In the LTE-A context, this can be seen as a generalization of LTE-A proposal of almost blank sub-frame (ABSF) during which the MBS does not schedule on any data channels. In other words, all SCs always schedule their transmissions whereas the MBS does not schedule its transmission for a certain proportion of time α (say). Clearly, by admitting only a sub-set of $\mathcal{I}^0$ such that the above condition is satisfied (meaning each SC is necessarily transmitting on its tAI all the time), our approach can easily model ABSF.*

No coordination (NC)

ON-OFF coordination involves a large set of independent sets (whose cardinality grows exponentially with the number of AIs). Such complexity might not always be desirable. In another extreme, we could employ no coordination at all. Under *no coordination* (NC), all transmit AIs in the network would stay scheduled all the time, as long as it is possible to do so. The only exception would be the case when a backhaul link and an access link in RN j are both operating on the same set of channels. In such a case, tAI of RN j has to be turned-off when the backhaul link to j is active. Such restrictions do not appear when $m \in T_j$ and $n \in R_j$ are allocated orthogonal sets of channels. By restricting the set of ISets to a subset of $\mathcal{I}^O$ that satisfies this condition, we can define the set of ISets $\mathcal{I}^{NC}$ for NC.

Incorporating wired backhaul links into the model

So far, the notions of physical as well as logical links, and the independent sets dealt with the relay cases, i.e., the cases with only the wireless links. We have not considered how one or more wired backhaul links can be incorporated into our general notions of links and ISets. Without incorporating these wired links, we will not be able to use our model for the scenarios with wired backhaul links. A wired link is different from a wireless link in the following ways:

- A wired backhaul link to SC j has a fixed capacity C_j, which is analogous to the link-rate in the wireless case.

- A wired link is always feasible and thus can be included in any independent set.

Let a wired backhaul link to SC j be represented as $(o(l), d(l), R(l))$ where $o(l) = 0$, $d(l) = j$, and $R(l) = C_j$. Since the capacity is independent of channel allocation, we assume that $|\mathcal{K}(\tilde{l})| = 1$ for all wired links l.

In order to allow for an identical treatment of the two types of links in our formulation, we will assume that 0 is a dummy tAI at the MBS and j is a dummy rAI at SC j, and thus the set of tAIs at the MBS is updated to include tAI 0, and the set of rAIs at SC j is updated to include rAI j. Let $\tilde{T}_j$ and $\tilde{R}_j$ respectively represent the set of tAIs and rAIs at node j after incorporating the dummy AIs.

Let $\mathcal{L}_{wired} \subseteq \{(0, j, C_j) : j \in \mathcal{P}\}$ be the set of all wired backhaul links. Then, we can expand the set of all relevant logical links in the network to

$$\tilde{\mathcal{L}} = \mathcal{L} \cup \mathcal{L}_{wired} \tag{3.7}$$

Now, if $\mathcal{I}$ was the set of ISets defined purely with the wireless links as before, the set of ISets after incorporating the wired links can be defined as follows.

$$\tilde{\mathcal{I}} = \{s \cup \mathcal{L}_{wired} : \quad s \in \mathcal{I}\} \tag{3.8}$$

This will allow us to consider either wired, or relay, or a mixed deployment where some SCs are pico BSs, and some others are RNs.

3.3.5 User association as flow routing: multi-association

User association determines whether user i is associated to BS j or not. We incorporate user association into our framework by introducing the "routing variables" x_l^f which represents the amount of flow f routed through logical link l. Typically a user associates to exactly one BS. Such a *single-association* would then impose single-path routing constraints on the routing variables which would thus result in an Integer Problem (IP), which is very hard to solve (since the problem that we will formulate later is non-linear). While formulating our optimization model, for tractability, we make the assumption that a user can associate to

multiple BSs. Clearly, such a *multi-association* can be modeled under a multipath routing framework. Such an assumption yields a much more tractable model and the solution based on optimal multipath routing is an upper bound to the optimal single-association solution[8]. It is however unclear a priori if such an upper-bound is tight. We will later show that it is indeed the case.

3.3.6 Problem formulation

Our aim is to obtain proportional fair throughput allocations $\{\lambda_f\}_{f\in\mathcal{F}}$ under optimal scheduling/transmission-coordination and flow-routing/user-association within a macro cell coverage. Given a set of nodes $\mathcal{N}$, a set of flows $\mathcal{F}$, a set of bands $\mathcal{B}$, the associated channels and the rate functions, channel gains $G_{m,n}$ between any two AIs, a set of tAIs $\{T_n\}_{n\in\mathcal{N}}$ and rAIs $\{R_n\}_{n\in\mathcal{N}}$, their directivity properties $\{D_m(\phi)\}_{m\in T\cup R}$, a set of wired links $\mathcal{L}_{wired}$, the adjacency indicators $A[m,n]$, and **given the channel allocations $\mathcal{K}(\tilde{l})$, and the power allocations $P(\tilde{l})$ for all physical links**, the set of ISets $\tilde{\mathcal{I}}$ can be constructed a priori. Our problem of proportional fair throughput allocation under a joint optimal scheduling/coordination and flow-routing/user-association, can then be stated as follows.

[8]The newer cellular standards (e.g., LTE-A) are considering the possibility of allowing a UE to be associated to more than one BS at the same time, in which case, our assumption of multi-association is applicable.

$$
\begin{aligned}
[\mathbf{P_{Joint}}(\boldsymbol{K},\boldsymbol{P})] \max_{\boldsymbol{\lambda},\boldsymbol{x},\boldsymbol{\beta}} & \sum_{f\in\mathcal{F}} \log(\lambda_f) \\
& \sum_{m\in\tilde{T}_n}\left(\sum_{l\in\tilde{\mathcal{L}}:o(l)\in m} x_l^f\right) - \sum_{m\in\tilde{R}_n}\left(\sum_{l\in\tilde{\mathcal{L}}:d(l)\in m} x_l^f\right) \\
& \qquad = \lambda_f \mathbf{1}_{\{n=f_s\}} - \lambda_f \mathbf{1}_{\{n=f_d\}}, \forall n\in\mathcal{N}, \forall f\in\mathcal{F} \qquad (3.9) \\
& \sum_{f\in\mathcal{F}} x_l^f \le |\mathcal{K}(\tilde{l})| \sum_{s\in\tilde{\mathcal{I}}:l\in s} \beta_s R(l), \quad \forall l\in\tilde{\mathcal{L}} \qquad (3.10) \\
& \sum_{s\in\tilde{\mathcal{I}}} \beta_s \le 1 \qquad (3.11) \\
& \beta_s \ge 0, x_l^f \ge 0, \lambda_f \ge 0, \quad \forall s\in\tilde{\mathcal{I}}, \forall f\in\mathcal{F}, \forall l\in\tilde{\mathcal{L}}
\end{aligned}
$$

where $\boldsymbol{\lambda}$ is a tuple containing the throughput variables λ_f, $\boldsymbol{x}$ is a tuple containing the flow-association variables x_l^f, and $\boldsymbol{\beta}$ is a tuple containing the user scheduling variables β_s. (3.9) is the flow-conservation constraint. (3.10) is the capacity constraint that limits the total amount of flow in a link l. (3.11) is the scheduling constraint. The above problem solves for optimal user scheduling (possibly with transmission coordination), and user association/flow-routing when channel and power allocations for all physical links are given.

Remark 3. *The problem* $[\mathbf{P_{Joint}}(\boldsymbol{K},\boldsymbol{P})]$ *is for a given realization* ω*. This explicit dependence is not mentioned, but is to be understood.*

Remark 4. *The problem is parameterized with* $(\boldsymbol{K},\boldsymbol{P})$*.* $\boldsymbol{K}$ *represents the tuple of the channel allocation variables* $\mathcal{K}(\tilde{l})$*, and* $\boldsymbol{P}$ *is the tuple of the power allocation variables* $P(\tilde{l})$*. Let us call them the model parameters. In order to solve the model, these parameters have to be chosen and fixed. A joint optimal resource allocation, user scheduling/transmission coordination, and user association can thus be obtained by solving a set of parameterized problems* [9] *to find the optimal model parameters:*

$$\arg\max_{\boldsymbol{K},\boldsymbol{P}} \mathbf{P_{Joint}}(\boldsymbol{K},\boldsymbol{P})$$

[9]Note that, we use the symbol $\mathbf{A}$ to represent the optimal value (i.e., the value of the objective function when the variables are chosen optimally) of problem $[\mathbf{A}]$.

The set of possible choices on the model parameters will depend on the deployed resource allocation scheme. For example, under co-channel deployment (introduced in the next chapter), there are no choices, where as under orthogonal deployment (also introduced in the next chapter), there are a discrete number of choices.

Maximizing the objective $\sum_{f\in\mathcal{F}} \log(\lambda_f)$ is known to yield a proportional fair throughput allocation [55]. A PF throughput allocation is known to maximize the geometric mean (GM) throughput $\left(\prod_{f\in\mathcal{F}} \lambda_f\right)^{1/|\mathcal{F}|}$ and hence we will use the GM throughput as the performance metric. We chose proportional fairness as a metric as it is known to strike a good trade-off between fairness and efficiency. The above formulation will be used as the main tool to perform studies in the next two chapters (Chapters 4 and 5). We will study a more general objective function, but with a restricted (non-flow based) model while presenting an in-depth study on user scheduling and user association under limited backhaul capacities in Chapters 6 and 8.

Chapter 4

Detailed Study: Wired SC Deployment

Summary: In this chapter, we use the optimization formulation obtained in Chapter 3 to study different scenarios of wired SC deployment, and present insights on the interplay between resource allocation, transmission coordination, and user association.

4.1 Introduction

In this chapter, we will use our framework introduced in the previous chapter to study the performance of HetNet under different choices of resource allocation, transmission coordination, and user association schemes, by restricting ourselves to the wired SC deployment scenarios. In the next chapter, we will present a detailed study for the relay deployment scenarios.

The wired SC deployment scenario corresponds to one LTE band ($S = 1$) with a total of M OFDM subchannels available for the given HetNet. The direct and the access links operate on this band. The MBS as well as each of the SCs have one LTE omni-directional

AI, used for the direct and the access links respectively. Let, o_0 represent the tAI at the MBS and o_j represent the tAI at the SC j. Also, UE i has one omni-directional AI for the reception on both the direct and the access links. Let i refer to this rAI. The set of wired backhaul links is given as $\mathcal{L}_{wired} = \{(0, j, C_j) : j \in \mathcal{P}\}$. We will assume that the backhaul capacities C_j are very large.

Given this set-up, we can solve $[\mathbf{P_{Joint}}(\boldsymbol{K}, \boldsymbol{P})]$ as long as the following parameters are given:

- **Channel Allocation** $\mathcal{K}(\tilde{l})$ for all $\tilde{l} \in \mathbf{L}_D \cup \mathbf{L}_A$ where $\mathbf{L}_D = \{(o_0, i) : i \in \mathcal{U}\}$ is the set of direct links and $\mathbf{L}_A = \{(o_j, i) : j \in \mathcal{P}, i \in \mathcal{U}\}$ is the set of access links
- Power Allocation $P(\tilde{l})$ for $\tilde{l} \in \mathbf{L}_D \cup \mathbf{L}_A$

We consider three types of RA schemes: co-channel deployment (CCD), orthogonal deployment (OD), and partially shared deployment (PSD), that dictate how the M subchannels are allocated to the direct and the access links. As mentioned earlier in Section 3.3.3, we consider a simple power allocation strategy, based on equal sharing of available power. Note that for each RA scheme, determining the channel allocation $\{\mathcal{K}(\tilde{l})\}$ and power allocation $\{P(\tilde{l})\}$ requires a number of *scheme-specific* parameters to be fixed. Let $\mathcal{V}_s$ represent the set of these parameters for RA scheme s. Then, by fine tuning $\mathcal{V}_s$, we can obtain the optimal performance for the corresponding RA scheme s, under joint optimal user scheduling/transmission coordination, and user association, as follows:

$$\Lambda_i^* = \max_{\mathcal{V}_s} \Lambda_s^*(\mathcal{V}_s)$$

where $\Lambda_s^*(\mathcal{V}_s)$ is the optimal value of $[\mathbf{P_{Joint}}(\boldsymbol{K}, \boldsymbol{P})]$ for RA scheme s when the scheme-specific parameter set is set to $\mathcal{V}_s$. Recall that the optimization model $[\mathbf{P_{Joint}}(\boldsymbol{K}, \boldsymbol{P})]$ allows us to study different transmission coordination schemes. We will focus on *ON-OFF* coordination (O) and no coordination (NC). Also, recall that our framework allows us to study the performance of, not only the optimal user association, but also a number of simple user association rules. By obtaining results for realistic networks, we will provide a number of interesting engineering insights[1]:

[1]Some of the results in this chapter were published in [39] and [42].

- The upper bounds obtained under the multi-BS association assumption are tight and hence allowing a user to associate to more than one BS will not offer significant performance gains.
- PSD/OD perform very well even in the absence of sophisticated transmission coordination whereas transmission coordination is essential for the satisfactory performance of CCD.
- A simple *small cell first* user association rule performs well even with ON-OFF TC if properly tuned. Its effectiveness under *no coordination* was shown earlier in [36].

The rest of this chapter is organized as follows. Section 4.2 presents the details of the three resource allocation schemes. In Section 4.3, we present different configurations based on the choice of RA and TC. After that, we present a number of user association rules. In Section 4.5, we present the numerical results before concluding the study.

4.2 Resource Allocation Schemes

Under wired deployment, RA affects two types of links, the *direct* and the *access* links[2]. We study the following three RA schemes.

Co-channel deployment (CCD)

Under CCD, all wireless (direct and access) links operate over all the M subchannels. Thus, the direct and the access links interfere with each other. Also, there is no resource allocation specific parameter to configure.

[2]Under relay deployment however, an RA affects all three types of links (i.e., the *backhaul* links in addition to the *direct* and the *access* links). This in effect requires more complex considerations to be taken while dealing with the relay deployment, as will be evident in the next chapter.

Orthogonal deployment (OD)

OD corresponds to *channel splitting* where a set of K subchannels is allocated for SC operation (i.e., the access links) and the remaining set of $M-K$ subchannels is dedicated for MBS operation (i.e., the direct links). Such an orthogonal set of frequencies at the two tiers allows for low interference operation. Additionally, a frequency reuse pattern could be used among the SCs so as to guarantee low interference at the SC-tier also. However, [36] has shown that if other network processes are chosen optimally, an aggressive full frequency reuse performs better than more conservative frequency reuse patterns. Accordingly, in our work, we consider that all K subchannels are used by each SC. Under this RA, K is a parameter to be configured and we call it the *channel split parameter*.

More formally, with $\mathcal{P} = \{1, 2, \cdots, X\}$, $K_{\tilde{l}}$, the number of subchannels on which wireless physical link $\tilde{l}$ can operate, is given as follows[3].

$$K_{\tilde{l}} = (M-K)\mathbf{1}_{\{o(\tilde{l})=o_0\}} + K\mathbf{1}_{\{o(\tilde{l})\in\{o_j : j\in\mathcal{P}\}\}} \tag{4.1}$$

where $\mathbf{1}_{\{A\}}$ is an indicator function evaluating to 1 if statement A is true, and 0 otherwise.

Partially shared deployment (PSD)

Under PSD, K subchannels are allocated to each SC and the remaining $M-K$ subchannels are dedicated to the MBS, as in OD. However, the MBS can also transmit in the K subchannels allocated to the SCs, albeit at a lower power. Clearly, OD can be viewed as a special case of PSD when the MBS does not transmit at the K subchannels.

For our modeling convenience, we introduce a dummy BS corresponding to the MBS when it is transmitting on the K shared subchannels. This dummy BS is represented as $0'$ and can be viewed as an additional SC that is connected to the MBS (node 0) with a wired link $(0, 0', C'_0)$ of infinite capacity, i.e., $C_{0'} = \infty$. Clearly, the set of SCs under PSD includes $X+1$ elements, i.e., $\mathcal{P}' = \{1, 2, \cdots, X\} \cup \{0'\}$. The channel gains of the dummy BS correspond to the channel gains of the MBS, i.e., $G_{0',n} = G_{0,n}$. Also, the set of access links has to be redefined as $\mathbf{L}'_A = \{(o_j, i) : j \in \mathcal{P}', i \in \mathcal{U}\}$.

[3]For relay deployment, OD can take multiple forms as we will explain in the next chapter.

4.2.1 Power allocation

MBS can transmit at the maximum total power of P_M and each SC can transmit at the maximum total power of P_S. Under CCD, the power per subchannel is chosen by assigning equal power to all of the allocated subchannels. Hence, it is simply given by,

$$p(\tilde{l}) = \frac{P_M}{M}; \qquad p(\tilde{l}) = P = \frac{P_S}{M}, \qquad \forall \tilde{l} \in \mathbf{L}_D \cup \mathbf{L}_A \tag{4.2}$$

Recall, $p(\tilde{l})$ represents the power per subchannel for physical link $\tilde{l}$.

Under PSD, MBS allocates P' for transmission on the shared K subchannels and the remaining power $(P_M - P')$ for transmission on the dedicated $M - K$ subchannels. The power per subchannel for different physical links is simply given by,

$$p(\tilde{l}) = \begin{cases} \frac{P'}{K} & \text{if } o(\tilde{l}) = o_{0'} \\ \frac{P_s}{K} & \text{if } o(\tilde{l}) \in \{o_j : j \in \mathcal{P}\} \end{cases} \qquad \forall \tilde{l} \in \mathbf{L'}_A \tag{4.3}$$

$$p(\tilde{l}) = \left(\frac{P_M - P'}{M - K}\right) \quad \forall \tilde{l} \in \mathbf{L}_D \tag{4.4}$$

Recall that we decomposed MBS into node 0' (resp. node 0) transmitting on K (resp. $M - K$) subchannels. Clearly, OD corresponds to PSD with $P' = 0$.

4.3 Configurations

We call a configuration the exact choice of resource allocation and the transmission coordination mechanism. Generically, ***[X-Y]*** denotes a configuration where X is the RA (either CCD, OD, or PSD), and Y is the type of employed transmission coordination mechanism (either O for ON-OFF TC or NC for no coordination). For example, ***[CCD-O]*** represents a configuration under CCD with ON-OFF TC. For each configuration, UA is either performed optimally or is based on some simple rules that are defined next.

4.4 User Association

Our optimization model [$\mathbf{P_{Joint}}(\boldsymbol{K}, \boldsymbol{P})$] can yield optimal user association under multi-association assumption. UA is captured by the flow variables $\{x_l^f\}$. We can also incorporate other user association schemes in the model. We study three different simple but sub-otpimal user association schemes, which are based on simple rules that a UE can use to perform its association decision.

1. *Best-SINR*: In this scheme, an UE associates to the BS that offers the highest SINR. This approach had been used often in homogeneous settings. In HetNet case, though, it is shown to perform poorly mainly due to the *power disparity* between the MBS and the SCs, thereby resulting in overloaded MBS [36].

2. *Range Extension* (RE): In RE, the problem of power disparity is addressed to some extent by associating a user to the BS with the smallest path-loss [57].

3. Small-cell First (SCF(δ)): UE i associates to small cell $j \in \mathcal{P}$ if j provides the best per-subchannel SINR γ_{ji} among all SCs and if this SINR is greater than δ, i.e., if $j = \arg\max_{j' \in \mathcal{P}} \gamma_{j'i}$ and if $\gamma_{ji} > \delta$. If no such small cell j exists, UE i goes to BS $\tilde{j}$ that provides the best SINR, i.e., $\tilde{j} = \arg\max_{j' \in \{0\} \cup \mathcal{P}} \gamma_{j'i}$ [36]. δ is the UA configuration parameter that can be adjusted to change the relative association bias between the MBS and the SCs.

All of these three rules are simple in the sense that they do not involve any real-time load-balancing and are easy to calculate (each UE can do it itself). They also provide feasible single-association solutions and thus provide the lower bounds on the optimal single-association solution. These UA rules can be applied to our earlier problem by translating the association structure into the routing variables (x_l^f) of our model. As an example, let UE i associates to BS j under the given association rule. Then the corresponding flow routing variable x_l^f (where flow f is the downlink flow to user i and thus $f_d = i$) will be 0 for all wireless links l that do not belong to BS j. Once x_l^f captures the user association structure imposed by this rule, we can easily compute the other parameters by using our problem formulation [$\mathbf{P_{Joint}}(\boldsymbol{K}, \boldsymbol{P})$].

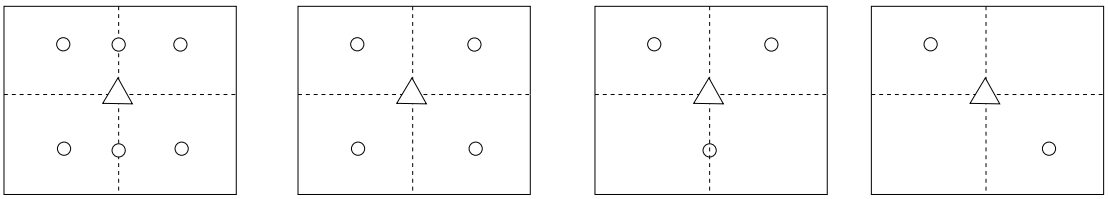


Figure 4.1: X SCs placed in a grid layout on a macro coverage of a $500m \times 500m$ square

Table 4.1: Path-loss model

Transmitter	Link (j,i)	Path-loss at the medium ($\phi_{j,i}$)	Antenna gain (AG_j)	Cable losses (ζ_j)
MBS	$(0,i)$	$128.1 + 37.6 \log_{10}\left(\frac{d_{0i}}{1000}\right), d_{0i} \geq 35m$	15	20
SC	$(j,i) : j \in \mathcal{P}$	$140.7 + 36.7 \log_{10}\left(\frac{d}{1000}\right)(dB), d_{ji} \geq 10m$	5	20
Total path-loss ($L_{j,i}$) (dB)				
$L_{j,i} = \phi_{j,i} + \zeta_j - AG_j$				

Studying these simple UA rules serves us with two purposes. The first is to obtain lower-bounds so that we can validate our upper-bounds. The second is to understand how these simple UA rules perform. In the absence of transmission coordination, [36] already shows that SCF(δ) works well. Our study allows us to see whether this observation extends to the case of ON-OFF TC as well.

4.5 Numerical Results

We consider a $500m \times 500m$ square as the user deployment area with an MBS placed at the center. We consider scenarios with $X = 2$, 3, 4 and 6 SCs deployed as shown in Fig. 4.1. The path loss $L_{j,i}$ for the transmitter-receiver pair (j,i) separated by a distance d_{ji} (m) is given in Table 4.1, together with the appropriate values of antenna-gains and miscellaneous losses. This is a path-loss model recommended by 3GPP [6]. We further apply a log-normal shadowing with zero mean and standard deviation of 8 dB to obtain the random path-loss $\overline{L}$, i.e., $\overline{L}_{j,i} = L_{j,i} + \mathbf{N}(0,8)$ where $\mathbf{N}(\mu,\sigma)$ is a normal random variable with mean μ and standard deviation σ. The channel gains can then be obtained as $G_{j,i} = 10^{-\frac{\overline{L}_{j,i}}{10}}$. We take $P_M = 46dBm$, a noise power of $N_0 = -112.4245dBm$ per subchannel (corresponding to

Table 4.2: Available rates and the corresponding SNR thresholds

Threshold SNR (dB)	-6.5	-4	-2.6	-1	1	3	6.6	10	11.4	11.8	13	13.8	15.6	16.8	17.6
Efficiency (η)	0.15	0.23	0.38	0.60	0.88	1.18	1.48	1.91	2.41	2.73	3.32	3.9	4.52	5.12	5.55

a subchannel bandwidth of $b = 180KHz$, with a noise figure of 9 dB), and $M = 100$ subchannels. While computing SINR, we do not consider interference from nearby macro cells. We have improved this limitation in the subsequent chapters.

We consider an adaptive MCS with 15 discrete rates, used in LTE. The rates (efficiencies) and the corresponding required threshold SNRs are listed in Table 4.2. The efficiency η is related to rate R as $R = \eta \frac{n_{sc} n_{ts}}{T_{symbol}}$ where n_{sc} is the number of sub-carriers per OFDM symbol, n_{ts} is the number of OFDM symbols in one subframe, and T_{symbol} is the duration of one OFDM subframe. For LTE, we have $n_{sc} = 12$, $n_{ts} = 14$, and $T_{symbol} = 1ms$.

We assume that the wired MBS-SC backhaul links are not the bottleneck and thus we consider these wired links to be of infinite capacity[4]. For each scenario of X SCs and N UEs, a network realization is obtained by generating N uniformly distributed random user positions in the deployment area. For each X and N, we have studied 100 such random realizations of the network. We obtain the numerical results by solving the convex optimization problem $[\mathbf{P_{Joint}}(\boldsymbol{K}, \boldsymbol{P})]$ formulated earlier to global maximum for each realization by using the commercial solver, Minos$^{\text{TM}}$[4]. PF is known to maximize the geometric mean (GM) of the throughput of the users, given as $\left(\prod_{f \in \mathcal{F}} \lambda_f^*\right)^{\frac{1}{N}}$. Thus, we take the GM throughput as the performance metric to compare different configurations. As mentioned earlier, CCD does not have any channel allocation parameter, whereas OD has the channel split parameter K. For each realization, the performance for OD is computed for the optimal value of channel-split parameter $K^* \in \{0, 1, \cdots, M\}$. Under PSD, in addition to the channel split parameter K, power P' also needs to be computed optimally, in order to obtain the best possible performance. However, solving for optimal P' is a difficult problem and our models developed so far can obtain the GM throughput only when P' is given. In order to obtain good performance gains for PSD, we coarsely tune P' by selecting the best power from the set of power choices from -10 dBm to 30 dBm at 1 dBm interval. All the results shown for PSD are obtained for the best P' from this set,

[4]We will later present studies that focus on the impact of backhaul capacity limitations.

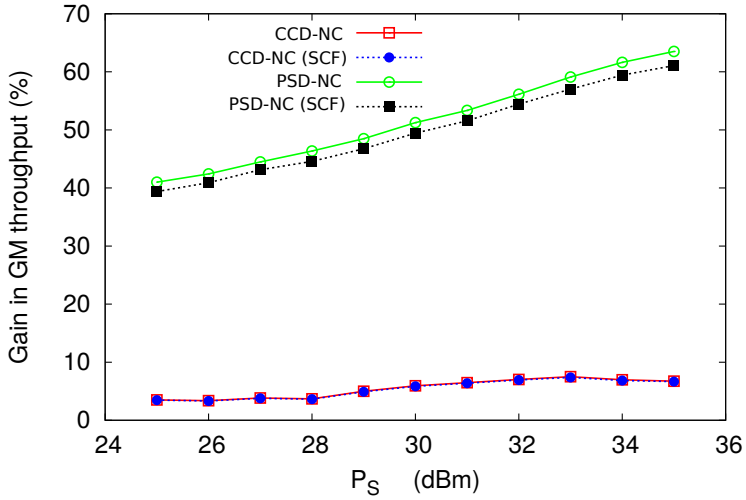
and for the optimal choice of K. Recall that, OD corresponds to PSD with $P' = 0$ W.

Throughput gain for each configuration on a particular realization is computed over the case when SCs are not deployed. This *MBS-only* configuration is thus a benchmark. For a particular realization i, the throughput gain obtained by configuration Y is given by $\mathcal{G}_Y(i) = 100 \times \frac{\chi_Y^{GM}(i) - \chi_0^{GM}(i)}{\chi_0^{GM}(i)}$ where $\chi_Y^{GM}(i)$ is the GM throughput of realization i under configuration Y. $Y = 0$ corresponds to the MBS only configuration. In order to characterize the average gain in throughput performance of each configuration, we obtained the average gain in GM throughput over the random realizations.

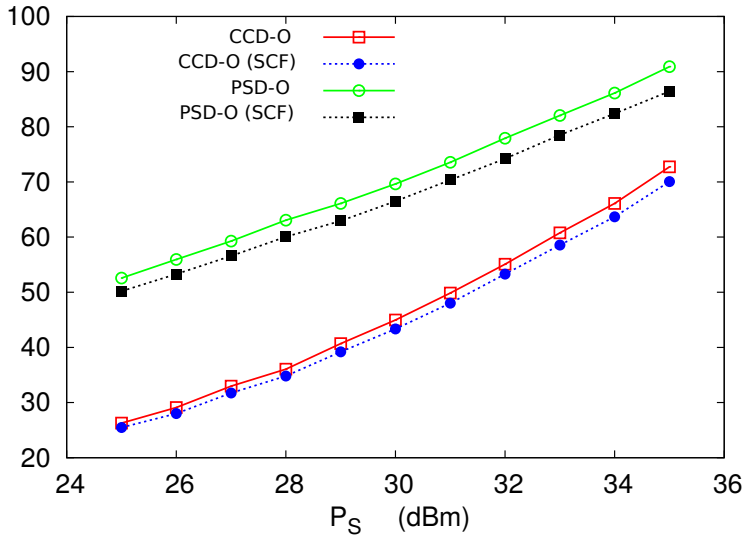
4.5.1 Validation of the upper bounds

Before continuing with the performance comparison of different configurations, we validate our assumption of multi-association with the help of a feasible single-association solution as discussed below.

As will be discussed later, SCF(δ) yields the best performance of the three UA schemes that we studied. Hence, we present the results obtained under SCF(δ) and show these results along-side the results obtained with optimal multi-association, averaged over the 100 realizations. In order to get the "best" lower bound, for each configuration and each realization, we select the value of δ from the set of SINR thresholds specified in Table 4.2 that provides the best performance in terms of the GM throughput. In Fig. 4.2, we plot the average gain in GM throughput for different configurations with a fine-tuned SCF(δ) as well as with the optimal multi-association for the scenario with $X = 4$ SCs and $N = 75$ UEs. Our optimal multi-association yields an upper bound to single-association whereas the (sub-optimal) SCF based association provides a feasible single-association and hence yields a lower bound to the optimal single-association. The results in Fig. 4.2 show that the performance of SCF, in terms of the gain in GM throughput with respect to the base-case, averaged over 100 realizations, is within 4% of the performance with optimal multi-association, across all configurations. Moreover, the gap between the lower-bound and the upper-bound was less than 5% for at least 95% of the realizations that we studied. The numerical closeness of the two bounds thus validates that the results obtained by considering multi-association are tight bounds for the optimal single-association. Moreover,



(a) No coordination



(b) ON-OFF TC

Figure 4.2: Average gain in GM throughput over 100 realizations for optimal and SCF association - $X = 4$ and $N = 75$

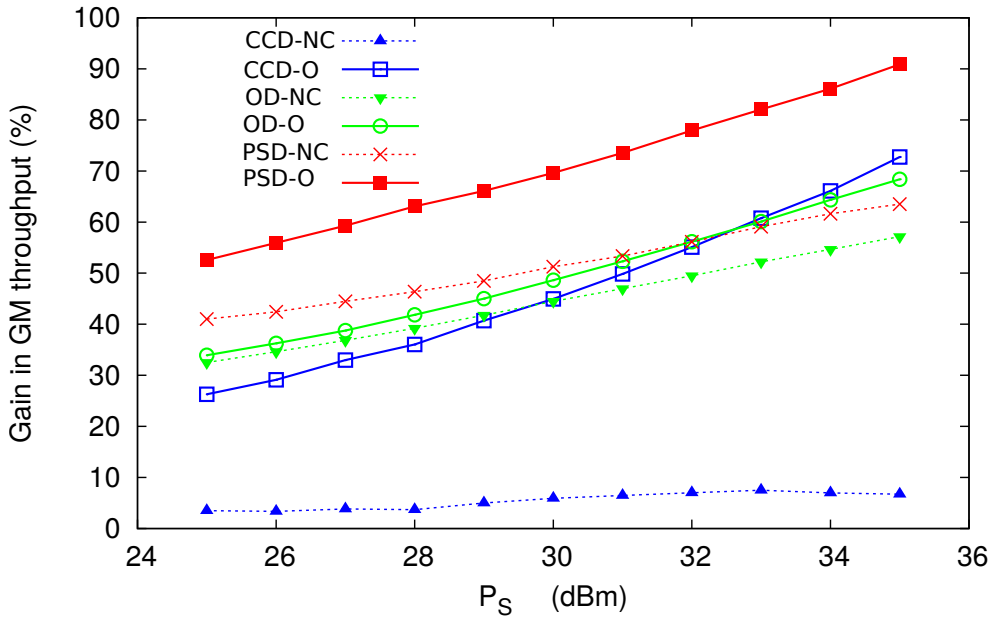


Figure 4.3: Average gain in GM Throughput over 100 realizations, optimal UA - $X = 4$ and $N = 75$

it also means that the optimal multi-association does not provide much performance gains over the optimal single-association. Hence, introducing multi-association capabilities will not offer significant performance gains[5].

4.5.2 Comparison between different RA schemes, and the need for transmission coordination

We present the average gain in GM throughput obtained by different configurations in Fig. 4.3 as a function of P_S for $X = 4$ and $N = 75$. Next, we discuss these results.

When P' is chosen properly, PSD clearly offers the best throughput performance among all the three RA mechanisms that we have considered. As evident from Fig. 4.3, PSD outperforms OD. The gains obtained by PSD over OD can simply be attributed to the

[5]We conducted similar computations for 100 cases of non-uniformly distributed users and randomly deployed SCs, and obtained similar results.

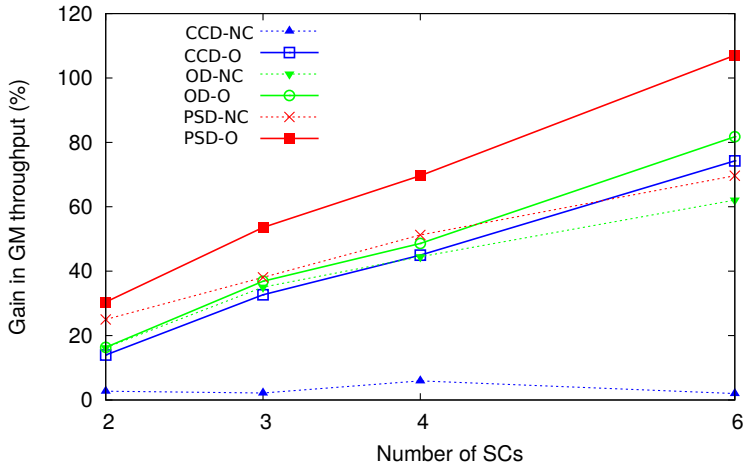
added flexibility of allowing MBS to use more channels at a carefully chosen power P'. It is however important to stress that any PSD is not guaranteed to perform better than OD if the power P' is not chosen carefully. CCD, on the other hand, performs very poorly in the case of no transmission coordination. Both PSD and OD outperform CCD significantly. In fact, the deployment of SCs under CCD provides very little gains (less than 8%) to the MBS-only deployment. Clearly, co-channel deployment, though attractive due to its simplicity, might perform poorly in the absence of transmission coordination.

ON-OFF transmission coordination case

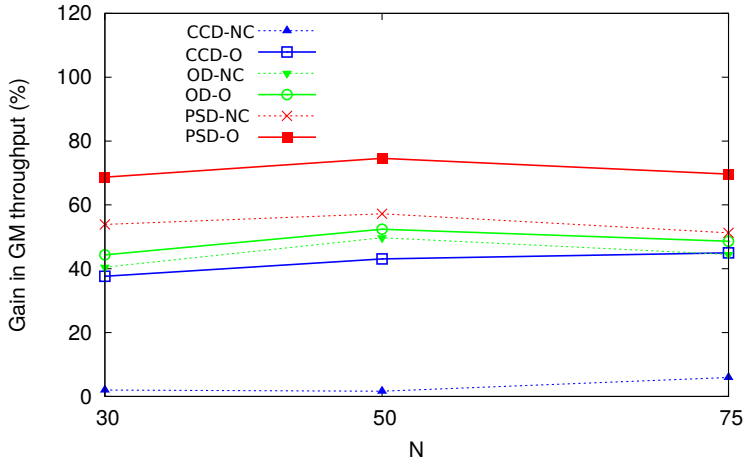
For a given RA, allowing ON-OFF TC can only improve over the case with no coordination. Our results show that the magnitude of improvements brought by ON-OFF TC are significant, especially for CCD. Under ON-OFF transmission coordination, PSD continues to perform significantly better (15 to 20 %) than CCD. More important perhaps is the observation that the relative performance of CCD under ON-OFF transmission coordination is very different from its performance under no coordination. CCD is a simple resource allocation mechanism as it does not require the configuration of any resource allocation parameter. The good performance of CCD under ON-OFF transmission coordination might motivate us to consider CCD as a favorable choice. However, we have seen that CCD requires transmission coordination, or otherwise performs too poorly to justify its simplicity. PSD as well as OD, on the other hand, perform very well even without transmission coordination as evident from the comparison of the performance of ***[PSD-NC]*** and ***[OD-NC]*** with the performance of ***[CCD-O]***. ON-OFF transmission coordination involves a problem of exponential complexity and requires a much fine-grained control as compared to computing the optimal channel split parameter K (with no coordination). Thus, our results favor PSD/OD over CCD.

Different number of SCs and UEs

Fig. 4.4a shows the performance of different configurations for $N = 75$ and $P_S = 30dBm$ for different numbers of SCs deployed. The performance of all configurations except ***[CCD-***



(a) Different number of SCs (X) with $N = 75$



(b) Different number of UEs (N) with $X = 4$

Figure 4.4: Average gain in GM throughput over 100 realizations, with $P_S = 30dBm$

NC] improve with more SCs deployed. Notable is the result that with increasing number of SCs, the gains due to ON-OFF TC increases for each RA scheme.

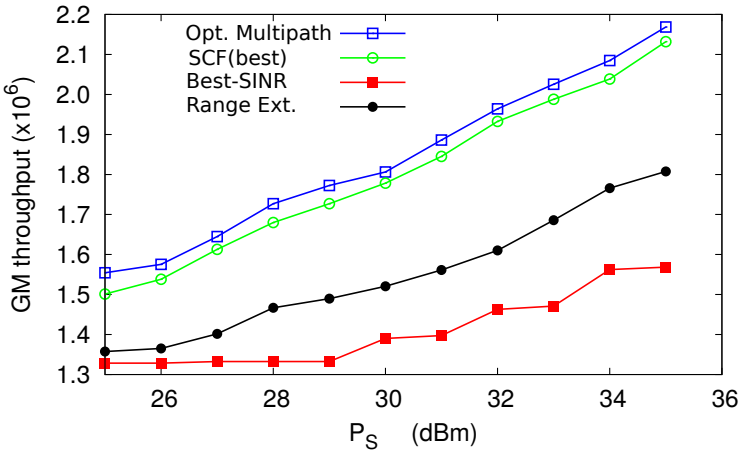
Fig. 4.4b shows the performance of different configurations for $X = 4$ and $P_S = 30dBm$ for different values of N. The results show that the performance in terms of throughput gains do not change significantly with the number of UEs in the system.

4.5.3 Performance of different UA rules

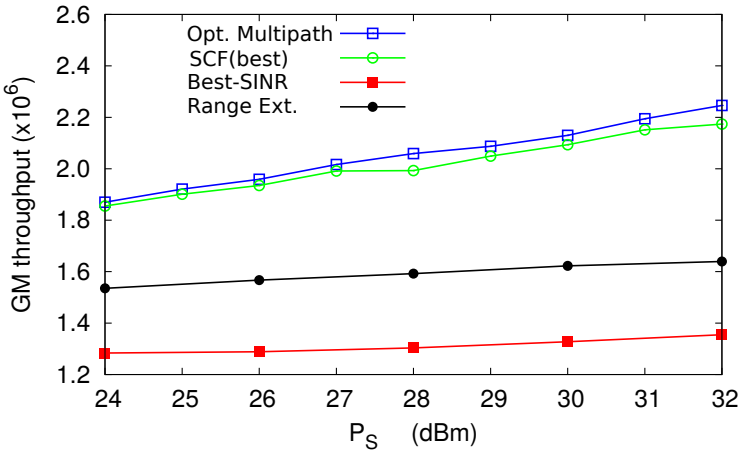
In Fig. 4.5, we show the performance of the three simple UA rules along with the optimal multi-association for ***[PSD-O]*** and ***[CCD-O]*** for $X = 4$ and $N = 75$. The results for SCF(δ) are obtained for a fine-tuned δ. This result shows that if properly configured, the performance of SCF is adequate and that it outperforms both best-SINR and range extension based UA rules. Similar conclusion was reported in [36] for the case of no coordination.

4.6 Conclusion

In this chapter, we used our flow-based optimization framework for the joint optimization of resource allocation, transmission coordination, and user association in a heterogeneous network comprising a macro base station and a set of SCs with wired backhaul links. This chapter demonstrates how our formulation can be used for the offline study of heterogeneous networks. We also obtained important engineering insights on the interplay of different network processes. Our results showed that the gain offered by multi-association as compared to the optimal single-association is small. Also, our numerical results showed that co-channel deployment requires transmission coordination for a satisfactory performance whereas partially shared deployment or orthogonal deployment perform well even in the absence of sophisticated transmission coordination mechanism. PSD/OD, thus can be a better practical approach as compared to CCD with ON-OFF transmission coordination.



(a) ***[CCD-O]***



(b) ***[PSD-O]***

Figure 4.5: Comparison of different UA rules with $X = 4$ and $N = 75$ - one realization (SCF is carried out for a fine-tuned δ)

Chapter 5

Detailed Study: Relay Deployment

Summary: In this chapter, we use the optimization formulation obtained in Chapter 3 to study different scenarios of relay deployment, and answer the following question: *what configurations of relay deployment can yield capacity gains?*

5.1 Introduction

In the previous chapter, we used our optimization framework introduced in Chapter 3 to study the performance of HetNet for the case of SCs with wired backhaul links in the presence of transmission coordination. In this chapter, we present a detailed study of different configurations of relay deployment with different resource allocation schemes, and transmission coordination schemes, under optimal user association settings. Note that, the relay deployment is more complicated than the wired SC deployment in terms of the set of configurations that are possible. Though not exhaustive, we have tried to incorporate a rich set of natural configurations that can be used for relay deployment. We also incorporate interference coming from nearby cells while computing the SINRs.

Recall that we can divide relay deployment into two different deployment scenarios:

- *user-band* relay deployment, where the relay (backhaul) links operate in the same band as the user links
- *dedicated-band* relay deployment, where the relay (backhaul) links operate in a dedicated band

Under the chosen framework of proportional fair throughput allocation, and under optimal user-association, we obtain the best performance for each configuration using $[\mathbf{P_{Joint}}(\boldsymbol{K}, \boldsymbol{P})]$. Based on these results, we obtain a number of interesting engineering insights on wireless backhauling[1]:

- Some configurations of user-band relay deployment scenario yield very little or even negative gains whereas some others can yield performances very close to the upper bounds corresponding to the wired scenario with infinite backhaul capacities. This highlights the importance of deploying the right configurations.
- Using a dedicated band for backhauling is a promising solution for small cells, in particular in the case of the mmWave band, since a small bandwidth is sufficient to satisfy the demand of a typical small cell backhaul link.

The rest of this chapter is organized as follows. In Section 5.2, we present the three deployment scenarios, and the details of the associated configurations. We describe the different requirements that each configuration imposes in terms of the node capabilities. In Section 5.3, we present the numerical results and finally present the conclusions of this study.

5.2 Scenarios

We assume that Band 1 is an LTE band with M_T LTE OFDM channels available for the entire multi-cell system. Further, we assume that all user links operate only in this band and thus a UE is equipped with an LTE AI, exclusively used as an rAI. Let the rAI for

[1]Some of these results were presented in our work in [43] and [42].

UE i be simply referred to as i for all $i \in \mathcal{U}$. The set of LTE channels available for user links are allocated to different macro cells by employing a reuse factor of 3.

We consider three scenarios. Scenario 1 corresponds to wired backhauling (the benchmark scenario). The other two scenarios correspond to wireless backhauling. In the first of the two wireless backhauling scenarios (Scenario 2), the backhauling is done on the same band as the user links (i.e., $S = 1$) while in the second (Scenario 3), backhaul links use a dedicated band. Scenario 3 is similar in many ways to Scenario 1. Scenario 2 on the other hand involves a number of different configurations. Thus, we first discuss Scenarios 1 and 3.

5.2.1 Scenario 1: wired scenario (benchmark scenario)

This scenario consists of one LTE band (i.e., $S = 1$), and one omni directional LTE AI at the MBS as well as at each SC, used for the direct and the access links, respectively. In addition, there are X wired backhaul links, each with a capacity of C. Since all M_T LTE channels are available for user links, a given macro cell gets a pool of LTE channels $\mathcal{M}^{(1)}$ simply written as $\mathcal{M}$ with $|\mathcal{M}| = \frac{M_T}{3} = M$. Note, that we have already presented this scenario in the previous chapter, for different channel allocation schemes. We will take OD (see Section 4.2 for details) as the resource allocation scheme, where the direct and the access links operate on orthogonal channels (respectively, $M - K$ and K channels). Note that with sufficiently large values of C, the wired backhauling scenario can be seen as a benchmark for wireless backhauling scenarios. In this study, we will use the performance of the wired scenario for a very large value of C, with optimal OD without coordination (OD-NC) as an upper benchmark.

5.2.2 Scenario 3: dedicated-band relay scenario

In this scenario, in addition to the LTE band (Band 1) for the user links, a separate mmWave band (Band 2) is available exclusively for the backhaul links (i.e., $S = 2$). We assume that the mmWave band has a bandwidth of F MHz. In order to exploit this new band, the MBS needs to have at least one additional AI in the mmWave band and each

RN needs to have one additional AI for receiving on the mmWave band. We consider two configurations for this scenario:
1) **mmWave-TDM**: MBS has one omnidirectional mmWave air-interface for transmitting to all backhaul links. Thus, the backhaul links operate in a time-shared fashion. We assume that all backhaul links in a given macro cell operate with a reuse factor of 3 (to manage the interference), and thus get a mmWave bandwidth of $B = \frac{F}{3}$.
2) **mmWave-SIMUL**: MBS has one directional mmWave air-interface for each backhaul link. Thus, the backhaul links operate as narrow-beam simultaneous links, all operating on the mmWave band (Band 2). In this case, we assume that the mmWave links do not interfere with each other, since the beams are very narrow. Hence, we can exploit full reuse, i.e., a backhaul link operates in entire mmWave band, i.e., $B = F$.

The mmWave band is assumed to comprise of one wide-band channel of bandwidth B and a logarithmic rate function, $\theta^{(2)}(\gamma) = B\log(1+\gamma)$. This scenario is very similar to Scenario 1 in the sense that the backhaul links do not steal channel resources from the user links. Thus, both channel allocations, OD and CCD, as defined before are relevant. Similar to the wired scenario, we consider OD only. However, unlike the wired scenario, the available transmit power budget at the MBS has to be divided between the direct links and the backhaul links. Let P_B be the power allocated to each mmWave backhaul link, then the power allocated to direct links will be $P_M - P_B$ for **mmWave-TDM** and $P_M - XP_B$ for **mmWave-SIMUL**. In other words, the values of P_B and K completely characterize the channel allocation and power allocation, which can be used to obtain the best GM throughput $\Lambda^*(P_B, K)$, corresponding to the optimal solution of $[\mathbf{P_{Joint}}(\boldsymbol{K}, \boldsymbol{P})]$. The best performance can then be obtained by fine tuning the power and channel allocations:

$$\max_{P_B \in P_\Delta, K \in \{1,2,\cdots,M\}} \Lambda^*(P_B, K)$$

where P_Δ is a discrete set of available power levels.

5.2.3 Scenario 2: user-band relay scenario

In this scenario, $S = 1$ and hence the backhaul links have to operate on the same LTE band as user links. We assume that an SC has one omni AI that it uses for both, transmitting

Table 5.1: Model parameters for Scenario 2 configurations

Num. of tAIs	Config.	Channels $C(\tilde{l})$	Channel allocation constraints	$P(\tilde{l})^{\dagger}$	$\mathcal{V}_i$
1 AI:	1	$= \mathcal{M}$, $\forall \tilde{l} \in \mathbf{L}_D \cup \mathbf{L}_A \cup \mathbf{L}_B$	$\lvert\mathcal{M}\rvert = \frac{M_T}{3} = M$	$= P_M$, $\forall \tilde{l} \in \mathbf{L}_D \cup \mathbf{L}_B$	(Given)
	2	$= \mathcal{M}_1, \forall \tilde{l} \in \mathbf{L}_D \cup \mathbf{L}_B$ $= \mathcal{M}_2, \forall \tilde{l} \in \mathbf{L}_A$	$\frac{M_T}{3} = M; \lvert\mathcal{M}_2\rvert = K$ $\lvert\mathcal{M}_1\rvert = M - K; \mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$	$= P_M$, $\forall \tilde{l} \in \mathbf{L}_D \cup \mathbf{L}_B$	$(K \in [1, M])$
	3	$= \mathcal{M}_1, \forall \tilde{l} \in \mathbf{L}_D \cup \mathbf{L}_A$ $= \mathcal{M}_2, \forall \tilde{l} \in \mathbf{L}_B$	$\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$ $\lvert\mathcal{M}_2\rvert = \frac{W_T}{3}; \lvert\mathcal{M}_1\rvert = \frac{M_T - W_T}{3} = M'$	$= P_M - P_B, \forall \tilde{l} \in \mathbf{L}_D$ $= P_B, \forall \tilde{l} \in \mathbf{L}_B$	$(W_T \in [1, M_T]$, $P_B \in P_\Delta)$
	4	$= \mathcal{M}_1, \forall \tilde{l} \in \mathbf{L}_D$ $= \mathcal{M}_2, \forall \tilde{l} \in \mathbf{L}_B$ $= \mathcal{M}_3, \forall \tilde{l} \in \mathbf{L}_A$	$\mathcal{M}_j \cap \mathcal{M}_{j'} = \emptyset, \forall j, j' \in \{1,2,3\}, j \neq j'$ $\lvert\mathcal{M}_2\rvert = \frac{W_T}{3}; M' = \frac{M_T - W_T}{3}$ $\lvert\mathcal{M}_3\rvert = K; \lvert\mathcal{M}_1\rvert = M' - K$	$= P_M - P_B, \forall \tilde{l} \in \mathbf{L}_D$ $= P_B, \forall \tilde{l} \in \mathbf{L}_B$	$(W_T \in [1, M_T]$, $P_B \in P_\Delta$, $K \in [1, M'])$
$X + 1$ **AIs**:	5	$= \mathcal{M}_1, \forall \tilde{l} \in \mathbf{L}_D \cup \mathbf{L}_B$ $= \mathcal{M}_2, \forall \tilde{l} \in \mathbf{L}_A$	$\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$ $\lvert\mathcal{M}_2\rvert = W_T; \lvert\mathcal{M}_1\rvert = \frac{M_T - W_T}{3} = M'$	$= P_M - XP_B, \forall \tilde{l} \in \mathbf{L}_D$ $= P_B, \forall \tilde{l} \in \mathbf{L}_B$	$(W_T \in [1, M_T]$, $P_B \in P_\Delta)$
	6	$= \mathcal{M}_1, \forall \tilde{l} \in \mathbf{L}_D$ $= \mathcal{M}_2, \forall \tilde{l} \in \mathbf{L}_B$ $= \mathcal{M}_3, \forall \tilde{l} \in \mathbf{L}_A$	$\mathcal{M}_j \cap \mathcal{M}_{j'} = \emptyset, \forall j, j' \in \{1,2,3\}, j \neq j'$ $\lvert\mathcal{M}_2\rvert = W_T; M' = \frac{M_T - W_T}{3}$ $\lvert\mathcal{M}_3\rvert = K; \lvert\mathcal{M}_1\rvert = M' - K$	$= P_M - XP_B, \forall \tilde{l} \in \mathbf{L}_D$ $= P_B, \forall \tilde{l} \in \mathbf{L}_B$	$(W_T \in [1, M_T]$, $P_B \in P_\Delta$, $K \in [1, M'])$

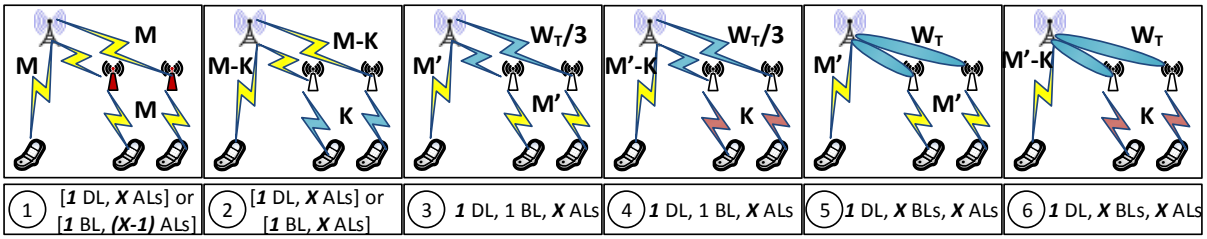


Figure 5.1: Configurations of Scenario 2 (DL: Direct Link, AL: Access Link, BL: Backhaul Link)

on an access link as well as receiving on the backhaul link. Note that this means an SC cannot simultaneously transmit and receive in the same set of channels (even though it can do so over orthogonal set of channels). If an SC had two AIs, such limitation could be avoided. However, in the absence of a mechanism to separate the interference between a tAI and an rAI of the same node (e.g., interference cancellation, spatial separation), the additional AI would not be beneficial.

We consider six configurations for this scenario which differ in terms of the number of AIs at the MBS and the way the LTE channels are allocated to the direct, access, and the backhaul links. In other words, each configuration is characterized by a given number of AIs at the MBS and the channel allocation scheme. The configurations are depicted in Fig. 5.1. Even though our selection of configurations is not exhaustive, we believe that we have included the most natural ones. Next, we discuss the implications of having a number of AIs at the MBS as well as the choice that we make in terms of channel allocation.

Number of AIs

In terms of the number of AIs at the MBS, we consider two possibilities: 1 AI and $X + 1$ AIs (recall that X is the number of SCs). We could also consider the case with 2 omni AIs, one for the direct links and the other for the backhaul links. However, having a simultaneous direct and backhaul link on the same set of channels would mean a lot of mutual interference due to the omni directional nature of both AIs.

- **1 AI**: The MBS has only one omni AI. This AI is used for both the direct and the backhaul links. This means that on a given channel, only one link can be activated at a time. Configurations 1 to 4 in Fig. 5.1 are such configurations. Let o_0 be the omni AI of the MBS, and let T_j and R_j respectively be the tAI and rAI in RN j, then the set of direct links is given as $\mathbf{L}_D = \{(o_0, i) : i \in \mathcal{U}\}$, the set of access links is given as $\mathbf{L}_A = \{(T_j, i) : j \in \mathcal{P}, i \in \mathcal{U}\}$, and the set of backhaul links is given as $\mathbf{L}_B = \{(o_0, R_j) : j \in \mathcal{P}\}$.

- $X + 1$ **AIs**: The MBS has one omni AI called o_0 for the direct links, and one directional AI D_{0j} for each backhaul link. This means that on a given channel, up to $X + 1$ links can be activated simultaneously. Configurations 5 and 6 in Fig. 5.1 are such configurations. The set of direct, access, and backhaul links are then given as $\mathbf{L}_D = \{(o_0, i) : i \in \mathcal{U}\}$, $\mathbf{L}_B = \{(D_{oj}, R_j) : j \in \mathcal{P}\}$, and $\mathbf{L}_A = \{(T_j, i) : j \in \mathcal{P}, i \in \mathcal{U}\}$ respectively.

Channel and power allocation

In addition to a given number of AIs at the MBS, each configuration in Fig. 5.1 has a specific channel allocation, which is illustrated in Fig. 5.1 and specified in details in Table 5.1. For Configuration 1, all (direct, access and backhaul) links are allocated all the available channels ($\mathcal{M}$). For Configuration 2, on the other hand, the direct and the backhaul links are allocated the same set of channels whereas the access links are allocated the remaining channels. Table 5.1 also shows the power allocated to each link.

Each channel allocation choice has its own impact:

1. **Is a direct link orthogonal to an access link?** If no, access links will receive large interference from the direct links and thus some transmission coordination (i.e., ON OFF TC) might be required. Configurations 1, 3 and 5 are such configurations where direct and access links interfere and thus we study both the NC and ON-OFF TC. For Configurations 2, 4 and 6, however, we only consider the case of no coordination.

2. **Is a backhaul link orthogonal to a direct link?** If yes, a backhaul link can operate in parallel to a direct link (Configurations 3 to 6). In that case, the MBS can

simultaneously have one direct link and either one backhaul link (for configurations with 1 AI at MBS) or X backhaul links (for configurations with $X+1$ AIs at MBS). In this case, power allocation for the backhaul links is crucial. Let P_B be the power allocated to a backhaul link. Then, the power allocated to each direct link is $P_M - P_B$ for Configurations 3 and 4 and it is $P_M - XP_B$ for Configurations 5 and 6.

3. **Is a backhaul link orthogonal to an access link?** If no, an RN cannot transmit while it is receiving on the backhaul link (e.g., Configuration 1). Configuration 1 is an example of the *in-band* RN deployment specified in LTE-A [1].

Recall that, for each configuration, given the channel and power allocation per physical link, our optimization model [$\mathbf{P_{Joint}}(\boldsymbol{K}, \boldsymbol{P})$] can be used to obtain the optimal geometric mean throughput. The set of parameters for determining a channel allocation and power allocation per physical link for configuration i is represented as $\mathcal{V}_i$ and is shown in Table 5.1. For example, for Configuration 1, there are no such parameters (in the sense that no channel/power allocation parameter has to be chosen). For Configuration 6 on the other hand, there are three parameters (namely, the number of channels allocated to backhaul links (W_T), the channel-split parameter between direct and the access links (K), and the power allocated to the backhaul links P_B). Let $\Lambda_i^*(\mathcal{V}_i)$ be the optimal GM throughput obtained for a given choice of channel and power allocation $\mathcal{V}_i$, then the best performance for configuration i is obtained by fine-tuning these parameters:

$$\Lambda_i^* = \max_{\mathcal{V}_i} \Lambda_i^*(\mathcal{V}_i)$$

.

5.3 Numerical Results

We consider a macro cellular layout as shown in Fig. 1, with a given inter-site distance ($ISD = 1732m.$), which corresponds to a rural settings. The central macro cell in Fig. 1 forms the HetNet system with its centrally placed MBS and $X = 4$ SCs at a radius of $d = 400m.$, symmetrically. $N = 50$ users are uniformly distributed in the central

Table 5.2: Available rates and the corresponding SNR thresholds (the last two are available for relay links only)

γ	-6.5	-4	-2.6	-1	1	3	6.6	10	11.4	11.8	13	13.8	15.6	16.8	17.6	21.04	24.07
η	0.15	0.23	0.38	0.60	0.88	1.18	1.48	1.91	2.41	2.73	3.32	3.9	4.52	5.12	5.55	7.0	8.0

cell. A given realization i of user positions (and the corresponding channel-gains across all communication links) is taken as a static snap-shot of the system. We study 100 such realizations, with a condition that each of them is *connected* even when the macro cell does not have any SCs. The physical layer parameters for LTE (shown in Table 5.3) correspond to the parameters recommended on the 3GPP evaluation recommendations [1]. The LTE path-loss models for MBS and small cells are used along with a log-normal shadowing of 8 dB standard deviation, for generating channel-gains G_{ji} for the direct and the access links.

We assume that the relays are outdoor and thus there is no penetration loss (pen. loss) for backhaul links. Also, we assume that there exists a line-of-sight (LOS) between the MBS and an RN of the same cell and thus we take LOS path-loss model between the serving MBS and its RNs. We use non-LOS (NLOS) path-loss model to compute channel gains between an MBS and an RN that are located in different cells (i.e., for calculating interference). Directional backhaul links have an additional directional gain of 20dB and we assume that the directional links do not interfere with each other.

Also, while calculating inter-cell interference, due to the small transmit power of SCs and a much faster power attenuation with distance, we ignore the interference from SCs in nearby macrocells. We however account for the interference from all surrounding MBSs.

We use an MCS with 15 rates for the user links [42]. For the LTE relay backhaul links (Scenario 2), we have two extra modulation schemes (corresponding to 256QAM with a rate of 1/2 and 2/2) (see Table 5.2). The efficiency η is related to rate R as $R = \eta \frac{n_{sc} n_{ts}}{T_{symbol}}$ where n_{sc} is the number of sub-carriers per OFDM symbol, n_{ts} is the number of OFDM symbols in one subframe, and T_{symbol} is the duration of one subframe.

The mmWave parameters are taken from [76] and are shown in Table 5.3. The path-loss model taken is considered to be a realistic model for links at 28 GHz. As mentioned already, a logarithmic rate function is assumed for the mmWave links.

Table 5.3: Physical layer parameters

UE Noise Power	-174 dBm/Hz	P_M	43dBm
P_S	30dBm	Channel BW	180 KHz
UE Noise-figure	9dB	RN Noise-figure	5 dB
UE Pen. Loss	20 dB		
MBS Ant. Gain	15 dBi	SC Ant. Gain	5 dBi
Directional Gain	20 dBi	$T_{subframe}$	1 ms
n_{sc}	12	n_{ts}	14
MBS-UE Path-loss	$128.1 + 37.6 \log_{10}(d/1000), d \geq 35m$		
SC-UE Path-loss	$140.7 + 36.7 \log_{10}(d/1000), d \geq 10m$		
MBS-SC Path-loss	LOS: $103.4 + 24.2 \log_{10}(d/1000)$ NLOS: $131.1 + 42.8 \log_{10}(d/1000)$		
mmWave:			
Tr. Gain	25dBi	Rcv. Gain	12 dBi
Impl. Loss	3dB	Noise-figure	7 dB
Path-loss	$157.4 + 32 \log_{10}(d/1000)$		
$P_\Delta = \{-10, -5, 0, 5, 10, 15, 20, 25, 30\}$ dBm			

For a given backhauling scenario and a specific configuration, we use our optimization framework to obtain the allocated throughputs for each realization i and obtain the best GM throughput by fine-tuning the channel and power allocations as explained before. Also, we take the scenario of MBS-only deployment as the base scenario and express the performance of different scenarios and their configurations in terms of the gain in performance w.r.t. that MBS-only deployment.

5.3.1 Scenario 2: user-band relay scenario

Fig. 5.2 shows the percentage gain in GM throughput (with respect to the MBS-only case) for each of the six configurations of the user-band relay scenario as well as the wired scenario. The results show that Configurations 1, 3 and 5 (all corresponding to

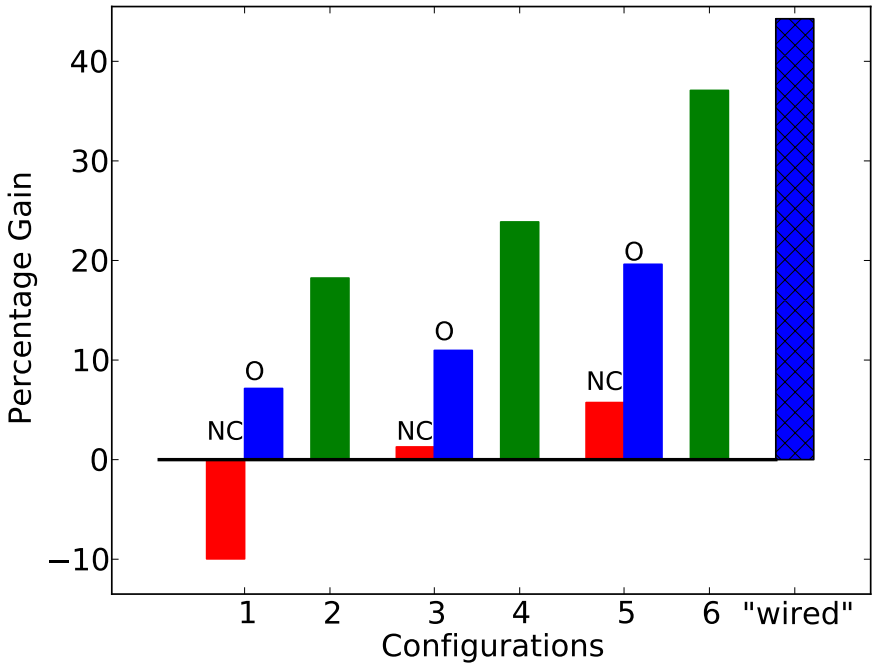


Figure 5.2: Scenario 2: Different configurations (NC means no coordination, O means ON-OFF coordination)

configurations where the access links get interference from the direct links) in the absence of interference coordination (NC) do not yield meaningful throughput gains. In fact, with Configuration 1, there is a negative gain in the performance w.r.t. the MBS-only case. This means the spatial reuse gain and SINR improvement brought to some poor users does not off-set the loss in performance due to an overall increased interference. Even for the configuration with $X + 1$ AIs (Configuration 5), a very small gain in performance is observed. These configurations (1, 3 and 5), however do much better in the presence of ON-OFF coordination. The results are not surprising since ON-OFF coordination is a means to combat the interference to an access link from a direct link.

The figure also shows performance results for Configurations 2, 4 and 6 without coordination. These configurations do not require the transmission coordination for protecting access links from the MBS interference. The performance of these configurations show that the number of AIs and the channel allocation scheme play a very important role in the performance of an RN deployment. With $X+1$ AIs and an appropriate channel allocation (Configuration 6), we obtain performance not very far from the upper bound (38% for Configuration 6, and 44% for the upper bound).

5.3.2 Scenario 3: mmWave backhaul

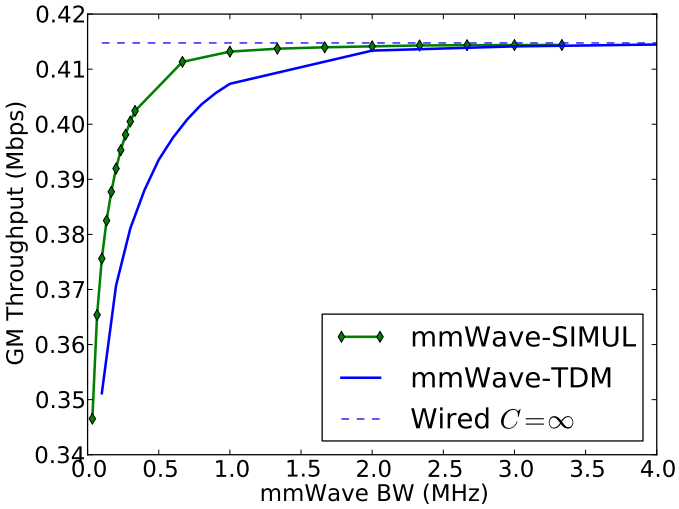
In Fig. 5.3a, we plot the GM Throughput performance of two configurations of mmWave backhauling (**mmWave-SIMUL** and **mmWave-TDM**), as a function of the available mmWave bandwidth F, assuming that the best power allocation P_B in P_Δ is chosen. We also show the upper-bound which corresponds to the wired scenario with infinite backhaul capacities. As we can see, a bandwidth of about $2.5MHz$ can yield a performance within 98% of the upper-bound. This is a very small bandwidth in a typical mmWave spectrum. This shows that a small fraction of available mmWave bandwidth can be sufficient to satisfy the load on backhaul links.

In Fig. 5.3b, we plot the GM throughput versus the backhaul link capacity for the wired scenario, and for the two values of bandwidth of **mmWave-SIMUL**. For the mmWave scenario, the backhaul capacity is determined by the power P_B we allocate to the backhaul links. For the wired deployment, it is obvious that the performance improves with an

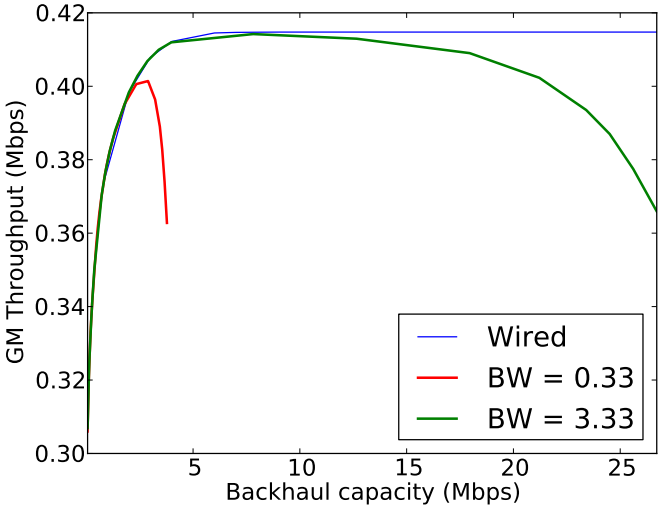
increase in backhaul capacity before it saturates. For the mmWave deployment, however, the results show that it is important to make sure that the relay backhaul power P_B is carefully chosen, or otherwise the performance can degrade significantly.

5.4 Conclusion

We used the optimization framework developed in Chapter 3 to evaluate different configurations of relay node deployment in a HetNet consisting of multiple bands and air interfaces per node. Our results show that some configurations of *user-band relay scenario* can yield negative or negligible performance gains where as some others can yield very good performances. Also, our results show that for the dedicated-band relay scenario, a small mmWave bandwidth is sufficient to satisfy the load on a typical small cell backhaul, provided that available parameters are chosen carefully. Thus, it is quasi-equivalent to the wired scenario.



(a) GM Throughput vs. F



(b) GM Throughput vs. backhaul link capacities

Figure 5.3: Scenario 3 (mmWave) along with Scenario 1 (Wired)

Chapter 6

User Scheduling under Backhaul Limitations

Summary: In this chapter, we study how backhaul capacity limitations impact the user scheduling. We consider a global α-fair user scheduling problem and characterize its solution under different scenarios of backhaul limitations.

6.1 Introduction

6.1.1 A different approach

The modeling approach taken so far focused on unifying different network processes together, and characterizing different set of configuration choices under the same footing. This approach led us to a very general and powerful optimization model [$\mathbf{P}_{\mathbf{Joint}}(\boldsymbol{K}, \boldsymbol{P})$]. This approach, however, has some limitations. Since all network processes are optimized jointly, it might not be able to reflect the reality of networks where different network processes are optimized at different time-scales. For example, user association are not necessarily jointly optimized across all users in the system. Also, the snapshot approach is limited to the offline study phase.

From this chapter onwards, in order to yield simple models providing useful insights, and results that can be used to obtain online algorithms, we take a different approach where we will study one network processes at a time while others are given and tuned. In this chapter, we will focus on user scheduling (by assuming that resource allocation and user association are given), and in the next chapter we will study user association.

We will restrict ourselves to the wired SC deployment[1] with ***[OD-NC]***, i.e., orthogonal deployment with no coordination. Our assumption of no transmission coordination, allows us to take a model much simpler than the flow-model that we formulated before.

6.1.2 Focus on backhaul limitations

Most of the studies in the literature focus on the wireless access end of the HetNets, and hence there is an implicit assumption that the backhaul infrastructure is not limiting. Our studies in Chapter 4 also made this assumption. Such an assumption could be justified in older cellular networks, where the access network (and not the backhaul network) was the bottleneck. In the emerging HetNet architecture, this assumption needs to be reexamined.

Network operators see small cell backhauling as an immediate challenge for the successful deployment of HetNets [29], [73]. The ultra-dense deployment of small cells with low average number of users per BS means that the *cost of backhauling* for small cells becomes a significant part of the total Capital Expenditure (CAPEX), in some cases exceeding the cost of the small cell BS equipment [73]. It is thus desirable that the backhauling cost for small cells is kept low. This economic consideration can often limit the capacity of the installed SC backhaul links. For example, a number of cheap solutions are being proposed, including ADSL [35], mesh networks [94], and even non-licensed microwave links [35]. Besides economy, *flexibility* is also a key requirement as there will be numerous SCs added or moved frequently. Fiber or copper infrastructures are often not flexible. The third constraint is *physical*. A small cell might be at an inaccessible street furniture where bringing a fiber link can be infeasible. A low capacity solution like non-line-of-sight (NLOS) wireless backhauling might be the only available option in such a case [29].

[1]Note that wireless backhaul links with dedicated spectrum are quasi-equivalent to wired backhaul links, as shown in Section 5.3.2 and thus this study can be easily adapted to such cases.

MBS backhaul limitations, on the other hand, are less likely to be a concern right now, since MBS backhauling is a small portion of the CAPEX [73], and thus can be well provisioned. However, the future networks are expected to operate with a high number of small cells per macro base station, with highly efficient wireless links (e.g., using massive MIMO [18]) and on very high bandwidth spectrum (e.g., mmWave [77]). This will translate to a huge increase in traffic load on the backhaul. Moreover, many multi-cell architectures are emerging where signaling for coordination between BSs is done via the backhaul links (e.g., Joint Processing (JP) CoMP [58]), which increases the traffic load on the backhaul links as well as pose more stringent delay requirements. The deployment of cloud-RAN (C-RAN) [26] like architecture is also going to put a lot of pressure on the MBS backhaul. So, it is possible that MBS backhaul limitation might also be a concern for future networks.

Finite capacity of a backhaul link translates into two types of limitations: *1) rate limitation:* the maximum amount of traffic (in bits per seconds) that can be carried via the backhaul link, and 2) *delay limitation:* the delay/jitter incurred by the backhaul link for a given traffic load. These two aspects are inter-related, usually via complex relationships, which are explored using various queuing models. The rate limitation directly affects the total throughput in the HetNet whereas the constraints imposed on delay are key in meeting control signaling deadlines. In this study, we focus only on the rate limitation of the backhaul links, where a backhaul link l has a maximum capacity of C_l Mbps. Note that, limiting the aggregate amount of traffic on a link to a given rate (lower than C_l) can also be used to guarantee a certain level of delay performance on that link.

Topology of the Backhaul Infrastructure

The exact topology of the backhaul system can have a major impact on the performance. We consider a hierarchical topology of the backhaul links where SC j is connected to the MBS via a backhaul link of capacity C_j and the MBS is connected to the core via a backhaul link of capacity C_{BH}. In other words, for a downlink system, an SC backhaul link has to carry the downlink traffic of its users only whereas the MBS backhaul link has to carry the aggregate traffic of all its users as well as the aggregate traffic from all other SCs in its cell.

6.1.3 Objective

The purpose of our study is to understand the impact of backhaul limitations on how user scheduling is to be performed on the downlink of HetNets. Our main message is that finite backhaul links have a fundamental impact on user scheduling, i.e., there is a need for backhaul-aware user schedulers.

We focus on a *macro cellular area* with one macro base station (MBS), and a number of small cells connected to the MBS within a macro cell. We only study the downlink and assume that the resource allocation and the user association scheme are given. For a given *network realization* of channel gains, our objective is to schedule the users at these BSs so as to guarantee fairness. We use the concept of α-fairness, and study user scheduling scheme that guarantees α-fairness in a global sense (i.e., over all users in the considered macro cellular area). By choosing the value of α, an operator can strike the trade-off she wants between fairness and efficiency.

6.1.4 Contributions

Our contributions , summarized in Table 6.1, can be stated as follows[2].

1) Our work builds on [37], where Fooladivanda and Rosenberg study the special case of α-fairness where $\alpha = 1$, also called proportional fairness (PF), under unconstraining backhaul capacities. Under this scenario, they have shown that, under some assumptions, the global proportional fair (PF) user scheduling problem decomposes into independent local PF user scheduling problems (one per BS). Additionally, they show that the local PF is equivalent to a local equal-time scheduling scheme. We generalize these results for the general α-fair utility function and in particular derive closed-form expressions for optimal schedules.

2) For the scenario where the MBS backhaul is sufficiently provisioned and hence is not the bottleneck, but where the SC backhaul links have limited capacities, we present

[2]Some of these results were presented in our work [39]. Our work [40], accepted for publication, contains the expanded version.

the results for a general value of $\alpha > 0$. Our findings for this scenario can be summarized as follows.

- Similar to the scenario of very large SC backhaul capacities, the global problem can be decomposed into independent local problems. The nature of the local α-fair scheduling is different from that of the scenario of very large backhaul capacities. For example, local PF scheduling under backhaul limitations is not always equivalent to the local equal-time scheduling.

- In order to achieve global α-fairness, we show that each small cell j has to schedule its users based on how its backhaul capacity C_j compares to two critical values c_j^* and $C_{j,\alpha}^*$, which are specific to a given network realization. We show that if $C_j \leq c_j^*$ then local α-fair scheduling is equivalent to local *equal-throughput* scheduling, while if $C_j \geq C_{j,\alpha}^*$ then it is equivalent to local α-*fair* scheduling under unconstraining backhaul capacities.

- Using numerical results, we quantify the impact of limited SC backhaul capacity on the system performance. We also propose a heuristic scheduler that is simple to compute and performs very well.

3) For the more general scenario, where the MBS backhaul is also of limited capacity, we perform a detailed analysis of the global scheduling problem, and obtain a number of results. Our findings for this scenario can be summarized as follows.

- We introduce a notion of *virtual backhaul capacity* that allows us to decompose the global problem into per-BS local problems. We present a simple bisection search based algorithm to compute the optimal values of the virtual backhaul capacities. However, these values are realization-dependent and have to be re-computed whenever the network realization changes. In other words, the user schedule at a BS is affected by the channel gains of users in other BSs, which we call the *global realization-dependence* of the optimal solution.

- We present two realization-agnostic heuristics where the *virtual backhaul capacities* are kept fixed all the time, thereby reducing the complexity of the scheduling problem

Table 6.1: Summary of contributions

$\alpha = 1$ (PF), Unlimited $\{C_j\}$ and C_{BH}	Prior art [37], [61]
$\alpha > 0$ (General), Unlimited $\{C_j\}$ and C_{BH}	Contribution 1
$\alpha > 0$ Finite $\{C_j\}$, Unlimited C_{BH}	Contribution 2
$\alpha > 0$ Finite $\{C_j\}$, Finite C_{BH}	Contribution 3

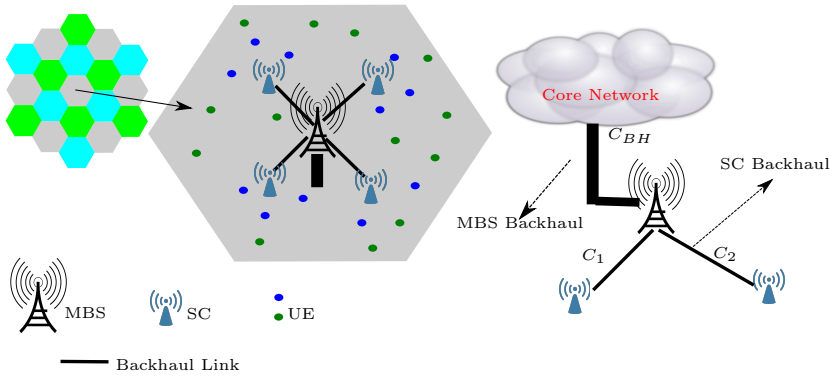


Figure 6.1: Our system.

greatly. We quantify the loss in performance due to these schemes and show that they both work well.

The rest of this chapter is organized as follows. In Section 6.2, we present the system model. Section 6.3 shows the formulation of the general optimization problem. In Section 6.4, we consider the scenario of unlimited backhaul capacities. In Section 6.5, we consider the scenario when the MBS backhaul is very large and thus SC backhaul links are the only limitations. Section 6.6 considers the general scenario where the MBS backhaul is also limited. Relevant results are presented in each section. Section 6.7 concludes this chapter. The relevant proofs are included in Appendix A.

6.2 System Model

We consider an OFDM-based cellular network consisting of multiple macro cells. Each macro cell comprises one macro base station (MBS), X small cells (SCs), and N user equipments (UE) (sometimes simply called users), see Fig. 6.1. We consider each macro cell, with its MBS, SCs, and UEs as a standalone HetNet system. However, we account for interference coming from nearby macro cells, as we will describe later. We focus on the macro cell in the middle. 0 represents the MBS, $\mathcal{P} \triangleq \{1, 2, \cdots, X\}$ represents the set of SCs, and $\mathcal{N}$ represents the set of all UEs.

In this study, we consider a *tree* topology of the backhaul network as shown in Fig. 6.1 where small cell $j \in \mathcal{P}$ is connected to the MBS via a backhaul link of capacity C_j. The capacity of the backhaul link between the MBS and the backbone is given as C_{BH}. Since the major portion of the traffic load is usually on the user plane, we ignore the traffic coming from the control plane[3].

We consider only the *downlink* of the HetNet and assume that all users are active, i.e., there exists a downlink flow from the MBS (source) to each UE (destination). We assume that the users are greedy in throughput and that the BSs have an infinite backlog of packets per UE. The MBS has a transmit power budget of P_M and each small cell has a transmit power budget of P_S. We assume that each BS transmits all the time with its available transmit power.

The system as a whole uses M' OFDM subchannels and each macro cell is allocated $M = \frac{M'}{r}$ subchannels, where $r > 1$ is the reuse factor. Thus, a total of M OFDM subchannels are available for the HetNet system under study (i.e., to be used by the MBS in the middle of Fig. 6.1 and its X SCs).

Different subchannel allocation schemes can be used inside the HetNet, with significant effect on the overall system performance. In this study, we consider Orthogonal Deployment (OD) (See Section 4.2), where K subchannels are allocated to the small cells and the remaining $M - K$ subchannels are allocated to the MBS. This exclusive partitioning of

[3]With more complex cooperative communication (like the CoMP) with joint processing, the control plane will also carry a large traffic load in the future.

subchannels between the MBS and the SCs means that the macro transmissions and SC transmissions do not interfere with each other. In this study, we assume that K is given. The analysis in this work can be applied to other variants, including the partially shared deployment (PSD) and co-channel deployment (CCD) (See Section 4.2).

The following assumptions[4] will allow us to simplify our subsequent formulations: **A1.** A BS transmits on all the subchannels allocated to it; **A2.** Power allocated to a given BS is equally divided among all the allocated subchannels; **A3.** Channels are flat, i.e., the channel gains across different subchannels between a BS and a UE are equal. These assumptions allow us to reduce a time and frequency domain scheduling to pure time domain *single user scheduling* problem, where a BS allocates all of its subchannels to one UE at a given time, as discussed in [42]. However, this means that the channel-dependent scheduling aspect of an OFDM system cannot be exploited in this framework.

A realization $\omega \triangleq \{G_{ji}(\omega)\}_{j\in\{0\}\cup\mathcal{P},i\in\mathcal{N}}$ represents a set of channel gains between all (BS,UE) pairs. Channel gain $G_{ji}(\omega)$ between BS j and UE i incorporates two random aspects of the network: *1)* the random locations of N users[5], which will result in random path-loss between the BSs and the users, *2)* a random slow fading at each location modeled by a log-normal shadowing of a given standard deviation. Note that, the notations for the model elements in Chapter 3 did not contain the explicit reference to realization ω, as it was understood by the context.

6.2.1 Physical interference model and link rates

Let $\gamma_{ji}(\omega)$ be the signal to interference plus noise ratio (SINR) between BS j and UE i on each allocated subchannel for a given realization ω, and for a given P_M and P_S. For all $j \in \mathcal{P} \cup \{0\}$ and for all $i \in \mathcal{N}$, we have

$$\gamma_{ji}(\omega) = \frac{P_j G_{ji}(\omega)}{N_0 + \mathbf{1}_{\{j\in\mathcal{P}\}} \sum_{k\in\mathcal{P}:k\neq j} P_k G_{ki}(\omega) + \mathbf{1}_{\{j=0\}} I^r_{0i}} \tag{6.1}$$

[4]Note that these assumptions are identical to the assumptions in Section 3.3.3.

[5]N (and hence $\mathcal{N}$) can also depend upon ω if we consider a random number of users.

where P_j is the power per subchannel for BS j given as

$$P_j = \mathbf{1}_{\{j=0\}} \frac{P_M}{(M-K)} + \mathbf{1}_{\{j\in\mathcal{P}\}} \frac{P_S}{K}. \tag{6.2}$$

I_{0i}^r is the interference coming to user i from macro BSs in the nearby macro cells using the same channel resources as MBS 0, based on the reuse factor of r employed among the macro cells. In order to compute this interference, we assume that the nearby HetNets have identical channel allocation scheme (i.e., OD with the same value of K) and transmit power budgets. Interference due to SCs in the nearby cells is often very small. So, for simplicity, we do not consider the interference from SCs in other macro cells, but we do consider interference from SCs in the same cell.

There is a discrete function $\theta(.)$ that maps the SINR $\gamma_{ji}(\omega)$ from BS j to user i to the maximum supportable data rate per subchannel. Then, the maximum supportable rate $R_{ji}(\omega)$ for user i associated to BS j (available only if the UE i is alone in BS j) is given as

$$R_{ji}(\omega) = K_j \theta(\gamma_{ji}(\omega)) \tag{6.3}$$

where K_j is the number of subchannels allocated to BS j, given by the OD channel allocation scheme, as follows.

$$K_j = K\mathbf{1}_{\{j\in\mathcal{P}\}} + (M-K)\mathbf{1}_{\{j=0\}}, \forall j \in \{0\} \cup \mathcal{P} \tag{6.4}$$

For a given realization ω, and given backhaul capacities (C_{BH} and $\boldsymbol{C} = [C_1, C_2, \cdots, C_X]$), we assume that the channel allocation parameter K as well as the rate-function $\theta(.)$ are given. In this case, the $R_{ji}(\omega)$'s can be computed a priori as input parameters using (6.1), (6.2), (6.3), and (6.4). Even though our model assumes that the value of K is given, note that choosing a good value of K is important (and in general not trivial) [42].

6.2.2 User association (UA)

We assume that the user association rule is given, with one UE associating to only one BS. Without loss of generality, we assume that we employ the *Small Cell First (SCF)* user

association rule, presented in Section 4.4. Recall that it has a tunable parameter δ. We choose it as it had the better performance than the other simple UA schemes. Thus, for a given K and ω, this rule with a given value of δ allows us to determine the set of UEs associated to BS j, represented as $A_j(\omega)$. Let $N_j = |A_j(\omega)|$ represent the number of UEs associated to BS j. We assume that the above stated user association scheme guarantees that each UE has a non-zero rate to its BS, i.e., $R_{ji}(\omega) > 0$ for all $j \in \{0\} \cup \mathcal{P}$ and for all $i \in A_j(\omega)$. Note that if $i \notin A_j(\omega)$, then by our definition, $R_{ji}(\omega) = 0$. It is important to note that, even for a fixed value of the UA parameter δ, the sets $A_j(\omega)$ change with the realization. The backhaul limitations also could have an impact on UA schemes. In this chapter, we take a simple UA scheme and thus do not consider this impact. Designing backhaul-aware UA scheme is however very important, and we address it in the next chapter.

6.3 Global User Scheduling Problem

We intend to schedule the users so as to guarantee a *global* fairness. This would entail fairness among all users in the entire system, i.e., over multiple cells. However, under our assumptions, the system-level global scheduling can be separated into independent per macro cell scheduling problems. So, in the following, when we mention the *global* problem, we mean the problem at the level of one macro cell, and thus global fairness deals with users within the macro cell under consideration[6]. These users might be associated to the MBS or one of the X SCs.

We use the notion of α-fairness, which was introduced in [71], and has been used often in throughput allocation frameworks usually under Network-Utility Maximization (NUM) formulations [90], [72]. If λ is the throughput offered to a given user, the *utility* corresponding to this allocation is given by $U_\alpha(\lambda) = \frac{\lambda^{1-\alpha}}{1-\alpha}$ if $\alpha > 0, \alpha \neq 1$ and is given by $U_\alpha(\lambda) = \log(\lambda)$ if $\alpha = 1$.

For tractability, we made the assumptions **A1**-**A3**, which allow us to reduce the scheduling problem to a pure time-domain *single user scheduling* at each BS. Thus, the user

[6]This is same as the approach we took while formulating the joint problem $[\mathbf{P_{Joint}}(\boldsymbol{K}, \boldsymbol{P})]$.

scheduling process is completely characterized by $\{\beta_{ji}\}_{j\in\{0\}\cup\mathcal{P},i\in\mathcal{N}}$, where β_{ji} denotes the fraction of time BS j schedules user i. Then, our global α-fair user scheduling problem corresponds to finding the values of $\{\beta_{ji}\}$'s such that $\sum_{i\in\mathcal{N}} U_\alpha(\lambda_i)$ is maximized, where λ_i is the throughput offered to user i. Of particular interest is the case of $\alpha = 1$, as used in [55] which yields the global proportional fair (PF) scheduling problem.

Formally, the global scheduling problem can be stated as follows: given network realization ω, RA (K), and UA ($\{x_{ji}\}$) (i.e., given the rates $R_{ji}(\omega)$ and the set of UEs associated to BS j ($A_j(\omega)$)), find the optimal values of $\{\beta_{ji}\}$ by solving the following[7].

$$[\mathbf{P}(\omega, \{x_{ji}\}, K)] \quad \max_{(\lambda_i),(\beta_{ji})} \sum_{i\in\mathcal{N}} U_\alpha(\lambda_i)$$

$$\text{subject to: } \lambda_i = \sum_{j\in\mathcal{P}\cup\{0\}} R_{ji}(\omega)\beta_{ji}, \forall i \in \mathcal{N} \tag{6.5}$$

$$\sum_{i\in\mathcal{N}} R_{ji}(\omega)\beta_{ji} \le C_j, \quad \forall j \in \mathcal{P} \tag{6.6}$$

$$\sum_{j\in\mathcal{P}\cup\{0\}} \sum_{i\in\mathcal{N}} R_{ji}(\omega)\beta_{ji} \le C_{BH} \tag{6.7}$$

$$\sum_{i\in\mathcal{N}} \beta_{ji} \le 1, \qquad \forall j \in \mathcal{P}\cup\{0\} \tag{6.8}$$

$$0 \le \beta_{ji} \le x_{ji}, \forall i \in \mathcal{N}, \forall j \in \mathcal{P}\cup\{0\} \tag{6.9}$$

For brevity, we omit $\{x_{ji}\}$ and K from the problem name, and call it $[\mathbf{P}(\omega)]$.

(6.5) relates user schedules to throughputs, (6.6) is the constraint due to finite backhaul capacities at each small cell. (6.7) is the constraint due to the limited capacity of the MBS backhaul, which limits the total flows on all BSs. (6.8) represents the scheduling constraints at each BS. Note that the mention of ω in the parenthesis of the optimization problem name is done to stress on the fact that the given problem is realization-dependent.

We can show that maximizing the sum of the α-fair utility is equivalent to maximizing

[7]Note that this problem is much simpler than the joint problem $[\mathbf{P}_{\mathbf{Joint}}(\boldsymbol{K}, \boldsymbol{P})]$.

the following throughput-based metric.

$$\overline{T}_\alpha(\{\lambda_i\}_{i\in\mathcal{N}}) = \left(\frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}}\lambda_i^{1-\alpha}\right)^{\frac{1}{1-\alpha}}, \alpha > 0, \alpha \neq 1$$
$$= \left(\prod_{i\in\mathcal{N}}\lambda_i\right)^{\frac{1}{|\mathcal{N}|}}, \qquad \alpha = 1 \tag{6.10}$$

For PF (i.e., $\alpha = 1$), this metric $\overline{T}_1(.)$ represents the geometric mean (GM) of user throughputs. We will refer to $\overline{T}_\alpha(.)$ simply as the *α-mean throughput*.

We identify three scenarios: *Scenario 0, Scenario 1, and Scenario 2*. Scenario 0 is the scenario where the capacities of both the MBS backhaul and the SC backhaul links are large enough not to be bottlenecks , this is true in particular if $C_j > Kr_{max}$, and $C_{BH} > (XK + (M - K))r_{max}$ where $r_{max} = \max_{\gamma\geq 0}\theta(\gamma)$ is the highest value of the rate function). Scenario 1 represents the scenario where the SC backhaul capacities are limited and the MBS backhaul capacity is not constraining. Scenario 2 is the most general scenario where all backhaul links have capacities that are constraining.

Prior work exists for a version of this problem for $\alpha = 1$ (i.e., global PF) for scenario 0 (i.e., without considering the backhaul limitations (6.6) and (6.7)). Fooladivanda and Rosenberg in [37] have shown that the following properties hold.
1) *Decomposability*: The global problem for $\alpha = 1$ can be decoupled into a set of $X + 1$ *independent* local PF problems, one per each BS. A local problem for BS j tries to maximize its own local sum of utilities ($\sum_{i\in A_j(\omega)} U_\alpha(\lambda_i)$), without regard to how the scheduling is done in other BSs. A local scheduling solution at BS j depends only on its local information (e.g., values of channel gains of its own users $A_j(\omega)$) which we will refer to as the *local realization dependence*, as opposed to the *global realization dependence* in which schedules in a BS would depend on channel gains in other BSs. *Local realization dependence* is a desirable property.
2) *Equal-time equivalence*: Under the stated assumptions, a local PF scheduling at BS j is equivalent to an *equal-time* scheduling where each user $i \in A_j(\omega)$ is allocated $\frac{1}{|A_j(\omega)|}$ fraction of time.

In this chapter, we build on this prior work and study the problem under a more general

α-fairness objective, and under limited backhaul capacities.

6.4 Scenario 0: $\{C_j\}$'s and C_{BH} are very large

The following theorem states our results for Scenario 0.

Theorem 1 (Scheduling under Scenario 0)**.** *If all backhaul links are very large,*
a) Decomposition: The global problem ***[P(ω)]*** *can be decoupled into a set of $X+1$ independent local α-fair problems, one per each BS, where the local problem for BS j is*

$$[\mathbf{P^j_{Local}}(\omega)] : \max_{\{\beta_{ji}\geq 0\}_{i\in A_j(\omega)}} \sum_{i\in A_j(\omega)} U_\alpha(R_{ji}(\omega)\times\beta_{ji})$$
$$s.\ t. \sum_{i\in A_j(\omega)} \beta_{ji} \leq 1; \quad \beta_{ji}\geq 0 \tag{6.11}$$

b) Closed-form solution: The following schedule is optimal for the local problem $[\mathbf{P^j_{Local}}(\omega)]$.

$$\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i'\in A_j(\omega)} T_{ji',\alpha}(\omega)}, \forall i\in A_j(\omega), \forall j\in\{0\}\cup\mathcal{P} \tag{6.12}$$

where $T_{ji,\alpha}(\omega) \triangleq R_{ji}(\omega)^{\frac{1-\alpha}{\alpha}}$.

Proof. The proof is shown in Appendix A.2. □

This result means that scheduling is very simple for Scenario 0. The result is the generalization of the known result for $\alpha = 1$, where the local scheduler is the *equal-time* scheduler[8] .

[8]The equivalence of PF scheduling to an equal-time solution has been the basis of many algorithms. We consider static channels. However, note that, many schedulers that exploit channel dynamics by selecting users with good instantaneous link-rates maintain the notion of proportional fairness by guaranteeing equal-time scheduling *asymptotically.*

6.5 Scenario 1: C_{BH} is very large while $\{C_j\}$'s are not

When C_{BH} is very large, the constraint (6.7) (*MBS backhaul constraint*) can be removed from the optimization problem $[\mathbf{P}(\omega)]$. Let us call this relaxed problem as $[\mathbf{P}_\infty(\omega)]$. $[\mathbf{P}_\infty(\omega)]$ can be decomposed into a set of local α-fair scheduling problems, one per BS. The local scheduling problem for the MBS is $[\mathbf{P^0_{Local}}(\omega)]$, which is the simple local α-fair scheduling problem without backhaul limitations, defined earlier. SC j should solve the local α-fair scheduling problem with backhaul limitations, shown below.

$$[\mathbf{P^j_{Local}}(\omega,C_j)] : \max_{\{\beta_{ji}\}_{i\in A_j(\omega)}} \sum_{i\in A_j(\omega)} U_\alpha(\beta_{ji}R_{ji}(\omega)) \text{ s.t.}$$

$$\sum_{i\in A_j(\omega)} \beta_{ji} \le 1, \qquad (\zeta_{j,\omega}) \tag{6.13}$$

$$\sum_{i\in A_j(\omega)} \beta_{ji}R_{ji}(\omega) \le C_j, \qquad (\mu_{j,\omega}) \tag{6.14}$$

$$\beta_{ji} \ge 0, \quad \forall i \in A_j(\omega) \qquad (l_{j,i,\omega}) \tag{6.15}$$

where $\zeta_{j,\omega}$, $\mu_{j,\omega}$, and $l_{j,i,\omega}$ are the dual variables of the scheduling constraint (6.13), the total-flow constraint (6.14), and the non-negativity constraint of user schedules, respectively.

In other words, under Scenario 1, BS j schedules its users independently of other BSs with only its local information (its own backhaul link capacity C_j, and channel gains G_{ji} of its own users only), and thus there is no need for a global entity to assist in the stated decomposition. (6.12) can be used to obtain the optimal solution of $[\mathbf{P^0_{Local}}(\omega)]$. In the next subsection, we will derive the solution to the local α-fair scheduling problem $[\mathbf{P^j_{Local}}(\omega,C_j)]$.

6.5.1 Local α-fair scheduling under backhaul limitation

If we define the following two critical values of the backhaul capacity for BS j, and realization ω,

$$c^*_j(\omega) \triangleq \frac{|A_j(\omega)|}{\sum_{i\in A_j(\omega)} \frac{1}{R_{ji}(\omega)}}; \quad C^*_{j,\alpha}(\omega) \triangleq \sum_{i\in A_j(\omega)} \frac{R_{ji}(\omega)^{\frac{1}{\alpha}}}{\sum_{i\in A_j(\omega)} T_{ji,\alpha}(\omega)} \tag{6.16}$$

then, the nature of the local α-fair scheduling can be characterized as follows.

Theorem 2. *The local α-fair scheduling $[\mathbf{P^j_{Local}}(\omega,C_j)]$ can be characterized based on how the backhaul capacity C_j compares to the two critical values. There are three regions:*
(a) If $C_j \geq C^*_{j,\alpha}(\omega)$, the scheduler is in Region 1 (which we refer to as backhaul-unlimited (BHU) scheduler), and is given as follows.

$$\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i' \in A_j(\omega)} T_{ji',\alpha}(\omega)}, \quad \forall i \in A_j(\omega) \ \textbf{\textit{[Region 1]}} \tag{6.17}$$

(b) If $C_j \leq c^*_{j,\alpha}(\omega)$, the scheduler is in Region 2 (which we refer to as local equal-throughput scheduler), and is given as follows.

$$\beta_{ji} = \frac{C_j}{|A_j(\omega)| R_{ji}(\omega)}, \qquad \forall i \in A_j(\omega) \quad \textbf{\textit{[Region 2]}} \tag{6.18}$$

(c) If $c^*_j < C_j < C^*_{j,\alpha}$, the scheduler is in Region 3. The optimal dual solution is obtained by solving the following equations for $\mu_{j,\omega} > 0$ and $\zeta_{j,\omega} > 0$.

$$\sum_{i \in A_j(\omega)} \frac{R_{ji}(\omega)^{\frac{1}{\alpha}}}{(\mu_{j,\omega} R_{ji}(\omega) + \zeta_{j,\omega})^{\frac{1}{\alpha}}} = C_j$$
$$\sum_{i \in A_j(\omega)} \frac{T_{ji,\alpha}(\omega)}{(\mu_{j,\omega} R_{ji}(\omega) + \zeta_{j,\omega})^{\frac{1}{\alpha}}} = 1 \quad \textbf{\textit{[Region 3]}}$$

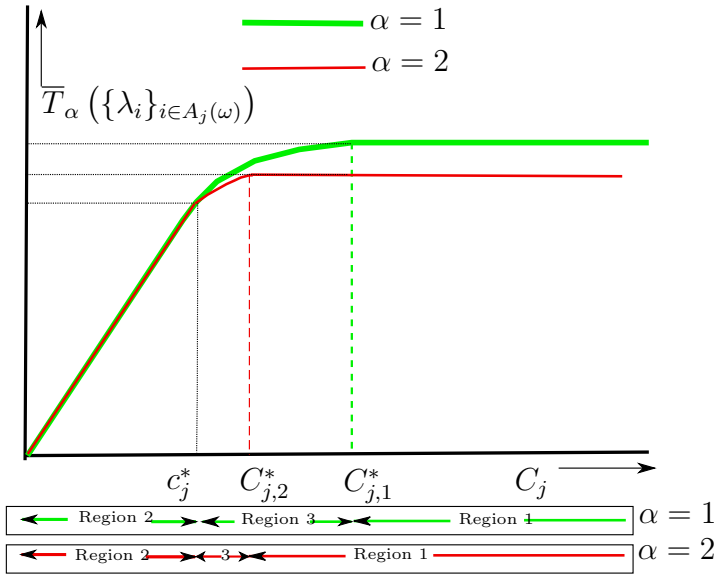
The primal solution is then given as $\beta_{ji} = T_{ji,\alpha}(\omega) \times (\mu_{j,\omega} R_{ji}(\omega) + \zeta_{j,\omega})^{-\frac{1}{\alpha}}$ for all $i \in A_j(\omega)$.

Proof. The proof can be found in Appendix A.1. □

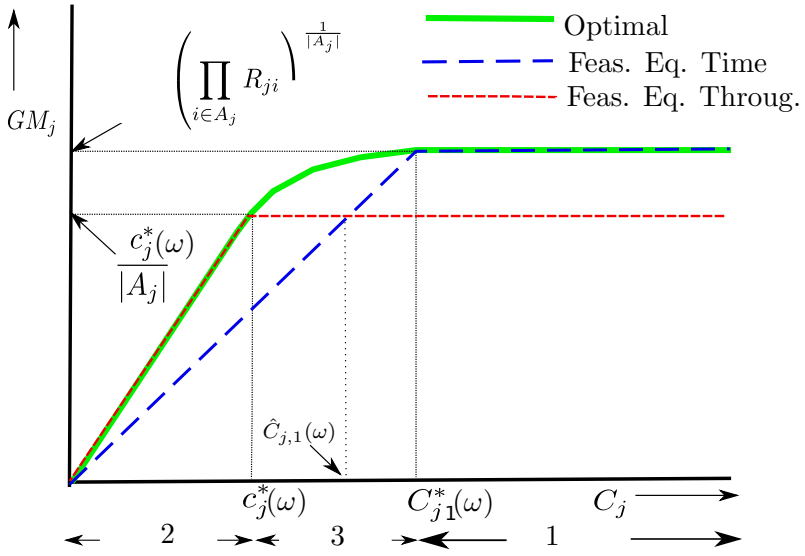
Note that the two critical values are realization-dependent which means that any change in the realization would trigger a need to recompute them.

Interpretation of Theorem 2

In Fig. 6.2a, we show curves that represent the typical shape of the plots of α-mean throughput ($\overline{T}_\alpha(.)$) as a function of the backhaul capacity C_j for a given value of α for one of the small cells $j \in \mathcal{P}$ when the local α-fair scheduling is performed. This figure clearly



(a) Illustration of Theorem 2



(b) Motivation for the heuristic, $\alpha = 1$

Figure 6.2: α-mean throughput versus SC backhaul capacity for a realization

shows the three scheduling regions (Regions 1, 2 and 3) as a function of the two critical values of the backhaul capacity.

For sufficiently large backhaul capacity $C_j \geq C^*_{j,\alpha}(\omega)$, we are in Region 1. For a very limited backhaul capacity $C_j \leq c^*_j(\omega)$, we are in Region 2. For intermediate values of the backhaul capacity $c^*_j(\omega) < C_j < C^*_{j,\alpha}(\omega)$, we are in Region 3.

Region 1: For each value of α, there is a critical value of the backhaul capacity $C^*_{j,\alpha}$ such that any more capacity of backhaul link does not translate to a better performance. This is shown as Region 1 in the figure. It is important to note that, for a given set of user rates, this critical value is different for different values of α. Note that in this region, the scheduler is the same as the *backhaul-unlimited* (BHU) scheduler defined for Scenario 0. As an aside, note that $C^*_{j,\alpha}(\omega)$ is also the smallest value of the backhaul capacity C_j for which the backhaul link is no longer a bottleneck on the performance.

Region 2: If $C_j \leq c^*_j(\omega)$, we have $\beta_{ji}R_{ji}(\omega) = \frac{C_j}{|A_j(\omega)|}$ for all $i \in A_j(\omega)$ (from Theorem 2(b)). This is a region where users in a given BS are offered equal throughput $\frac{C_j}{|A_j(\omega)|}$. Thus for $C_j \leq c^*_j(\omega)$, a local *equal-throughput* scheduling is equivalent to the local α-fair scheduling. It is interesting to note that, unlike $C^*_{j,\alpha}(\omega)$, this critical value is independent of α and so is the scheduler. In other words, all α-fair local schedulers operate identically when $C_j \leq c^*_j(\omega)$. In Fig. 6.2a, they would all have the same Region 2.

Region 3: For $c^*_j(\omega) < C_j < C^*_{j,\alpha}(\omega)$, neither *local equal-throughput* nor *backhaul-unlimited* *α-fair* scheduling is optimal. The optimal solution to the local α-fair scheduler has to be obtained by computing the solution to the equations in Theorem 2(c). Note that for $\alpha \to \infty$ (i.e., the *max-min* case), Region 3 does not exist.

6.5.2 Simple heuristic

When the scheduling is in Region 1 or Region 2, the variables have closed-form solutions as given in (6.17) and (6.18), and hence are very easy to compute. In Region 1, backhaul-unlimited α-fair scheduling is optimal whereas in Region 2, a local equal-throughput scheduling with throughput of $\frac{C_j}{|A_j(\omega)|}$ is optimal. We do not have closed-form solutions for Region 3, where we need to numerically solve the set of non-linear equations in Theorem 2(c). A scheduler preferably with closed-form solutions for all regions would

be desirable.

We propose the following simple heuristic: *take the best of two easy-to-compute feasible schedulers.*

1. The first one is a *feasible* version of the equal-throughput scheduler, i.e., a solution to the local problem with the following constraint $R_{ji}(\omega)\beta_{ji} = R_{ji'}(\omega)\beta_{ji'}$ for all $i, i' \in A_j(\omega)$. The solution to this feasible local equal-throughput scheduling is $\beta_{ji} = \min\{\frac{C_j}{|A_j(\omega)|R_{ji}(\omega)}, \frac{c_j^*(\omega)}{|A_j(\omega)|R_{ji}(\omega)}\}$ for all $i \in A_j(\omega)$. Note that this scheduler is optimal for Region 2.

2. The second one is a *feasible* (scaled-down) version of the backhaul-unlimited scheduler, i.e., $\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i' \in A_j(\omega)} T_{ji',\alpha}(\omega)} k$, where k is a strictly positive scaling constant that corresponds to the largest value less or equal to 1 that guarantees feasibility of the local problem. This problem is solved by $\beta_{ji} = \min\left\{\frac{C_j T_{ji,\alpha}(\omega)}{\sum_{i' \in A_j(\omega)} R_{ji'}(\omega)^{\frac{1}{\alpha}}}, \frac{T_{ji,\alpha}(\omega)}{\sum_{i' \in A_j(\omega)} T_{ji',\alpha}(\omega)}\right\}$ for all $i \in A_j(\omega)$. This scheduler is optimal for Region 1.

The rational behind our heuristic is illustrated in Fig. 6.2b. This approach results in a much simpler scheduler as compared to the optimal one because of the closed-form scheduling solutions. Of course, we need to verify that this simplification does not result in a significant loss in performance. We will see how this scheme performs in realistic network settings while presenting the numerical results in the next subsection.

Further properties of the local problem: We now present some properties of the local problem that will be used in the analysis of Scenario 2. Let, $f_{j,\omega}(C_j)$ be the optimal value of $[\mathbf{P^j_{Local}}(\omega, C_j)]$. Also, let $\frac{\partial f_{j,\omega}(C_j)}{\partial C_j} \triangleq f'_{j,\omega}(C_j)$ be the rate at which the optimal value changes with C_j. Then, we can show that the following holds.

Lemma 1. *The rate of change of $f_{j,\omega}(C_j)$ with respect to the backhaul capacity C_j is given*

as follows:

$$\frac{\partial f_{j,\omega}(C_j)}{\partial C_j} = \begin{cases} \left(\frac{|A_j(\omega)|}{C_j}\right)^{\alpha} & \text{if } C_j \leq c_j^*(\omega) \\ \mu_{j,\omega}^*(C_j) & \text{if } c_j^*(\omega) < C_j < C_{j,\alpha}^*(\omega) \\ 0 & \text{if } C_j \geq C_{j,\alpha}^*(\omega) \end{cases} \tag{6.19}$$

where $\mu_{j,\omega}^*(C_j)$ is the optimal value of the dual variable $\mu_{j,\omega}$ for backhaul capacity C_j.

Proof. Please see Appendix A.3. □

Also, note that $f_{j,\omega}(C_j)$ is a concave, non-decreasing function of C_j in $(0, \infty)$. In particular, $f_{j,\omega}(C_j)$ is strictly increasing in $(0, C_{j,\alpha}^*(\omega)]$. $\frac{\partial f_{j,\omega}(C_j)}{\partial C_j} = f'_{j,\omega}(C_j)$ is a strictly decreasing function of C_j in $(0, C_{j,\alpha}^*(\omega)]$.

6.5.3 Numerical results

We consider a hexagonal HetNet deployment area with each side equal to $500/\sqrt{3}\, m$, which corresponds to the scenario of an inter-site distance (ISD) of $500m$ (urban setting). The centrally placed MBS is overlaid with $X = 4$ *symmetrically placed* small cells ($j = 1, 2, 3, 4$) at a distance of $d = 178\ m.$ from the MBS. An MBS transmit power budget P_M of 46 dBm and an SC transmit power budget P_S of 30 dBm are considered. The overall system has $M' = 99$ subchannels and the reuse factor of $r = 3$. Hence there are $M = 33$ subchannels available to each macro cell, out of which K subchannels are allocated to each small cell and the remaining $M - K$ subchannels are allocated to the MBS. The interference from the outer macro cells is calculated by considering 18 identical macro cells around the given macro cell, and by assuming that identical channel splitting (K) is employed in the interfering macro cells. We only consider the interference the 4 small cells create for each other. We use the distance-based path-loss model recommended by 3GPP [6], as shown in Table 6.2.

The channel gains G_{ji} are obtained by further applying a log-normal shadowing of 8 dB standard deviation. A random realization ω corresponds to a realization of channel gains

Table 6.2: Physical layer parameters

UE Noise Power	-174 dBm/Hz	Channel BW	180 KHz
UE Noise-figure	9dB	n_{sc}	12
UE Pen. Loss	20 dB	n_{ts}	14
T_{symbol}	1 ms		
MBS Ant. Gain	15 dBi	SC Ant. Gain	5 dBi
MBS-UE Path-loss	$128.1 + 37.6 \log_{10}(d/1000), d \geq 35m$		
SC-UE Path-loss	$140.7 + 36.7 \log_{10}(d/1000), d \geq 10m$		

for a random instance of uniformly deployed N equal to 30 user positions and randomly generated shadowing coefficients.

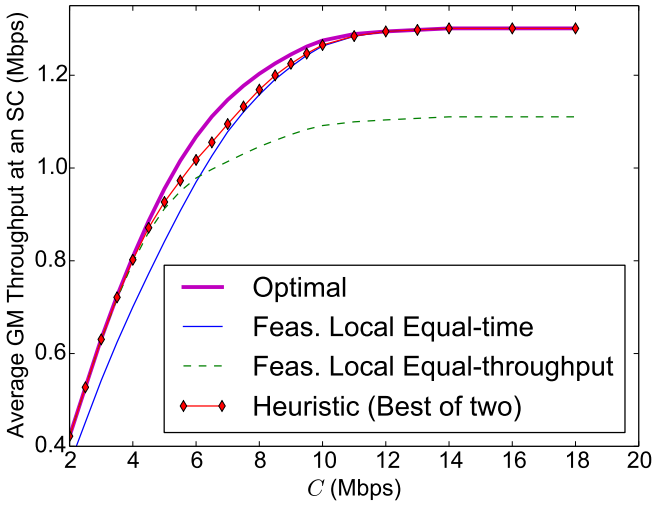
The rate function $\theta(.)$ is taken as the 15-rate MCS available in LTE, as shown in Table 4.2. We take $N_0 = -112.45dBm$ as the noise power per subchannel (i.e., a noise of $-174dBm/Hz$ with a noise figure of 9 dB).

We consider scenarios where the small cells are identical, i.e., they all have the same backhaul capacities C_j equal to C. Also, recall that in this scenario, C_{BH} is sufficiently large.

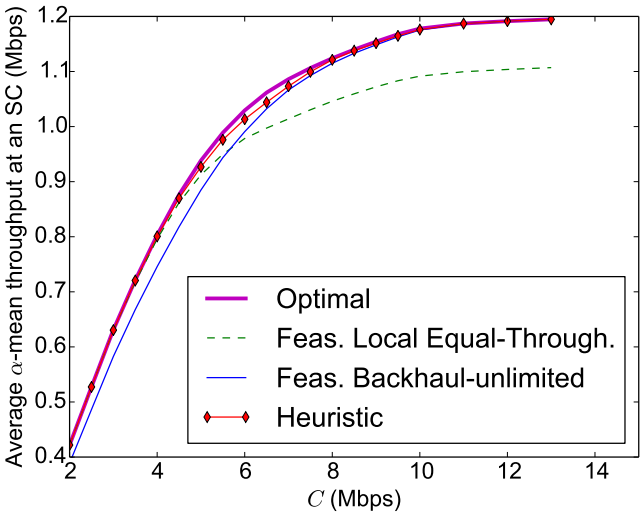
We study 100 random realizations $\omega \in \Omega$ of user positions. The average of the α-mean throughput $\overline{T}_\alpha(.)$ over these realizations is the metric for comparison of the different schemes (optimal and sub-optimal).

PF Scheduling ($\alpha = 1$)

In Fig. 6.3a, we plot the average GM throughput (which is the α-mean throughput for $\alpha = 1$) of users in a given SC as a function of SC backhaul capacity C for $K = 15$, $\delta = 6.6dB$. The impact of limited SC backhaul capacity on the throughput performance is significant. Let us concentrate on the optimal scheduling scheme. We can see that after a certain point, increasing capacity C does not translate to a significant improvement on the throughput performance. This is expected due to the concavity of $f_{j,\omega}(C_j)$, and the fact that $\frac{\partial f_{j,\omega}(C_j)}{\partial C_j}$ is equal to 0 for $C_j \geq C^*_{j,\alpha}(\omega)$. Also, it can be observed that, for this particular scenario, there is a value of C (about 12 Mbps, shown by the vertical dashed line) after which there is effectively no improvement in system performance as we increase



(a) Average GM throughput versus C, Local PF ($\alpha = 1$)



(b) Average α-mean throughput versus C, $\alpha = 2$

Figure 6.3: Comparison of the optimal and the sub-optimal local α-fair schedulers

the backhaul capacity. This value of C can be considered as the *sufficient capacity* of the SC backhaul link to guarantee that the SC backhaul link is no longer a bottleneck to system performance.

The figure also illustrates that using either a local equal-throughput or a local equal-time scheduling regardless of the backhaul capacity can result in a significant loss in performance. The same figure shows that our heuristic, where a BS chooses the best of local feasible equal-throughput and local feasible equal-time for each realization, works remarkably well. This scheduler achieves a performance very close to optimal and yet is quite simple to compute. This heuristic can thus be seen as good *backhaul-aware* local PF scheduler.

$\underline{\alpha = 2}$

In Fig. 6.3b, we plot results for $\alpha = 2$. This value of α maximizes the harmonic mean of user throughputs and is often called the *minimum potential-delay scheduling*. The plot shows similar results, in particular, it shows that our proposed heuristic is a good approximation of the local α-fair scheduler.

6.6 Scenario 2: $\{C_j\}$'s and C_{BH} are not very large

6.6.1 Optimal scheduler

Similar to Scenario 1, it would be desirable to decompose the global problem for Scenario 2 into independent local problems. However, unlike Scenario 1, decomposing the global problem into local problems is not straightforward, mainly due to the coupling constraint (6.7) (the MBS backhaul constraint). Indeed, allowing each BS j to independently schedule based on its own local problem ($[\mathbf{P^j_{Local}}(\omega,C_j)]$) could lead to violation of the MBS backhaul constraint (6.7). Thus, in order to obtain a decomposition that is feasible, we need to guarantee that the local problems do not violate the MBS backhaul constraint. This can be accomplished by defining the notion of *virtual backhaul capacities* $\tilde{C}_j$ for each BS $j \in \{0\} \cup \mathcal{P}$ where $\tilde{C}_j$ is used by the local scheduler at BS j as the actual available local capacity (as opposed to C_j). The vector $\tilde{\boldsymbol{C}} = [\tilde{C}_0, \cdots, \tilde{C}_j, \cdots, \tilde{C}_X]$ is considered to be

feasible if it satisfies the following conditions.

$$\sum_{j\in\{0\}\cup\mathcal{P}} \tilde{C}_j \leq C_{BH}; \quad \tilde{C}_j \leq C_j, \quad \forall j \in \mathcal{P} \tag{6.20}$$

Given a feasible vector $\tilde{\boldsymbol{C}}$, if the local scheduler at BS j solves $[\mathbf{P^j_{Local}}(\omega,\tilde{C}_j)]$ with $\tilde{C}_j$ as its local backhaul constraint without regard to other local problems, the end result is a feasible solution to the global problem $[\mathbf{P}(\omega)]$. Hence, as long as a *master* problem can provide one such feasible $\tilde{\boldsymbol{C}}$, the solutions due to the local schedulers would yield a feasible solution to the global problem, thereby yielding a *feasible* decomposition. Our first goal is to find an *optimal* decomposition, i.e., one that would yield the optimal solution of problem $\mathbf{P}[\omega)]$. Solving the following master problem will provide the values of $\tilde{C}_j$ corresponding to the optimal decomposition.

$$\max_{\tilde{\boldsymbol{C}}\geq 0} \sum_{j\in\{0\}\cup\mathcal{P}} f_{j,\omega}(\tilde{C}_j) \text{ s.t. } (6.20) \tag{6.21}$$

Recall that $f_{j,\omega}(\tilde{C}_j)$ is the value of the local problem $[\mathbf{P^j_{Local}}(\omega,\tilde{C}_j)]$ at BS j where $\tilde{C}_j$ is the backhaul capacity.

Without loss of generality, we will assume that the MBS solves this master problem. In our tree topology, it is indeed the most natural place to compute the solution to the master problem. This formulation can be seen as a *two-level* problem in which small cells report their user channel gain information $(\{G_{ji}(\omega)\}_{i\in A_j(\omega)})$ to the MBS which computes and reports back to them the optimal values of the virtual capacities $\tilde{C}_j$. BS j can then perform the local scheduling by considering the reported $\tilde{C}_j$ as the available backhaul capacity (as opposed to C_j), i.e., solving $[\mathbf{P^j_{Local}}(\omega,\tilde{C}_j)]$. This two-level decomposition solves the global problem optimally, and can be seen as an alternative formulation to $[\mathbf{P}(\omega)]$. Note that presenting the original global problem as a two step problem does not simplify its computational complexity but allows us to propose a good heuristic later.

Alternatively, a distributed approach could be used where the small-cells would report the subgradients and the MBS would update them with the value of Λ. Such process would eventually converge to the optimal solution. This way, there is no need to collect the channel gains, but now the overhead would be on the exchange of Λ, and one gradient

per BS at each step until convergence. The efficiency would depend on how fast the process converges.

Alternative dual-based formulation

We can rewrite the master problem as: $\max_{\tilde{\boldsymbol{C}}\geq 0} \sum_{j\in\{0\}\cup\mathcal{P}} \tilde{f}_{j,\omega,C_j}(\tilde{C}_j)$ s.t. $\sum_{j\in\{0\}\cup\mathcal{P}} \tilde{C}_j \leq C_{BH}$, where $\tilde{f}_{j,\omega,C_j}(\tilde{C}_j) = \min\{f_{j,\omega}(\tilde{C}_j), f_{j,\omega}(C_j)\}$ for all $j \in \{0\}\cup\mathcal{P}$, and where C_0, which is not a physical constraint but is here for consistency, is equal to a large value (e.g., greater than C_{BH}). This modified problem has only one dual variable, which we call Λ. Then, solving the following dual problem is equivalent to solving the master problem,

$$\min_{\Lambda\geq 0} \quad \max_{\tilde{\boldsymbol{C}}\geq 0} \quad L(\tilde{\boldsymbol{C}};\Lambda) \tag{6.22}$$

where $L(\tilde{\boldsymbol{C}};\Lambda) = \sum_{j\in\{0\}\cup\mathcal{P}} \tilde{f}_{j,\omega,C_j}(\tilde{C}_j) - \Lambda\left(\sum_{j\in\{0\}\cup\mathcal{P}} \tilde{C}_j - C_{BH}\right)$ is the Lagrangian function.

The following result allows us to obtain the solution to the dual problem.

Theorem 3.

$$\Lambda^*(\omega) = \min\left\{\Lambda\geq 0 : \sum_{j\in\{0\}\cup\mathcal{P}} \tilde{C}^D_{j,\omega,C_j}(\Lambda) \leq C_{BH}\right\}$$

is the optimal value of Λ in problem (6.22), where $\tilde{C}^D_{j,\omega,C_j}(\Lambda) \triangleq \min\{f'^{(-1)}_{j,\omega}(\Lambda), C_j\}$ for all $j \in \{0\}\cup\mathcal{P}$ is a mapping from dual variable Λ to primal variable $\tilde{C}_j$, $f'^{(-1)}_{j,\omega}(\Lambda)$ is the inverse mapping of $f'_{j,\omega}(C_j)$ defined in (6.19), and is given as follows:

$$f'^{(-1)}_{j,\omega}(\Lambda) = \begin{cases} \frac{|A_j(\omega)|}{\Lambda^{\frac{1}{\alpha}}} & \Lambda \geq \left(\frac{|A_j(\omega)|}{c^*_j(\omega)}\right)^{\alpha} \\ \mu^{*(-1)}_{j,\omega}(\Lambda) & 0 < \Lambda < \left(\frac{|A_j(\omega)|}{c^*_j(\omega)}\right)^{\alpha} \\ C^*_{j,\alpha}(\omega) & \Lambda = 0 \end{cases} \tag{6.23}$$

where $\mu^{*(-1)}_{j,\omega}(\Lambda)$ is the inverse of $\mu^*_{j,\omega}(C_j)$ which is the dual variable of the local problem $[\mathbf{P}^{\mathbf{j}}_{\mathbf{Local}}(\omega,C_j)]$ as defined earlier, and $(c^*_j(\omega),\ C^*_{j,\alpha}(\omega))$ are the critical values defined in (6.16).

Proof. Please see Appendix A.4. □

Even though computationally similar, there is a benefit of looking at the dual version as opposed to the primal problem: we can find a mapping from the dual variable to the virtual backhaul capacity $\tilde{C}_j$ allowing us to express the primal optimal solutions based on the optimal value of Λ.

This problem can be solved for one scalar value of Λ by employing a simple *bisection-search* for the smallest feasible value of Λ. This is because, the $\tilde{C}^D_{j,\omega,C_j}(\Lambda)$ are non-decreasing as we decrease Λ, as shown in Appendix A.4 (and hence $\sum_{j\in\{0\}\cup\mathcal{P}} \tilde{C}^D_{j,\omega,C_j}(\Lambda)$ is non-decreasing as we decrease Λ).

The details of the bisection search algorithm is presented in Algorithm 1.

Algorithm 1 Compute Optimal Dual Variable $\Lambda^*(\omega)$

Input: $\omega, \{\tilde{C}^D_{j,\omega,C_j}(.)\}_j, C_{BH}$
Define: $g(\Lambda) \triangleq \sum_{j\in\{0\}\cup\mathcal{P}} \tilde{C}^D_{j,\omega,C_j}(\Lambda)$
Begin:
$(\Lambda_{Max}, \Lambda_{Min}) \leftarrow (L, 0)$ ▷ *L :A sufficiently large number*
while $|\Lambda_{mid} - \Lambda_{Max}| < \epsilon$ **do** ▷ *ϵ :A small positive number*
$\Lambda_{Mid} \leftarrow \frac{\Lambda_{Max}+\Lambda_{Min}}{2}$
if $g(\Lambda_{Mid}) \leq C_{BH}$ **then** $\Lambda_{Max} \leftarrow \Lambda_{Mid}$
else $\Lambda_{Min} \leftarrow \Lambda_{Mid}$
end if
end while
Return Λ_{Max}

In the dual framework, we can thus view the global optimization as follows: The MBS computes the optimal dual variable $\Lambda^*(\omega)$ and sends this value to the SCs. Each SC computes its virtual backhaul capacity $\tilde{C}_j = \tilde{C}^D_{j,\omega,C_j}(\Lambda^*(\omega))$ for the given dual variable and then performs its local scheduling using this computed value. Besides this concise representation of the optimal solution, the dual formulation also serves as the basis for a very good heuristic that we will discuss later.

6.6.2 Complexity and overhead versus performance trade-off

The optimal values of $\tilde{C}_j$ (either computed using the primal master problem or by using the dual version) are **global realization-dependent**. By global realization-dependence, we mean that the values of $\tilde{C}_j$ change with a change in the network realization. Such a change in realization could be due to various factors: mobility, change in channel gains, user arrival or departure etc. There are at least two aspects of the optimal scheduler that are undesirable:
1) **Computational Complexity:** Computationally, the master problem is as complex as the global problem $[\mathbf{P}(\omega)]$ which is a convex optimization problem of size $\Theta(XN)$ with $\Theta(XN)$ constraints. The complexity of an interior-point method for solving a convex optimization problem is known to be polynomial on the problem size [20]. In our case, the problem size (the number of variables or constraints) increases linearly with the number of users N. Thus, for medium to large values of N, quick computation of the optimal $\tilde{C}_j$ can be a challenge.
2) **Amount and frequency of information exchange (overhead)**: The master problem needs the information of the channel gains from all users in all BSs. The optimal problem (either the primal or the dual version) is complex as it requires re-computation of the master problem each time a realization changes.

Thus, in a highly dynamic scenario, the approach of recomputing the optimal values of $\{\tilde{C}_j\}$ for every change in network realization will not be practical due to the large and frequent information exchanges required between the SCs and the MBS.

Note that other key parameters such as the resource allocation parameter K or the UA parameter δ can also change with time. However, the time-scale at which these parameters change is usually much larger than the time-scale at which the realizations change. So, in the remainder of this section, we assume that K and δ are fixed and do not change with time.

It would be desirable to have a scheme that overcomes the aforementioned issues by: 1) having a simpler master problem (with reduced problem size), and 2) requiring less overhead (i.e., less amount of and less frequent information exchange between the MBS and the SCs) for solving the master problem. These simplifications come at the expense of

some loss in throughput performance. Finding the right amount of trade-off between the throughput performance and the complexity/overhead is important.

We can define a class of *realization-agnostic* schemes where the virtual backhaul capacities are kept fixed all the time, even when the network realization changes, with changing channel gains as well as changing number of users in the system. Given these fixed values of the virtual backhaul capacities, the global problem can then be decomposed into per-BS local problems, thereby not requiring any information exchange between the MBS and the SCs.

We will take two approaches to choosing the realization-independent values of the virtual backhaul capacities:

Average virtual backhaul capacity based approach: We could generate offline a set of realizations in Ω' where each realization $\omega \in \Omega'$ has a random number of users N. We can then compute the optimal values of the virtual backhaul capacities for each of these realizations. We can then take the average of these virtual capacities as our fixed values, i.e.,

$$\overline{C}_j = \frac{1}{|\Omega'|} \sum_{\omega \in \Omega'} \tilde{C}^*_j(\omega), \forall j \in \mathcal{P} \cup \{0\}$$

Dual-based approach: We can replace the master problem by an approximate problem that is realization-independent. For example, in the dual version of the master problem, we can replace the realization-dependent mapping $\tilde{C}^D_{j,\omega,C_j}(.)$ by a realization-independent mapping $\overline{C}^D_{j,C_j}(.)$. Below, we derive one such mapping.

From Theorem 3, we know that in the interval $[\left(\frac{N_j(\omega)}{c^*_j(\omega)}\right)^\alpha, \infty)$, $C^D_{j,\omega,C_j}(\Lambda)$ is a non-increasing function of Λ with dependence on $N_j(\omega)$, and is equal to $\min\{C_j, \frac{N_j(\omega)}{\Lambda^{\frac{1}{\alpha}}}\}$. Also, $C^D_{j,\omega,C_j}(\Lambda)$ is a non-linear non-increasing function in the interval of $[0, \left(\frac{N_j(\omega)}{c^*_j(\omega)}\right)^\alpha]$ (that depends on the actual rates $\{R_{ji}(\omega)\}$), decreasing from $\min\{C_j, C^*_{j,\alpha}(\omega)\}$ for $\Lambda = 0$ to $\min\{C_j, c^*_j(\omega)\}$ for $\Lambda = \left(\frac{N_j(\omega)}{c^*_j(\omega)}\right)^\alpha$. If we replace the instantaneous values of $c^*_j(\omega)$, $C^*_{j,\alpha}(\omega)$, and $N_j(\omega)$ by the average values of these quantities, we could achieve our goal of replacing $C^D_{j,\omega,C_j}(\Lambda)$ by functions of Λ that do not depend on the realization, as follows:

Given channel allocation parameter K, UA parameter δ, and a set of realizations Ω, we

can compute the *average* values of $C^*_{j,\alpha}(\omega)$, $N_j(\omega)$, and $c^*_j(\omega)$: $\overline{N}_j \triangleq \lim_{|\Omega'|\to\infty} \frac{1}{|\Omega'|} \sum_{\omega\in\Omega'} N_j(\omega)$, $\overline{C}^*_{j,\alpha} \triangleq \lim_{|\Omega'|\to\infty} \frac{1}{|\Omega'|} \sum_{\omega\in\Omega'} C^*_{j,\alpha}(\omega)$, and $\overline{c}^*_j \triangleq \lim_{|\Omega'|\to\infty} \frac{1}{|\Omega'|} \sum_{\omega\in\Omega'} c^*_j(\omega)$. We can then use the following simple relationships between the (approximate) dual variable Λ and the primal variables $\tilde{C}_j$: $\overline{C}^D_{j,C_j}(\Lambda) = \min\{\overline{f}'^{(-1)}_j(\Lambda), C_j\}$ where

$$\overline{f}'^{(-1)}_j(\Lambda) = \begin{cases} \frac{\overline{N}_j}{\Lambda^{\frac{1}{\alpha}}}, & \Lambda \geq \left(\frac{\overline{N}_j}{\overline{c}^*_j}\right)^\alpha \\ \left(\overline{C}^*_{j,\alpha} - \Lambda \times \Delta_j\right), & \Lambda < \left(\frac{\overline{N}_j}{\overline{c}^*_j}\right)^\alpha \end{cases}$$

and $\Delta_j \triangleq \left(\frac{\overline{C}^*_{j,\alpha} - \overline{c}^*_j}{\left(\frac{\overline{N}_j}{\overline{c}^*_j}\right)^\alpha}\right)$.

The dual-based scheme works as follows: The small cells report the measurements on the average values of $(\overline{c}^*_j, \overline{C}^*_{j,\alpha}, \overline{N}_j)$. With these values, the MBS uses the bisection-search algorithm in Algorithm 1 to compute the realization-agnostic values of the virtual backhaul capacities which it sends to the SCs. These values are then kept fixed.

Remark 5. *Note that the dual-based heuristic can be implemented easily as an online algorithm (with no offline tuning required). This can be done by each BS learning the required averages, and reporting these averages once the measurements converge.*

6.6.3 Numerical results

We study how the realization-agnostic schemes work over a set of 500 realizations Ω', where each realization $\omega \in \Omega'$ has a number of users chosen uniformly at random in the interval $[10, 30]$. The users are distributed uniformly at random in the deployment area. Note that, in Section 6.5.3, we considered a set of realizations Ω with a fixed number of users ($N = 30$). But, in this section, we consider realizations with different number of users. This setup encompasses a large set of random realizations in a dynamic network with varying number of users and thus allows us to see if the realization-agnostic scheme works well in a dynamic context. Other than this, we take the same physical layer and network level parameters and setup as in Section 6.5.3.

In Fig. 6.4a-c, we consider the case of proportional fairness (PF) ($\alpha = 1$) with three different values of C_{BH}, i.e., 7, 16, and 30 Mbps, for $K = 15$ and $\delta = 6.6dB$. We present

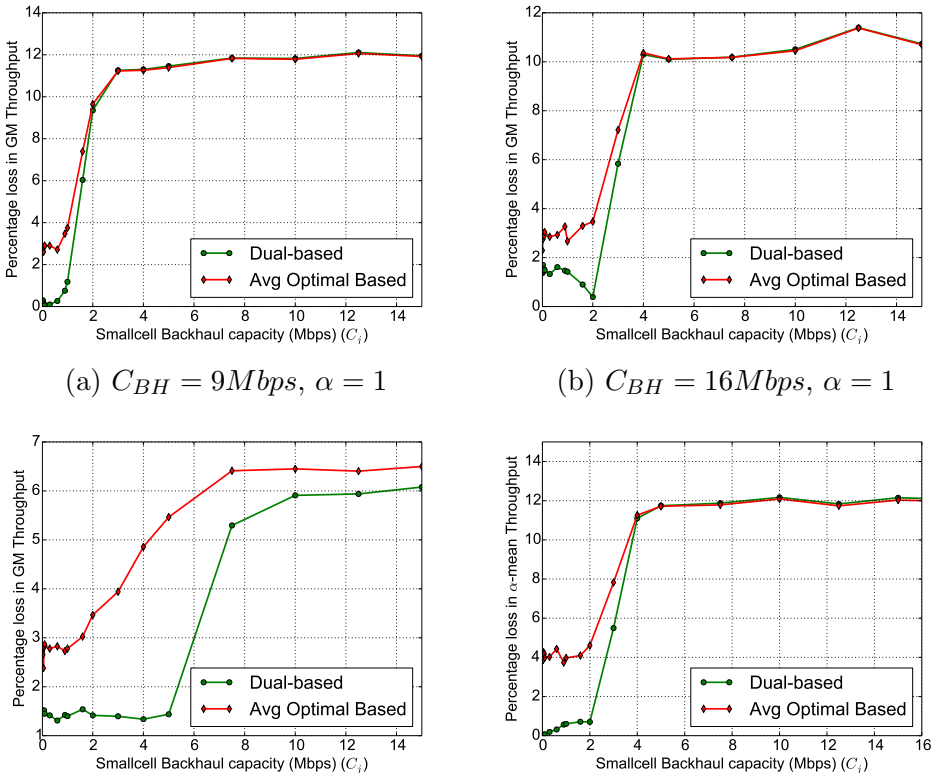
the performance of the sub-optimal schemes in terms of the loss in α-mean throughput performance incurred due to these schemes with respect to the optimal one, for different values of the backhaul capacities. Let $\chi^s(\omega)$ be the α-mean throughput for realization ω for scheme s. Then, the average loss in α-mean throughput for scheme s over the set of realizations Ω' is given as $100 \times \frac{1}{|\Omega'|} \sum_{\omega \in \Omega'} \frac{\chi^{(Opt)}(\omega) - \chi^{(s)}(\omega)}{\chi^{(Opt)}(\omega)}$ where $\chi^{(Opt)}(\omega)$ is the α-mean throughput of the optimal scheme for realization ω.

Observation (*Realization-agnostic schemes work well for* $\alpha = 1$): The results show that the price of using a realization-agnostic scheme is less than 12% for small C_{BH} and decreases when C_{BH} increases. A degradation of less than 12% is a reasonable price to pay, especially since the optimal scheme would be much more complex, and would require a lot of information exchange and a frequent global computation of the optimal solutions. A realization-agnostic scheme, on the other hand, yields independent scheduling at each BS, and thus is a good candidate for an online algorithm.

In Fig. 6.4d, we show similar results for $\alpha = 2$ for $C_{BH} = 16Mbps$. This shows the effectiveness of our heuristic schemes for another value of α.

6.7 Conclusion

In this chapter, we studied the impact of limited backhaul capacity on user scheduling in a heterogeneous network with a macro base station (MBS) overlaid with a number of small cells, inter-connected via a backhaul network deployed in a tree topology. We generalize the results available for proportional fairness under unlimited backhaul capacities to a more general objective of α-fairness, and under different scenarios of backhaul limitations. If each BS could perform its own scheduling locally, it would result in a simple operation of a HetNet. This decoupling of user scheduling processes in different BSs is obtained naturally in a network where the backhaul links do not have capacity limitations, and in such case, each BS can use a simple local scheduler. We have shown that if the limiting factor is the backhaul links between the MBS and the SCs, then each BS can still schedule locally and independently from the other BSs but the local scheduler can take different forms based on the level of capacity limitation. We propose a very simple scheduler that performs well



(a) $C_{BH} = 9Mbps$, $\alpha = 1$

(b) $C_{BH} = 16Mbps$, $\alpha = 1$

(c) $C_{BH} = 30Mbps$, $\alpha = 1$

(d) $C_{BH} = 16Mbps$, $\alpha = 2$

Figure 6.4: Performance of the two realization-agnostic heuristic schemes w.r.t. the optimal scheme, $N \in [10, 30]$

under backhaul limitations.

When the link between the MBS and the core network is also a limiting factor, scheduling becomes much more complex. Each BS can still perform a local scheduling as in the previous case as long as there is a master problem that allocates feasible virtual backhaul capacities to each BS. Doing so in an optimal way is complex and expensive in terms of the amount and frequency of information exchanges but we show that a relatively simple heuristic works very well.

Chapter 7

User Association under Backhaul Limitations

Summary: In this chapter, we study online user association algorithms under the global α-fairness framework, for different scenarios of backhaul limitations.

7.1 Introduction

In the previous chapter, we focused on the impact of backhaul limitations on user scheduling, under the assumption that user association and resource allocation were given. This resulted in the global α-fair scheduling problem $[\mathbf{P}(\omega)]$. We also presented schemes that would make the computation of the user schedules very simple and inexpensive. The optimization model $[\mathbf{P}(\omega)]$ did not consider user association and resource allocation parameters as system variables. In this chapter, we look at the problem of user association, given a scheduling scheme and the resource allocation, using an online approach that we will present next. We present optimal, and sub-optimal user association schemes and discuss various aspects of these schemes. This study provides some interesting insights on the *backhaul-aware α-fair user association*. We also briefly discuss the connections to resource

allocation. The system under consideration is similar to the one in the previous chapter. We thus reuse many notations and definitions from the previous chapter.

7.2 Online approach

In Chapter 3, we presented a model to optimize user association under a static snapshot approach, where all users in a given snapshot are jointly associated to their respective BSs, at once. Such an approach helped us obtain optimal performance. However, such approach leads to an offline study, and does not necessarily reflect the behavior of *online* user association algorithms where the system does not have the opportunity to "reorganize" the way all its users are associated on a regular basis. As a departure from such an approach, we now take an online approach to user association where the UA process is *user-centric* and is called only when necessary, i.e., UA decision for a user is taken by the user itself at the arrival time and at other instants whenever a condition for re-association is satisfied. This is very different from our snapshot problem $[\mathbf{P_{Joint}}(\boldsymbol{K}, \boldsymbol{P})]$ where all user association variables (x_l^f) are optimized together.

The effectiveness of a particular user association scheme is highly dependent on how the network is operating, in terms of, for example, which user scheduling and resource allocation algorithms are used. For example, a UA scheme that is optimal for a network that is using a scheduling algorithm with a particular objective function (say, *maximize the sum-throughput*) is not necessarily optimal for a network using a scheduling algorithm with another objective function (say, *maximize the minimum throughput*). In this chapter, we will study online UA algorithms in systems that are using global optimal α-fair user scheduling.

7.2.1 Node-specific roles, and time-scales

Even though user association is user-centric, many other important processes are network-centric. The distinction between *network-centric and user-centric* is important, and is relevant for our study. For example, user scheduling is often carried out locally at each

BS, whereas resource allocation is often carried out at the cell level. A re-computation of the associated variables of these processes often affects a number of users and nodes (e.g., re-computation of user schedules at BS j affects all users associated to it, whereas re-computation of channel allocation parameter affects many BSs, and their associated users). User association, on the other hand, is often computed by a user based on its measurements and the available information, usually provided periodically by the BSs. Even though a user association decision uses some network-provided information, the decision itself is carried out at a user level. Clearly the association of a new user will affect the users associated to the same BS through scheduling.

An online approach naturally requires the understanding of *time-scales* at which different processes operate. The consideration of time-scales allows us to obtain a certain level of decoupling between different processes (which happen potentially at different nodes, as discussed before). For example, resource allocation parameters (K for OD) are expected to be changed relatively rarely, whereas user scheduling variables are recomputed extremely often. User association, on the other hand, are computed at user arrival instances, and perhaps when the channel conditions change drastically.

7.2.2 State of the art and the general framework

We will first look at some existing examples of online UA schemes. This will allow us to understand the limitations, and lay a ground for a more general framework.

Existing UA schemes and their limitations

Many existing UA schemes are user-centric in nature. The most common approach requires the UE to make physical layer measurements (for example, the SINR) from each candidate BS, and associate to one of them based on some simple criteria/rule. In order to facilitate these measurements, each BS transmits periodic information (e.g., *reference signals* and transmit power level). The main limitation with these UA schemes is that they do not consider:

1. Load-balancing, as they are based solely on the physical layer characteristics,

2. The network's choice of objective function (for example, a change in the α parameter for α-fairness would not affect the user association decision),

3. The backhaul limitations.

Our study is intended to tackle all of these three limitations by *incorporating load-balancing, the network's objective function, and the backhaul limitations* into the user association decision rules. We will first outline our generic framework for user association, before presenting the detailed system model.

A general framework

Even though a UA scheme is user-centric, it will perform its decision with the help of some information available from the BSs. As an example, we discussed how each BS has to transmit periodic signals so that a user can make the channel-related measurements. Most, if not all, of the modern communication systems require their BSs to transmit such periodic signals to facilitate users with their physical layer measurements. In addition to such "mandatory" periodic information, one can design systems where BSs would broadcast some other periodic information about the network which could then be exploited by the UA schemes. The design of a UA scheme is thus affected by the set of available information from the BSs. We can outline a common framework for all user-centric UA schemes, as follows: *the user n, at an instant t (which could be the arrival instant a_n, or an instant at which re-association is deemed necessary), performs the association decision with the help of a set of network-provided information in addition to its own link measurements, and based on a given rule.* The roles played by the BSs and the user are outlined below.

BS j broadcasts a set of BS-specific information $Info_j$ periodically to assist a user to make its UA decision. The set of information available is part of the system design, and is influenced by the UA scheme of choice. We will discuss this in more details later.

UE n:

1. Makes channel-related measurements periodically to estimate G_{jn} and γ_{jn} from a subset of BSs $j \in \{0\} \cup \mathcal{P}$.
2. Averages these values in some way, and decides if there is a need to re-associate or not.
3. If deemed necessary to (re)-associate, uses the available BS-specific information as well as its own measurements to find the best BS j_n^{ua}, based on some predefined rule $\zeta^{ua}(.)$:

$$j_n^{ua} = \zeta^{ua}\left(\{Info_j\}_{j\in\{0\}\cup\mathcal{P}}, \{G_{jn}\}_{j\in\{0\}\cup\mathcal{P}}\right)$$

   We will study three different choices of such a rule in Section 7.4.
4. Sends the association request to BS j_n^{ua}.

7.2.3 Three design aspects of UA schemes

We can identify the following three aspects that are important while designing online UA schemes:

1. *Performance*: Performance of a UA scheme can mean different things, based on the metric of choice. In our study, we assume that the operator choses global α-fairness as the network's objective, and thus two UA schemes are compared based on how well they perform in terms of the α-mean user throughput.
2. *Amount of information* that a UA scheme needs from each BS: Each UA scheme requires a certain amount of information from the BSs. For example, the *best-SINR* UA scheme requires only the SINR-measurement related information, whereas SCF (discussed in Chapter 4) requires the biasing parameter δ also. The best-SINR scheme

is desirable in terms of the amount of information it needs, but can perform very poorly as it has no ability to perform load balancing, nor does it address other network properties like the backhaul limitations. The SCF, on the other hand, can bring some improvement, by fine-tuning the biasing factor. However, this is a heuristic that does not directly address load balancing, network objective function, and backhaul limitations. We can design UA schemes that need more information from the BSs and that can exploit the available information to yield better performance. The right trade-off between performance and the amount of information is a design choice. Finding the right trade-off is often difficult, and one of the main goals of this chapter is to inform this design choice.

3. *Complexity* of the association rule $\zeta^{ua}(.)$: User association has to be done quickly and efficiently. We prefer solutions that have low computation complexity. In general, though, a less complex scheme might come at the cost of performance.

Based on these three aspects, we can list the following three desirable properties of a UA scheme.

1. *Scalable*: A UA scheme will be considered scalable if there are a constant number of BS-specific parameters required to be broadcast by each BS. If the number of parameters a BS has to broadcast increases with an increase in the number of users, the scheme is not scalable. We want a UA scheme to be scalable with as few parameters per BS as possible, as long as the efficiency is good.

2. *Computationally simple*: We want a UA scheme that is very easy to compute, preferably with a constant complexity with respect to the number of users in the system.

3. *Good performance*.

A UA scheme with good performance might require more information and a higher computational complexity, whereas a simple UA scheme might yield poor performance. Next, we present our detailed system model and explore the above-mentioned trade-offs in a more formal settings.

7.3 System Model

Recall the following from Section 6.2

- 0 is the MBS, $\mathcal{P} = \{1, 2, \cdots, X\}$ is the set of SCs, C_j is the capacity of the backhaul link to SC j, and C_{BH} is the capacity of the MBS backhaul. The backhaul infrastructure has a tree topology.
- Orthogonal deployment with channel-split parameter K is used with equal power per-subchannel.
- A realization ω comprises a set of users $\mathcal{N}(\omega)$, with a set of channel-gains $\{G_{ji}(\omega)\}_{j\in 0\cup\mathcal{P}, i\in\mathcal{N}(\omega)}$.

In this chapter, we will introduce time parameter t and represent the network dynamics in terms of t. Let $\omega(t)$ represent the network realization at time t, and let $\mathcal{N}(\omega(t))$ represent the set of users at time t. In other words, given time t, the network realization is uniquely determined. Note that, in the previous chapters, we did not mention the notion of time since we viewed the network in terms of its snapshots ω. By studying a large number of independent snapshots, we hoped to obtain a reasonable approximation of a dynamic behavior. We have already discussed how such an approach can have limitations when it comes to studying user-centric UA schemes.

Users arrive at the deployment area at random times and depart after a certain time when the work they requested from the network is done. Let a_n and d_n respectively be the arrival and the departure times of user n. The system, in terms of the set of users, changes at these instants as follows.

$$\mathcal{N}(\omega(a_n)) = \mathcal{N}(\omega(a_n^-)) \cup \{n\}$$

$$\mathcal{N}(\omega(d_n)) = \mathcal{N}(\omega(d_n^-)) \setminus \{n\}$$

where a_n^- and d_n^- represent the time just before the arrival and the departure of the n^{th} user, respectively. We assume that two arrival or departure events do not occur simultaneously. An UE makes channel-related measurements including the channel gain and the SINR, frequently. Let the SINR and rate seen by user i from BS j at time t be represented as $\gamma_{ji}(\omega(t))$ and $R_{ji}(\omega(t)) = K_j\theta(\gamma_{ji}(\omega(t)))$, respectively defined in (6.1) and (6.3).

7.3.1 Assumptions

We will make the following assumptions:

1. User schedules $\{\beta_{ji}(t)\}$ are (re)-computed by the network based on the backhaul-aware global α-fair scheduling problem studied in the previous chapter. The user schedules are computed whenever the network realization changes.

2. Orthogonal deployment parameter K is changed at a slow time-scale and hence is assumed to be constant in the following.

3. The rate-function $\theta(.)$ is known to the users in advance. In that case, a user i can compute $R_{ji}(\omega(t))$ from all BSs with the knowledge of K (channel allocation parameter) and by measuring its SINR values. SINR measurements are carried out with the help of reference signals.These reference signals are part of the periodic transmissions from the BSs, and are always available.

4. Each newly arriving user has a non-zero rate from at least one BS, i.e., the system provides full coverage.

A user n associates to BS $j_n^* \in \{0\} \cup \mathcal{P}$ whenever an *association event* of user n occurs. Such an event is triggered at the arrival instant a_n, as well as at later times. The k^{th} (re)-association event occurs at time $d_n > \tilde{t}_{nk} > a_n$, and it is triggered whenever some conditions are met (for example, the UE-related measurements suggest a drastic change in the channel conditions). Let $\boldsymbol{\tau}_n \triangleq \{a_n\} \cup \{\tilde{t}_{nk}\}_{k=1,2,\cdots}$ represent the set of instants where an association event occurs for user n. Let, $\tilde{t}_{nk}^-$ represent the time just before the association event at time $\tilde{t}_{nk}$. Let, $x_{ji}(t)$ be an indication of the association between user i and BS j, at time t. A value of 1 represents the fact that user i is associated to BS j, and a value of 0 represents otherwise. The association variable at instance t is determined by the decision carried out at the most recent association event, i.e., $x_{ji}(t) = x_{ji}(\hat{t}_n)$ where $\hat{t}_n = \max\{\tilde{t}_{nk} \in \boldsymbol{\tau}_n : \tilde{t}_{nk} \leq t\}$. We assume that *no two association events for different users occur exactly at the same time.*

7.4 Optimal UA scheme

In the previous section, we presented a generic user association scheme *ua* with a generic rule $\zeta^{ua}(.)$ that determines to which particular BS a user should associate to. In this section, we present the optimal UA scheme for different scenarios and discuss how it can be simple to implement for some scenarios, whereas very complex for others.

Let us define a user association scheme where user n associates to the BS $j \in \mathcal{P} \cup \{0\}$, at time $\tilde{t} \in \boldsymbol{\tau}_n$ (which corresponds to a time where an association event is triggered for user n), so that the global sum of user utilities is maximized, i.e.,

$$\forall n \in \{1, 2, \cdots\}, \forall \tilde{t} \in \boldsymbol{\tau}_n$$
$$j_n^{ua}(\tilde{t}) = \underset{j \in \mathcal{P} \cup \{0\}}{\arg\max} \left\{ \mathbf{P}_{jn}(\omega(\tilde{t})) \right\} \tag{7.1}$$

where $\mathbf{P}_{jn}(\omega(\tilde{t}))$ is the optimal value of the problem $[\mathbf{P}(\omega(\tilde{t}), \{x_{ji}(\tilde{t})\}, K)]$ (defined in Section 6.3) with $x_{jn}(\tilde{t}) = 1$ and $x_{j'n}(\tilde{t}) = 0$ for all other $j' \neq j$. Note that we only associate user n while all the other UEs keep their current association. We call this the *optimal user association scheme*. In other words, each user at the association event chooses the BS such that the global α-fair objective function is maximized[1].

7.4.1 Backhaul-unlimited scenario

Let us first look at the simple scenario of very large backhaul capacities. In this case, we can ignore (6.7), and (6.6) from the optimization problem $[\mathbf{P}(\omega(\tilde{t}), \{x_{ji}(\tilde{t})\}, K)]$. We first consider the case of proportional fairness (i.e., $\alpha = 1$).

$\underline{\alpha = 1}$

The optimal UA scheme for $\alpha = 1$, when there are large backhaul capacities has a simple form, as presented in the following proposition.

[1]Note that this is a *myopic approach* where "optimality" concerns the decision instant only.

Proposition 1. *The optimal UA scheme for the case with unlimited backhaul capacities, and $\alpha = 1$ takes the following form:*

$$\forall n \in \{1, 2, \cdots\}, \forall \tilde{t} \in \boldsymbol{\tau}_n$$
$$j_n^{ua}(\tilde{t}) = \arg\max_{j \in \{0\} \cup \mathcal{P}} \log(R_{jn}(\omega(\tilde{t}))) + \log\left(\frac{N_j^{N_j}}{(N_j+1)^{N_j+1}}\right) \tag{7.2}$$

where $N_j = |A_j(\omega(\tilde{t}^-)) \setminus \{n\}|$ is the number of users in BS j just before the association event at time $\tilde{t}$, excluding user n.

Proof. First let us restrict ourselves to the association event of user n corresponding to the arrival instant a_n. $R_{jn}(\omega(a_n))$ is the rate that it sees from BS j. If it decides to go to BS j, only the local utility of BS j is affected (due to the decoupling property as detailed in the previous chapter). So, the difference in local α-fair utility $\Delta(f_j)$ can be written as

$$\Delta(f_j) = \sum_{i \in A_j(\omega(a_n))} \log\left(\frac{R_{ji}(\omega(a_n))}{N_j+1}\right) - \sum_{i \in A_j(\omega(a_n^-))} \log\left(\frac{R_{ji}(\omega(a_n^-))}{N_j}\right)$$

where N_j is the number of users in BS j before user n's arrival. Due to the decoupled nature, it is easy to see that the optimal UA is equivalent to $\arg\max_{j \in \{0\}\mathcal{P}} \Delta(f_j)$. The proposition can be derived by simple manipulation of the terms in $\Delta(f_j)$. The proof for a general association event is very similar. □

In this particular case, the optimal UA scheme requires each BS to broadcast the number of users associated to it, $n_j = |A_j(\omega(\tilde{t}^-))|$. A UE can perform the UA decision based on this *load* information[2]. This scheme is scalable, and has a very simple computation (constant complexity in the number of users in the system). This is a very simple rule that gives the optimal trade-off between network load and link rate. This favors BSs that can provide higher rates, and with fewer users. We call this the backhaul-unaware optimal (*BHU-Optimal*) UA scheme.

[2]This is in addition to the required information for user n to make its link measurements (reference signals for SINR and rate-mapping for link-rate computation), and K.

$\underline{\alpha \neq 1}$

Unlike for the special case of $\alpha = 1$, the case of general $\alpha > 0$ does not yield a simple solution for the optimal UA scheme even when the backhaul is not a bottleneck. This is due to the fact that user scheduling variables $\{\beta_{ji}\}_{i\in\mathcal{N}_j}$ at BS j depend on the individual rate components $\{R_{ji}\}_{i\in\mathcal{N}_j}$, as shown in Theorem 1(b). This is unlike the case of $\alpha = 1$, where the association of user n to BS j affects the existing users' throughput in a way that is independent of user n's rate R_{jn} (i.e., the throughput of user i changed from $\frac{R_{ji}}{N_j}$ to $\frac{R_{ji}}{N_j+1}$). This means that the computation of $\Delta(f_j)$ (i.e., the change in local α-fair utility due to association decision) requires the individual rates of the existing users in BS j. This requirement complicates the implementation of the optimal UA scheme for a general value of α. Even though the user-centric implementation of this scheme is not as simple as the case of $\alpha = 1$, we can solve it to obtain the benchmark with which we can compare simpler rules.

7.4.2 The general backhaul-limited scenario

In the most general scenario, where the backhaul links are of limited capacity, the optimal UA scheme does not have a simple form. This is due to the fact that we can no longer decouple the local utilities at each BS, in order to determine the optimal choice. There is the coupling between multiple BSs, which we have explained in the previous chapter.

The general form of the optimal UA scheme can be solved by user n at association events $\tilde{t} \in \boldsymbol{\tau}_n$ if it has the following *global* information:

$$C_{BH}, \{C_j\}_{j\in\mathcal{P}}, \{\{R_{ji}(\omega(\tilde{t}^-))\}_{i\in A_j(\omega(\tilde{t}^-))}\}_{j\in\mathcal{P}\cup\{0\}}, \alpha, K.$$

If this UA scheme were to be implemented, we could use the MBS to broadcast C_{BH}, α and K as part of the system-specific information, and each BS to broadcast its backhaul capacity C_j and the individual user rates $\{R_{ji}(\omega(\tilde{t}^-))\}_{i\in A_j(\omega(\tilde{t}^-))}$. The amount of information that a BS has to broadcast, in this case, increases with the number of users attached to it. Moreover, the computation of the optimal solution requires to solve a set of $X + 1$

global α-fair scheduling problems. Each of these problems is computationally quite complex since it involves an interior-point algorithm to solve a convex optimization problem that is polynomial in the number of users $|\mathcal{N}(\omega(\tilde{t}^-))|$. Thus, the optimal UA scheme is not a good candidate for an online UA algorithm, except for the special case described above. Note however, that we can compute it to use as a benchmark.

Next, we present an approach that attempts to strike a good trade-off between performance and implementation complexity by taking a backhaul-unaware throughput-selfish approach.

7.5 Backhaul-unaware throughput-selfish UA scheme

The complexity of the optimal UA scheme mainly comes from two aspects: *1)*a user needs to know the change in local utilities of the BSs (which except for the case of $\alpha = 1$ requires a lot of information), and *2)*the backhaul-limitations couple the change in user utilities across different BSs. So, if a user ignores backhaul limitations, and concentrates only on maximizing its own throughput, we can obtain a very simple rule which does not require a lot of information. We call it the *BHU-Selfish* scheme. In this scheme, we allow a user to take a *selfish* approach where it associates to the BS that it believes will provide the best throughput, under the assumption that the system had no backhaul limitations. From Theorem 1, we know that if there were no backhaul limitations, and if user n associates to BS j during the association event $\tilde{t}$, it would be scheduled for $\beta_{jn} = \frac{T_{jn,\alpha}(\omega(\tilde{t}))}{\sum_{i\in A_j(\omega(\tilde{t}^-))\setminus\{n\}} T_{ji,\alpha}(\omega(\tilde{t}^-))+T_{jn,\alpha}(\omega(\tilde{t}))}$ proportion of time, and hence would get a throughput of

$$\frac{R_{jn}(\omega(\tilde{t}))^{\frac{1}{\alpha}}}{\sum_{i\in A_j(\omega(\tilde{t}^-))\setminus\{n\}} T_{ji,\alpha}(\omega(\tilde{t}^-)) + T_{jn,\alpha}(\omega(\tilde{t}))}.$$

Recall that $T_{ji,\alpha}(\omega) \triangleq R_{ji}(\omega)^{\frac{1-\alpha}{\alpha}}$. Thus, a Backhaul-Unaware, Throughput-Selfish (BHU-Selfish) UA scheme, can be written as follows.

$$\forall n \in \{1, 2, \cdots\}, \forall \tilde{t} \in \boldsymbol{\tau}_n$$
$$j_n^{ua}(\tilde{t}) = \arg\max_{j\in\mathcal{P}\cup\{0\}} \left\{ \frac{R_{jn}(\omega(\tilde{t}))^{\frac{1}{\alpha}}}{\Theta_j(\tilde{t}^-) + R_{jn}(\omega(\tilde{t}))^{\frac{1-\alpha}{\alpha}}} \right\} \tag{7.3}$$

where $\Theta_j(\tilde{t^-}) = \sum_{i \in A_j(\omega(\tilde{t^-}))\setminus\{n\}} T_{ji,\alpha}(\omega(\tilde{t^-}))$. In that case, each BS $j \in \mathcal{P} \cup \{0\}$ needs to broadcast *one* scalar value $\hat{\Theta}_j = \sum_{i \in A_j(\omega(\tilde{t^-}))} T_{ji,\alpha}(\omega(\tilde{t^-}))$, in addition to the system-specific parameters α and K broadcast by the MBS[3].

This scheme is scalable (with only one information per BS) in addition to two more from the MBS. This scheme is also computationally very simple and has a constant complexity in the number of users in the system. It allows us to apply the same approach to different values of the fairness parameter α. For the case of $\alpha = 1$, though, there is no need to go with this selfish approach because BHU-Optimal UA scheme has the same complexity as the BHU-Selfish UA scheme.

Note that a user makes its UA decision based on an estimate of the throughput it will receive. This estimate, however, is wrong because it does not take the backhaul into account. In Section 7.7, we use simulations to compare how this heuristic works with respect to the optimal benchmark.

7.6 Physical-layer based UA schemes

We have already mentioned that there are a class of UA schemes where a user performs its association based only on some physical layer parameters. We introduced them in Section 4.4. As a benchmark, we will consider small-cell first user association rule, as a representative of such rules. *Small cell first* (SCF) rule, introduced in Section 4.4, can be defined as follows.

$$\forall n \in \{1, 2, \cdots\}, \forall \tilde{t} \in \boldsymbol{\tau}_n$$
$$j_n^{ua}(\tilde{t}) = \begin{cases} \arg\max_{j \in \mathcal{P}} \left\{\gamma_{jn}(\omega(\tilde{t}))\right\} \text{ if } \max_{j \in \mathcal{P}} \gamma_{jn}(\omega(\tilde{t})) > \delta \\ \arg\max_{j \in \mathcal{P} \cup \{0\}} \gamma_{jn}(\omega(\tilde{t})), \text{ otherwise} \end{cases}$$

where δ is the SINR threshold, also called the biasing parameter. Without loss of generality, we assume that the MBS broadcasts this threshold δ. User n can compute the SINRs from

[3]This approach can be easily applied to the case where each BS has a different choice of fairness parameter α, in which case, each BS needs to broadcast its fairness parameter.

the BSs and make the association decision using this threshold. This scheme is very simple from the user's point of view as it involves comparing at most $X + 1$ physical layer measurements. The selection of the right threshold δ, however, is not easy. The burden of selecting a good value of threshold lies in the network. This UA scheme in principle is similar to other schemes which perform *biased* comparison of some physical layer measurements to make association decision, e.g., the cell range-expansion scheme [62].

In this next section, we will present our simulation set-up and the simulation results.

7.7 Simulation

7.7.1 Simulation set-up

We consider a HetNet with a deployment area in the form of a regular hexagonal cell with each side having a length of $500/\sqrt{3}$ m. This is the 3GPP urban setting with an inter-site distance (ISD) of $500m$ [6]. The deployment area is covered by a centrally placed MBS and $X = 4$ SCs which are *symmetrically placed* at a distance of $r = 178$ m. from the center. The MBS has a transmit power budget of $P_M = 43$ dBm and each SC has a transmit power budget of $P_S = 30$ dBm. A total of $M = 33$ subchannels are available to each macro cell, out of which SCs get K subchannels and the MBS gets the remaining $M - K$ subchannels. We take into account the interference coming from the outer macrocells by taking 18 identical macro cells around the given macro cell with a reuse factor of 3. We assume that all SC backhaul links have equal capacity, i.e. $C_j = C$.

We use the distance-based path-loss model recommended by 3GPP [6], as shown in Table 6.2 and assume perfect measurements. The interference can be easily computed in view of our assumptions. The rate function $\theta(.)$ is taken as the 15-rate MCS available in LTE, as shown in Table 4.2.

7.7.2 Key assumptions

Even though our framework allows us to consider a broad types of online user association algorithms, we have taken a number of assumptions for generating the numerical results.

- Each arriving user's location is randomly chosen at its arrival instant with a certain distribution. The user does not move and we assume that the channel condition of a given user does not change with time. This does not mean that the network is static. The network realization changes due to the arrival and departure processes.
- Each user n is associated to a BS upon arrival, and remains with the same BS as long as it is in the system, i.e., $\boldsymbol{\tau}_n = \{a_n\}$.
- The user inter-arrival times are exponentially distributed with mean $\frac{1}{a}$. A user stays in the system for an exponentially distributed amount of time with mean τ after which it leaves the system. This traffic model is different from a traffic model where users come with a given file-size and the amount of time a user stays in the system is dependent on the future evolution of the system. This user behavior is suitable to model users watching (or listening to) media streams where the quality of a stream (i.e., its coding rate) is adjusted to match with the available end-to-end throughput. In this traffic model, higher throughput to a particular user translates to a better quality of experience, but the amount of time a user spends in the system is independent of the allocated throughput. Note that the average number of users in the system is given by the M/M/∞ formula, i.e., $\overline{N} = a\tau$, where $a = 2$ users/minute is the user arrival rate. We take the value of τ such that an average of $\overline{N} = 30$ users are in the HetNet deployment area.

Regarding the spatial distribution of users, we consider two user distributions: uniform distribution (UD) and non-uniform distribution (NUD). In UD, an arriving user's location is chosen uniformly at random in the deployment area. In NUD, we construct small hexagons of radius $r_{hotspot} = ISD/10.0$ centered at each SC, as shown in Fig. 7.1. An incoming user selects one of the $X + 1$ hexagons (X small ones around the SCs, and the big one corresponding to the deployment area centered at the MBS) uniformly at random.

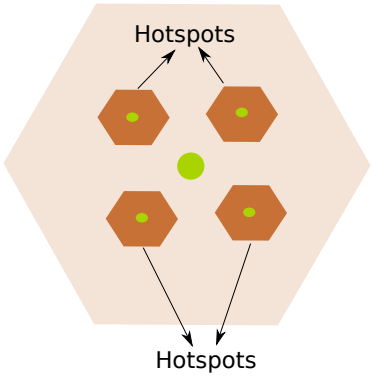


Figure 7.1: Hotspots in the non-uniformly distributed case

Once it selects one of the five hexagons, it chooses a point within the selected hexagon uniformly at random. In other words, we create hot-spots around the SCs.

7.7.3 Performance metric

We take the average value of the *α-mean throughputs (as defined in (6.10)) at arrival instants* as our performance metric, calculated as follows:

$$\frac{1}{L}\sum_{n=1}^{L}\overline{T}_{\alpha}\left(\{\lambda_i\}_{i\in\mathcal{N}(\omega(a_n))}\right) \tag{7.4}$$

where L is the number of user arrivals simulated in the system. Our simulation results are computed with $L = 1000$ user arrivals.

Let $\chi(ua, \alpha, C, C_{BH}, K)$ represent the average α-mean throughput for user association scheme ua and channel split parameter K, for a given values of backhaul capacities (C_{BH}, C). Choosing a good value of the channel split parameter K is very important and is expected to affect the performance of different UA schemes. For example, if $K^*(ua, \alpha, C, C_{BH})$ is the optimal choice of K for a given fairness parameter α, a given SC backhaul capacity C, a given MBS backhaul capacity C_{BH}, and a given UA scheme ua, we expect that $K^*(ua, \alpha, C', C_{BH})$ is less than or equal to $K^*(ua, \alpha, C, C_{BH})$ for $C' \leq C$.

This intuition is based on the fact that if the SC backhaul capacity is small, we would dedicate smaller amount of channel resources to the SCs. In other words, if the system had a process to fine-tune K, it would choose the value of K as a function of the employed UA scheme and the backhaul capacities, among other things. If we want to compare the different UA schemes fairly, we would compare the average α-mean throughputs when the system chooses a good value of K for a given UA scheme and backhaul capacities. Let us define this upper-bound in performance of each UA scheme as follows.

$$\chi^*(ua, \alpha, C, C_{BH}) = \max_{k \in \{0,1,\cdots,M\}} \chi(ua, \alpha, C, C_{BH}, K)$$

Next, we will use this metric to compare different UA schemes.

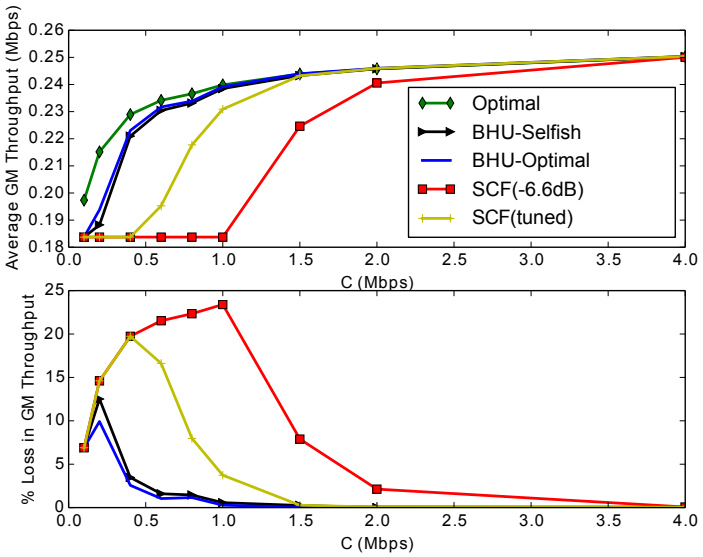
7.7.4 Results for fine-tuned K

Fig. 7.2 shows the average α-mean throughput χ^* (with K chosen optimally) for $\alpha = 1$ for different UA schemes, for the case of uniformly distributed (UD) users for $L = 1000$ user arrivals. For SCF, we fine-tune the threshold parameter δ by choosing the best threshold among the set of threshold SNRs given in Table 4.2 for a given set of backhaul capacities. We call this fine-tuned version of SCF as *SCF(tuned)*. For a sub-optimal scheme ua, the percentage loss in performance with respect to the optimal scheme is calculated as follows.

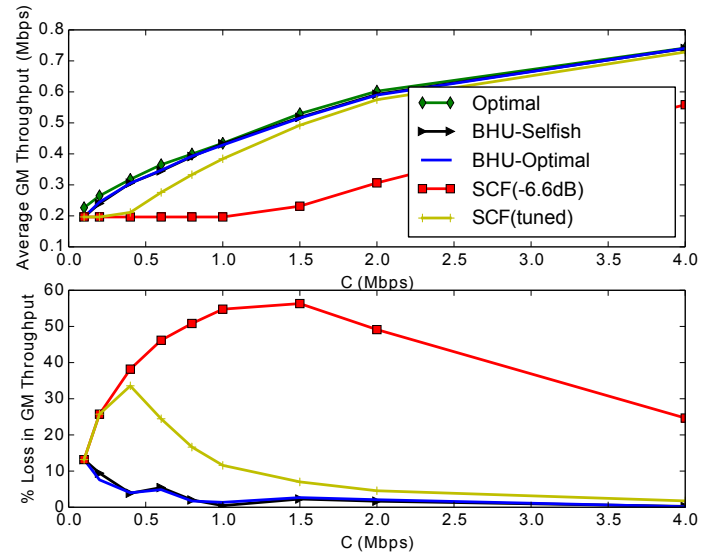
$$100 \times \frac{\chi^*(\text{OPTIMAL}, \alpha, C, C_{BH}) - \chi^*(ua, \alpha, C, C_{BH})}{\chi^*(\text{OPTIMAL}, \alpha, C, C_{BH})}$$

As we see from the figure, the loss in performance due to a sub-optimal UA scheme decreases with an increase in C. In this case (i.e., $\alpha = 1$), both BHU-Optimal and BHU-Selfish are equally complex, as discussed before. From performance point of view, the figure shows that the performance of both of these backhaul-unaware schemes are similar.

Unless the backhaul capacity is very small, the performance loss due to the backhaul-unaware schemes (i.e., BHU-Optimal or BHU-Selfish) is small. The backhaul-unaware schemes outperform SCF(tuned) when the backhaul capacity C is very small. Even though the SCF(tuned) scheme works well for $C > 2Mbps$, SCF requires the system to fine-tune its threshold, whereas the backhaul-unaware UA schemes have no parameter that needs to



(a) $C_{BH} = 6.0Mbps$



(b) $C_{BH} = 20.0Mbps$

Figure 7.2: Performance as a function of SC backhaul capacity, $\alpha = 1$, $\overline{N} = 30$, $K = K^*(ua, \alpha, C, C_{BH})$

Table 7.1: Comparison of optimal, BHU-selfish and SCF UA schemes: $\alpha = 1$, NUD.

NUD	$(C_{BH}, C) = (20.0\ 2.0)$		$(C_{BH}, C) = (20.0, 20.0)$		$(C_{BH}, C) = (6.0, 2.0)$		$(C_{BH}, C) = (6.0, 6.0)$	
UA Scheme	**α-mean**	**Loss**	**α-mean**	**Loss**	**α-mean**	**Loss**	**α-mean**	**Loss**
Optimal	0.30	-	0.74	-	0.22	-	0.22	-
BHU Selfish	0.30	0.00%	0.74	0.05%	0.22	0.22%	0.22	0.00%
SCF(tuned)	0.16	48.95%	0.74	0.53%	0.11	50.56%	0.22	0.00%

be optimized by the system. If SCF is not fine-tuned, the performance can be very bad, as depicted by the poor performance of SCF($\delta = -6.6dB$).

Table 7.1 shows the results for the case when users have a non-uniform distribution (NUD). For backhaul capacity $C \geq 2.0$, a backhaul-unaware scheme is practically as good as the optimal scheme. SCF(tuned), on the other hand, does not perform well when the SC backhaul link is severely limited ($C = 2.0$ Mbps). Similar results were obtained for $\alpha = 2$. We have listed these results in Table 7.2. The results presented so far allow us to make the following observations.

Observations 1) If the system chooses the *backhaul-aware* optimal value of K, and performs a backhaul-aware optimal α-fair scheduling, a backhaul-unaware user association scheme performs very well, both for the uniform and non-uniform deployment of users, and for different choices of fairness parameter α.
2) SCF scheme performs well only if the backhaul is not a bottleneck, and if the biasing parameter is well-chosen. This is consistent with the results shown in [37]. Tuning the biasing parameter can be very difficult. On the other hand, either of the two backhaul-unaware schemes for $\alpha = 1$ and the BHU-Selfish scheme for $\alpha \neq 1$ do not have any UA-specific parameter to be tuned. This is an additional benefit of the backhaul-unaware schemes over the parameterized physical-layer based UA scheme.

7.7.5 Impact of K

All our results are based on the assumption that the network optimizes the value of K, for a given UA scheme and for a given set of backhaul capacities. This could be achieved by

Table 7.2: Comparison of optimal, BHU-selfish and SCF UA schemes: $\alpha = 2$.

UD	$(C_{BH}, C) = (20.0\ 2.0)$		$(C_{BH}, C) = (20.0, 20.0)$		$(C_{BH}, C) = (6.0, 2.0)$		$(C_{BH}, C) = (6.0, 6.0)$	
UA Scheme	**α-mean**	**Loss**	**α-mean**	**Loss**	**α-mean**	**Loss**	**α-mean**	**Loss**
Optimal	0.46	-	0.63	-	0.22	-	0.22	-
BHU Selfish	0.45	3.79%	0.63	0.00%	0.22	0.00%	0.22	0.00%
SCF(tuned)	0.44	6.26%	0.60	6.03%	0.22	1.32%	0.22	2.18%
NUD								
Optimal	0.28	-	0.74	-	0.22	-	0.22	-
BHU Selfish	0.28	1.20%	0.74	0.00%	0.22	0.00%	0.22	0.00%
SCF(tuned)	0.14	51.34%	0.71	3.56%	0.11	52.07%	0.22	2.18%

finding an appropriate function $K^*(ua, \alpha, C, C_{BH})$ for a given distribution of users. Since the system chooses K to reflect the backhaul limitation (for example, the chosen K is small when the SC backhaul link capacities are small, and is large when the SC backhaul link capacities are large), our selfish scheme gets an *indirect signal* about backhaul limitation (as small K would generally translate to reduced throughput from the small cell). This is a reason why a user-centric backhaul unaware approach to user association would work on a system that properly configures its resource allocation parameter. If K was not chosen properly, the performance penalty due to a backhaul-unaware scheme as well as the SCF scheme can be higher, as depicted in Table 7.3. This loss is more pronounced in the NUD case.

A good configuration of K could be achieved by a self-organization (SON) algorithm, or by an offline computation. The difficulty of choosing a good value of K depends on the sensitivity of the network's performance as a function of K. In order to understand this, in Fig. 7.3, we plot the range of values of K that can yield a performance within 95% of the optimal choice, for each UA scheme and a given choice of backhaul capacities, when the users are distributed uniformly. As we can see, there is a range of values of K that work well. So, we do not have a very stringent requirement in optimizing the channel allocation parameter. We could employ any scheme that can tune the channel allocation parameter to any value in the quasi-optimal range shown in the figure.

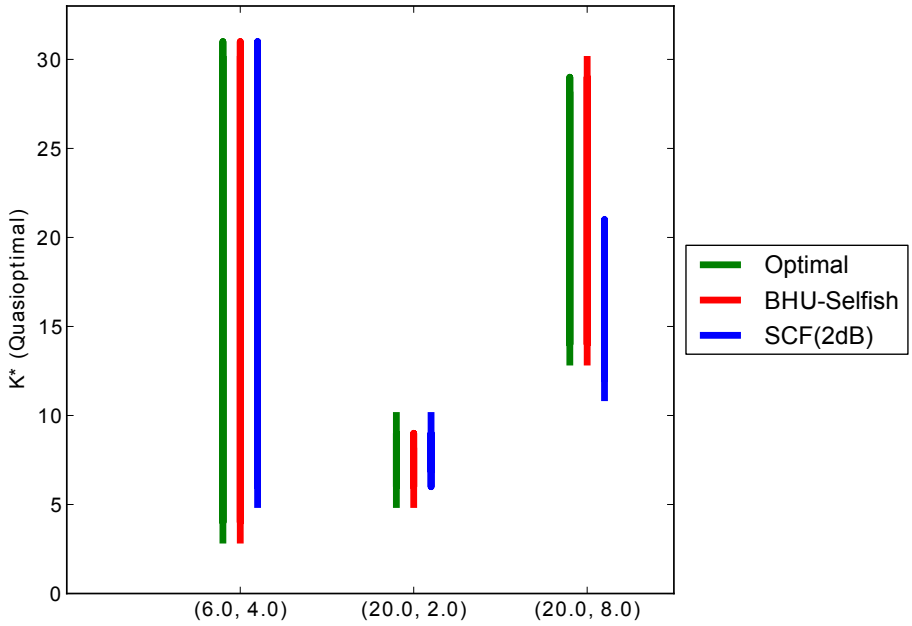


Figure 7.3: Quasi-optimal values of K for different backhaul capacities.

Table 7.3: Loss in performance for an arbitrarily-chosen value of K, $(C_{BH}, C) = (20.0, 2.0)$.

$\alpha = \mathbf{1}$	UD			NUD		
	$K = 5$	$K = 15$	$K = 25$	$K = 5$	$K = 15$	$K = 25$
BHU Selfish	1.45%	5.37%	2.35%	3.30%	20.42%	18.69%
SCF(tuned)	6.92%	7.28%	16.30%	49.55%	40.22%	31.67%
$\alpha = \mathbf{2}$	UD			NUD		
	$K = 5$	$K = 15$	$K = 25$	$K = 5$	$K = 15$	$K = 25$
BHU Selfish	0.00%	6.45%	3.68%	10.04%	29.70%	22.91%
SCF(tuned)	17.65%	8.38%	22.96%	54.51%	42.05%	33.46%

7.8 Conclusion

We took a user-centric approach to user association, and discussed important design considerations for online UA schemes. Under a global α-fair throughput allocation framework, we introduced optimal user association scheme that is carried out by a user whenever association events are triggered. We showed that, for some cases, the optimal UA scheme can be simple to implement. However, in general, it requires a large amount of information from the BSs. As a solution, we presented a *backhaul-unaware throughput-selfish* UA scheme, which is simple to implement, and can be used for all choices of the fairness parameter. By using numerical results, we showed that such a throughput-selfish scheme can yield performance close to the optimal, provided that the network fine-tunes the resource allocation parameter. We also showed that the parameterized physical-layer based UA scheme can be difficult to tune, and can often lead to poor performance.

Chapter 8

Conclusion

8.1 Summary

In this thesis, we studied the downlink of an OFDM-based heterogeneous cellular network from a throughput perspective. We studied four different radio resource management related processes, namely resource allocation, transmission coordination, user scheduling, and user association. The first part of our thesis (Chapters 3 - 5) focused on exploring the interplay between these processes under different scenarios of HetNet deployment. The main research challenge was to come up with a unified framework that allows us to characterize the throughput-performance of different combinations of these processes and deployment scenarios (also called the deployment *choices*). By adopting a flow-based approach popular in the literature of wireless mesh networks, we were able to formulate a joint optimization problem that can yield optimal throughput performance for different deployment choices. The need for a flow-based modeling arises due to two-hop wireless set-up under relay deployment scenarios, and due to on-off transmission coordination mechanism. Via numerical results, we obtained a number of interesting engineering insights:

Wired small cell deployment :

- Allowing a user to associate to more than one BS will not offer significant performance gains.

- Partially-shared deployment/Orthogonal deployment perform very well even in the absence of sophisticated transmission coordination whereas transmission coordination is essential for the satisfactory performance of co-channel deployment.

Relay deployment :

- Some configurations of user-band relay deployment yield very little or even negative gains whereas some others can yield performances very close to the upper bounds (corresponding to the wired deployment with sufficiently large backhaul capacities). This highlights the importance of deploying the right configurations.
- Using a dedicated band for backhauling is a promising solution for small cells, in particular in the case of the mmWave band, since a small bandwidth is sufficient to satisfy the demand of a typical small cell backhaul link.

Even though a joint optimization framework enabled us to study diverse choices of HetNet deployments, it has some limitations. The most important of these limitations perhaps is that it is based on a "snapshot" approach, where the system variables (like resource allocation parameter, user schedules, and user association variables) are solved for a given snapshot of the network and many independent snapshots are studied to get some average performance. In other words, the dynamics of a network is approximated by a set of independent snapshots, which might not be an accurate modeling approach. Moreover, with the snapshot approach, we optimize the system variables *jointly and simultaneously*, which can give us upper-bound in performance but is often limited to the offline study. As far as finding good online radio resource management algorithms is concerned, we would need a different approach. This is the motivation for the second part of the thesis.

In the second part of our thesis (Chapters 6 - 7), rather than jointly optimizing different network processes together, we studied problems where we optimize one process at a time. The first problem we studied is the global α-fair user scheduling problem under backhaul limitations. We characterized optimal user scheduling under different scenarios of backhaul limitations. This problem is still based on a snapshot. However, the analytical insights

obtained by solving the problem allowed us to come up with heuristics that can be implemented as very simple online α-fair user scheduling algorithms. In Chapter 7, we studied α-fair user association under backhaul limitations. We introduced user arrival/departure processes into our model and presented performance results of the optimal and sub-optimal α-fair user association schemes. For backhaul unlimited case and with $\alpha = 1$, we showed how a very simple rule can be used to achieve optimal user association. In the general case, the optimal algorithm can be very complex, but we showed that if other network processes are optimized, a very simple user association scheme can be employed, without a huge penalty in performance.

8.2 Future Research Directions

In our study, we presented performance of ON-OFF transmission coordination mechanism, which can be seen as a special case of a more general coordination mechanism that involves power control at the BSs. There has been a growing interest in understanding the feasibility of such an approach. Incorporating the general coordination mechanism into our model without losing tractability can be an interesting related problem.

Interference cancellation is often seen as an alternative to interference coordination. Incorporating interference cancellation to our model could be another interesting extension. This will allow us to answer questions like "*can interference cancellation avoid the need for transmission coordination?*"

We have restricted ourselves to the downlink. However, uplink is equally important. Optimization of the network processes in the presence of both uplink and downlink flows is important, and can be a future research direction.

As far as the backhaul limitations is concerned, we only take the rate limitations. Delay is another equally important limitation. Incorporating both the rate and the delay limitations of the backhaul infrastructure, and characterizing the optimal α-fair user scheduling could be a very interesting research problem.

Appendix A

Proofs

A.1 Proof of Theorem 2

The following important property of $[\mathbf{P^j_{Local}}(\omega,C_j)]$ will be useful in the ensuing analysis.

Proposition 2. *If $C_j > 0$, there exists a unique optimal solution to $[\mathbf{P^j_{Local}}(\omega,C_j)]$ with $\beta_{ji} > 0$ for all $i \in A_j(\omega)$.*

The proof is similar to Proposition 1 in [39].

The Lagrangian function of the local problem can be defined as follows.

$$L(\boldsymbol{\beta}_j;\mu_j,\zeta_j,\boldsymbol{l}_j) = -\sum_{i\in A_j(\omega)} U_\alpha(\beta_{ji}R_{ji}(\omega)) +$$
$$\mu_{j,\omega}(\sum_{i\in A_j(\omega)} R_{ji}(\omega)\beta_{ji} - C_j) + \zeta_{j,\omega}(\sum_{i\in A_j(\omega)} \beta_{ji} - 1) - \sum_{i\in A_j(\omega)} l_{j,i,\omega}\beta_{ji}$$

where $\boldsymbol{\beta}_j$ and $\boldsymbol{l}_j$ are respectively the vectors comprising of all β_{ji} and all $l_{j,i,\omega}$ for $i \in A_j(\omega)$. The Karush-Kuhn-Tucker (KKT) conditions [20], necessary for optimality of $[\mathbf{P^j_{Local}}(\omega,C_j)]$,

can be written as follows.

$$\frac{\partial L}{\partial \beta_{ji}} = 0 \implies \beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{(\mu_{j,\omega} R_{ji}(\omega) + \zeta_{j,\omega} - l_{j,i,\omega})^{\frac{1}{\alpha}}} \quad \forall i \in A_j(\omega) \tag{A.1}$$

$$\zeta_{j,\omega}(\sum_{i \in A_j(\omega)} \beta_{ji} - 1) = 0; \quad \mu_{j,\omega}(\sum_{i \in A_j(\omega)} R_{ji}(\omega)\beta_{ji} - C_j) = 0 \tag{A.2}$$

$$l_{j,i,\omega}\beta_{ji} = 0, \quad \forall i \in A_j(\omega) \tag{A.3}$$

$$\mu_{j,\omega} \geq 0; \quad \zeta_{j,\omega} \geq 0; \quad \boldsymbol{l}_j \geq 0; \quad Eq.(6.13); \quad Eq.(6.14); \quad Eq.(6.15);$$

(A.1) are the first-order necessary conditions for optimality. (A.2) and (A.3) are the so-called complementary-slackness conditions. The primal problem involves maximization of a concave function over a convex set, and hence any tuple of primal and dual variables $(\{\beta_{ji}\}, \mu_{j,\omega}, \zeta_{j,\omega}, \{l_{j,i,\omega}\})$ that satisfies all of the KKT conditions is optimal [20]. Also, from Proposition 2, we know that such a solution is unique. Moreover, since the optimal solution is known to satisfy $\beta_{ji} > 0$, we have $l_{j,i,\omega} = 0$ for all $i \in A_j(\omega)$ from (A.3). Using this fact on the first order condition (A.1), we get

$$\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{(\mu_{j,\omega} R_{ji}(\omega) + \zeta_{j,\omega})^{\frac{1}{\alpha}}}, \quad \forall i \in A_j(\omega) \tag{A.4}$$

Note that the optimal dual variables obey one of the *three* conditions: $(\mu_{j,\omega} = 0, \zeta_{j,\omega} > 0)$, $(\mu_{j,\omega} > 0, \zeta_{j,\omega} = 0)$, and $(\mu_{j,\omega} > 0, \zeta_{j,\omega} > 0)$. This is because, (A.4) imposes $\mu_{j,\omega} R_{ji}(\omega) + \zeta_{j,\omega} \neq 0$, for $\alpha > 0$. Hence, $(\mu_{j,\omega} = 0, \zeta_{j,\omega} = 0)$ is not possible.

We will make use of the following lemmas to establish our main result.

Lemma 2. *(a) If $C_j \geq C^*_{j,\alpha}(\omega)$, then $(\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i \in A_j(\omega)} T_{ji,\alpha}(\omega)}, \forall i \in A_j(\omega))$ is the unique optimal solution to $[\mathbf{P^j_{Local}}(\omega, C_j)]$. (b) If $C_j < C^*_{j,\alpha}(\omega)$, then $(\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i \in A_j(\omega)} T_{ji,\alpha}(\omega)}, \forall i \in A_j(\omega))$ is not feasible.*

Proof. It is easy to verify that $\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i \in A_j(\omega)} T_{ji,\alpha}(\omega)}$ for all $i \in A_j(\omega)$, $\mu_{j,\omega} = 0$ and $\zeta_{j,\omega} = \left(\sum_{i \in A_j(\omega)} T_{ji,\alpha}(\omega)\right)^{\alpha}$ satisfy all KKT conditions if $C_j \geq C^*_{j,\alpha}(\omega)$. It is thus an optimal solution consistent with the backhaul capacity value $C_j \geq C^*_{j,\alpha}(\omega)$. Proposition 2 implies that this is in fact the only optimal solution. If $C_j < C^*_{j,\alpha}(\omega)$, substituting $\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i \in A_j(\omega)} T_{ji,\alpha}(\omega)}$ for all $i \in A_j(\omega)$ in $\sum_{i \in A_j(\omega)} \beta_{ji} R_{ji}(\omega) \leq C_j$ results in a contradiction. □

Lemma 3. *(a) If $C_j \leq c_j^*(\omega)$, then $(\beta_{ji} = \frac{C_j}{|A_j(\omega)|R_{ji}(\omega)}, \forall i \in A_j(\omega))$ is the unique optimal solution to $[\mathbf{P^j_{Local}}(\omega, C_j)]$. (b) If $C_j > c_j^*(\omega)$, then $(\beta_{ji} = \frac{C_j}{|A_j(\omega)|R_{ji}(\omega)}, \forall i \in A_j(\omega))$ is not feasible.*

Proof. We can easily verify that $\beta_{ji} = \frac{C_j}{|A_j(\omega)|R_{ji}(\omega)}$ for all $i \in A_j(\omega)$, $\mu_{j,\omega} = \left(\frac{|A_j(\omega)|}{C_j}\right)^{\alpha}$ and $\zeta_{j,\omega} = 0$ satisfy all KKT conditions if $C_j \leq c_j^*(\omega)$. It is thus an optimal solution consistent with the backhaul capacity value $C_j \leq c_j^*(\omega)$. Proposition 2 implies that this is also the only optimal solution. If $C_j > c_j^*(\omega)$, then substituting $\beta_{ji} = \frac{C_j}{|A_j(\omega)|R_{ji}(\omega)}$ for all $i \in A_j(\omega)$ in $\sum_{i \in A_j(\omega)} \beta_{ji} \leq 1$ results in a contradiction. □

Lemma 4. *If $c_j^*(\omega) < C_j < C_{j,\alpha}^*(\omega)$, the optimal dual solution is obtained by solving the following equations for $\mu_{j,\omega} > 0$ and $\zeta_{j,\omega} > 0$.*

$$\sum_{i \in A_j(\omega)} \frac{R_{ji}(\omega)^{\frac{1}{\alpha}}}{(\mu_{j,\omega} R_{ji}(\omega) + \zeta_{j,\omega})^{\frac{1}{\alpha}}} = C_j \tag{A.5}$$

$$\sum_{i \in A_j(\omega)} \frac{T_{ji,\alpha}(\omega)}{(\mu_{j,\omega} R_{ji}(\omega) + \zeta_{j,\omega})^{\frac{1}{\alpha}}} = 1 \tag{A.6}$$

Proof. We will first show that the optimal dual variables have to satisfy $\mu_{j,\omega} > 0$ and $\zeta_{j,\omega} > 0$. First, we assume that there exists a dual optimal solution such that $\mu_{j,\omega} = 0$. $\mu_{j,\omega} = 0$ implies $\zeta_{j,\omega} > 0$, and hence

$$\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\zeta_{j,\omega}^{\frac{1}{\alpha}}} \text{ and } \sum_{i \in A_j(\omega)} \beta_{ji} = 1$$

$$\implies \beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{\sum_{i \in A_j(\omega)} T_{ji,\alpha}(\omega)}, \forall i \in A_j(\omega)$$

We know from Lemma 2(b) that this is an infeasible solution since $C_j < C_{j,\alpha}^*(\omega)$. Thus, we require $\mu_{j,\omega} > 0$.

Similarly, we assume that there exists a dual optimal solution such that $\zeta_{j,\omega} = 0$.

$\zeta_{j,\omega} = 0$ implies $\mu_{j,\omega} > 0$, and hence

$$\beta_{ji} = \frac{T_{ji,\alpha}(\omega)}{(\mu_{j,\omega} R_{ji}(\omega))^{\frac{1}{\alpha}}} \text{ and } \sum_{i \in A_j(\omega)} \beta_{ji} R_{ji}(\omega) = C_j$$

$$\implies \beta_{ji} = \frac{C_j}{|A_j(\omega)| R_{ji}(\omega)}, \forall i \in A_j(\omega)$$

We know from Lemma 3(b) that this is an infeasible solution since $C_j > c_j^*(\omega)$. Thus, we require $\zeta_{j,\omega} > 0$.

Thus, the optimal solution has to satisfy $\mu_{j,\omega} > 0$ and $\zeta_{j,\omega} > 0$. In such case, (A.2) mandates that the primal constraints (6.13) and (6.14) are satisfied with equality, i.e.,

$$\sum_{i \in A_j(\omega)} R_{ji}(\omega) \beta_{ji} = C_j \text{ and } \sum_{i \in A_j(\omega)} \beta_{ji} = 1 \tag{A.7}$$

Substituting the value of β_{ji} from (A.4) in these equalities, we get the required equations. A strictly positive solution of $(\mu_{j,\omega}, \zeta_{j,\omega})$ should exist due to Proposition 2. □

Proofs for Lemma 2, 3, and 4 complete the proof for Theorem 2.

A.2 Proof of Theorem 1

Note that the optimal schedules for Scenario 0 have to be equal to the solutions for sufficently large values of C_j. So, the proof of Lemma 2 contains the proof for Theorem 1.

A.3 Proof of Lemma 1

The results for $c_j^*(\omega) \leq C_j$ and $C_j \geq C_{j,\alpha}^*(\omega)$ are immediate from the closed-form solutions of $f_{j,\omega}(C_j)$ from Theorem 2.

For $c_j^*(\omega) < C_j < C_{j,\alpha}^*(\omega)$, we know that an optimal dual variable $\mu_{j,\omega}^*(C_j)$ is a subgradient of $f_{j,\omega}(C_j)$ at C_j. We need to show that this is unique and is the only subgradient, or alternatively we need to show that $f_{j,\omega}(C_j)$ is differentiable.

The differentiability of $f_{j,\omega}(C_j)$ can be shown by noting that the local problem has a unique optimal dual solution $\mu^*_{j,\omega}(C_j)$ for $c^*_j(\omega) < C_j < C^*_{j,\alpha}(\omega)$. Applying this uniqueness in Corollary 5(ii) of [70] proves differentiability.

A.4 Proof of Theorem 3

We first establish the following proposition which allows us to compute the primal variables $\{\tilde{C}_j\}_{j\in\{0\}\cup\mathcal{P}}$ that maximize the lagrangian function for a given dual variable Λ.

Proposition 3. $\tilde{C}^D_{j,\omega,C_j}(\Lambda) = \min\{f'^{(-1)}_{j,\omega}(\Lambda), C_j\}, \quad \forall j \in \{0\}\cup\mathcal{P}$ *give the values of virtual capacities* $\{\tilde{C}_j\}$ *that maximize the Lagrangian function* $L(\tilde{\boldsymbol{C}};\Lambda)$ *for a given* Λ *where* $f'^{(-1)}_{j,\omega}(\Lambda)$ *is defined in (6.23), with* $\mu^{*(-1)}_{j,\omega}(\Lambda)$ *representing the inverse mapping of* $\mu^*_{j,\omega}(C_j)$ *in the interval of* $(0, \left(\frac{|A_j(\omega)|}{c^*_j(\omega)}\right)^\alpha)$.

Proof. *Case 1: $C_j \geq C^*_{j,\alpha}(\omega)$* We first prove the proposition for the case of large $\{C_j\}$ (specifically, $C_j \geq C^*_{j,\alpha}(\omega)$ for all j). In this case, $\tilde{f}_{j,\omega,C_j}(\tilde{C}_j) = f_{j,\omega}(\tilde{C}_j)$. The Karush-Kuhn-Tucker (KKT) first-order conditions ($\frac{\partial L}{\partial \tilde{C}_j} = 0$) give us the following.

$$f'_{j,\omega}(\tilde{C}_j) = \Lambda \quad \forall j \in \{0\}\cup\mathcal{P} \tag{A.8}$$

Thus, for all $\Lambda > 0$, we require that a primal variable $\tilde{C}_j$ has to be less than or equal to $C^*_{j,\alpha}(\omega)$ (or, otherwise $f'_{j,\omega}(\tilde{C}_j)$ would be 0, which means $\Lambda = 0$). Together with this, the strictly decreasing nature of $f'_{j,\omega}(\tilde{C}_j)$ for $0 < \tilde{C}_j \leq C^*_{j,\alpha}(\omega)$ allows us to compute an inverse function of $f'_{j,\omega}(C_j)$, defined as $f'^{(-1)}_{j,\omega}(\Lambda)$, for all $\Lambda > 0$ and that, by definition, it should satisfy (A.8). Finding the exact description of this inverse function is not difficult, as outlined below.

The inverse function of $f'_{j,\omega}(C_j)$ with an image in $(0, c^*_j(\omega)]$ has a domain of $\Lambda \in [\left(\frac{|A_j(\omega)|}{c^*_j(\omega)}\right)^\alpha, \infty)$, whose expression, shown in (6.23), is immediate from (6.19). This inverse function with an image in $(c^*_j(\omega), C^*_{j,\alpha}(\omega)]$ has a domain of $\Lambda \in (0, \left(\frac{|A_j(\omega)|}{c^*_j(\omega)}\right)^\alpha)$, and is given by the inverse of dual variable $\mu^*_{j,\omega}(\tilde{C}_j)$, since $\frac{\partial f_{j,\omega}(\tilde{C}_j)}{\partial \tilde{C}_j} = \mu^*_{j,\omega}(\tilde{C}_j)$.

For $\Lambda = 0$, $f'_{j,\omega}(\tilde{C}_j) = \Lambda$ does not have a unique solution as $f'_{j,\omega}(\tilde{C}_j) = 0$ is true for all $\tilde{C}_j \geq C^*_{j,\alpha}(\omega)$. Choosing $\tilde{C}_j = C^*_{j,\alpha}(\omega)$ as the unique map of the inverse function for $\Lambda = 0$ thus does not affect optimality.

Case 2: $C_j < C^*_{j,\alpha}(\omega)$ For $C_j < C^*_{j,\alpha}(\omega)$, the additional requirement of the inverse mapping is that the value of primal variables as a function of Λ have to be feasible. A bounded version of the inverse mapping, with an upper-bound of C_j would satisfy the primal feasibility constraints, which is exactly what $\tilde{C}^D_{j,\omega,C_j}(\Lambda)$ guarantees. □

Since $\tilde{C}^D_{j,\omega,C_j}(\Lambda)$ is a non-increasing function of Λ in $[0, \infty)$, and since $\tilde{f}_{j,\omega,C_j}(\tilde{C}_j)$ is non-decreasing in $\tilde{C}_j$, $\sum_{j\in\{0\}\cup\mathcal{P}} \tilde{f}_{j,\omega,C_j}(\tilde{C}_j)$ can be solved my taking the smallest value of Λ so that the *MBS backhaul constraint* is satisfied. This is exactly what Theorem 3 states.

Bibliography

[1] 3GPP TR 36.814 v9.0.0, evolved universal terrestrial radio access (e-utra); further advancements for e-utra physical layer aspects (release 9).

[2] 3GPP TS 25.304 v10.3.0, 3rd generation partnership project; technical specification group radio access network; user equipment (ue) procedures in idle mode and procedures for cell reselection in connected mode (release 10).

[3] Cisco virtual networking index: Global mobile data traffic forecast update, 2011-2016. [Online], Website. `http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html`.

[4] Minos 5.5. [Online], Website. `http://www.sbsi-sol-optimize.com/asp/solproductminos.htm`.

[5] LTE Advanced: Heterogeneous networks. *Qualcomm Incorporated, White paper*, Feb. 2010.

[6] 3GPP-TSG-RAN-WG1. Evolved universal terrestrial radio access (EUTRA). *3GPP, Tech. Rep. TR 36.814*, 2010.

[7] R. Agrawal, A. Bedekar, RJ La, and V. Subramanian. Class and channel condition based weighted proportional fair scheduler. *Teletraffic Engineering in the Internet Era, Proc. ITC*, 17:553–565, 2001.

[8] Woo-Geun Ahn and Hyung-Myung Kim. Proportional fair scheduling in relay enhanced cellular OFDMA systems. In *Personal, Indoor and Mobile Radio Commu-*

nications, 2008. PIMRC 2008. IEEE 19th International Symposium on, pages 1 –4, Sept. 2008.

[9] I.F. Akyildiz, W.Y. Lee, M.C. Vuran, and S. Mohanty. Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey. *Computer Networks*, 50(13):2127–2159, 2006.

[10] I.F. Akyildiz, W.Y. Lee, M.C. Vuran, and S. Mohanty. A survey on spectrum management in cognitive radio networks. *Communications Magazine, IEEE*, 46(4):40–48, 2008.

[11] Alcatel-Lucent. Mobile backhaul. [Online], Website, 2014. `http://www.alcatel-lucent.com/solutions/mobile-backhaul`.

[12] David Amzallag and Danny Raz. Resource allocation algorithms for the next generation cellular networks. In *Algorithms for Next Generation Networks*, pages 99–129. Springer London, 2010.

[13] J.G. Andrews. Interference cancellation for cellular systems: a contemporary overview. *Wireless Communications, IEEE*, 12(2):19 – 29, april 2005.

[14] A. Bedekar and R. Agrawal. Optimal muting and load balancing for eicic. In *Modeling Optimization in Mobile, Ad Hoc Wireless Networks (WiOpt), 2013 11th International Symposium on*, pages 280–287, May 2013.

[15] T. Beniero, S. Redana, J. Hamalainen, and B. Raaf. Effect of relaying on coverage in 3GPP LTE-advanced. In *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*, pages 1–5. IEEE, 2009.

[16] F. Berggren and R. Jannit. Asymptotically fair scheduling on fading channels. In *Vehicular Technology Conference, 2002. Proceedings. VTC 2002-Fall. 2002 IEEE 56th*, volume 4, pages 1934–1938. IEEE, 2002.

[17] BLiNQ. Small cell base station backhaul. [Online], Website, 2014. `http://www.blinqnetworks.com/solutions/small_cell_base_station_backhaul`.

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

[19] S. Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 1, pages 321–331. IEEE, 2003.

[20] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge Univ Pr, 2004.

[21] T. Bu, L. Li, and R. Ramjee. Generalized proportional fair scheduling in third generation wireless data networks. In *Proc. IEEE INFOCOM 2006*, pages 1 –12, april 2006.

[22] D. Bultmann, T. Andre, and R. Schoenen. Analysis of 3GPP LTE-Advanced cell spectral efficiency. In *Personal Indoor and Mobile Radio Communications (PIMRC), 2010 IEEE 21st International Symposium on*, pages 1876–1881, 2010.

[23] V. Chandrasekhar and J. Andrews. Spectrum allocation in tiered cellular networks. *Communications, IEEE Transactions on*, 57(10):3059 –3068, october 2009.

[24] V. Chandrasekhar, J. Andrews, and A. Gatherer. Femtocell networks: a survey. *Communications Magazine, IEEE*, 46(9):59 –67, september 2008.

[25] Chung Shue Chen, F. Baccelli, and L. Roullet. Joint optimization of radio resources in small and macro cell networks. In *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, pages 1 –5, may 2011.

[26] China Mobile. C-RAN: the road towards green RAN. [Online], Website, Oct 2011. `http://labs.chinamobile.com/cran/wp-content/uploads/CRAN_white_paper_v2_5_EN.pdf`.

[27] G. Cili, H. Yanikomeroglu, and F.R. Yu. Energy efficiency and capacity evaluation of lte-advanced downlink comp schemes subject to channel estimation errors and system

delay. In *Vehicular Technology Conference (VTC Fall), 2013 IEEE 78th*, pages 1–5, Sept 2013.

[28] H. Claussen. Performance of macro- and co-channel femtocells in a hierarchical cell structure. In *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on*, pages 1 –5, sept. 2007.

[29] Mikael Coldery, Ulrika Engstrom, Ke Wange Helmersson, Mona Hashemi, Lars Manholm, and Pontus Wallentin. Wireless backhaul in future heterogeneous networks. *Ericsson Review*, 91, Nov 2014. `http://www.ericsson.com/ae/res/thecompany/docs/publications/ericsson_review/2014/er-wireless-backhaul-hn.pdf`.

[30] M. Coupechoux, J.M. Kelif, and P. Godlewski. Network controlled joint radio resource management for heterogeneous networks. In *Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE*, pages 1771–1775. IEEE, 2008.

[31] A. Damnjanovic et al. A survey on 3GPP heterogeneous networks. *IEEE Wireless Commun. Mag.*, 18(3):10 –21, June 2011.

[32] Suman Das, Harish Viswanathan, and G. Rittenhouse. Dynamic load balancing through coordinated scheduling in packet data systems. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 1, pages 786 – 796 vol.1, march-3 april 2003.

[33] Thiago Martins de Moraes, Muhammad Danish Nisar, Arturo Antonio Gonzalez, and Eiko Seidel. Resource allocation in relay enhanced LTE-Advanced networks. *EURASIP Journal on Wireless Communications and Networking*, 2012(1):1–12, 2012.

[34] Ericsson. Ericsson mobility report. [Online], Website, June 2014. `http://www.ericsson.com/res/docs/2014/ericsson-mobility-report-june-2014.pdf`.

[35] Ericsson. It all comes back to backhaul. [Online], Website, August 2014. `http://www.ericsson.com/res/docs/whitepapers/WP-Heterogeneous-Networks-Backhaul.pdf`.

[36] D. Fooladivanda, A. Al Daoud, and C. Rosenberg. Joint channel allocation and user association for heterogeneous wireless cellular networks. In *Proc. IEEE PIMRC*, Sept. 2011.

[37] D. Fooladivanda and C. Rosenberg. Joint resource allocation and user association for heterogeneous wireless cellular networks. *IEEE Trans. Wireless Commun.*, 12(1):248–257, 2013.

[38] H. Galeana-Zapien and R. Ferrus. Design and evaluation of a backhaul-aware base station assignment algorithm for ofdma-based cellular networks. *Wireless Communications, IEEE Transactions on*, 9(10):3226–3237, 2010.

[39] J. Ghimire and C. Rosenberg. Impact of limited backhaul capacity on user scheduling in heterogeneous networks. In *Wireless Communications and Networking Conference (WCNC), 2014 IEEE*, pages 2480–2485, April 2014.

[40] Jagadish Ghimire and Catherine Rosenberg. Revisiting scheduling in heterogeneous networks when the backhaul is limited. *IEEE Journal on Selected Areas in Communications, Special Issue on Recent Advances in Heterogeneous Cellular Networks*, Accepted, Feb 2015.

[41] Jagadish Ghimire and Catherine Rosenberg. On the need for coordination among base stations in a heterogeneous network. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2012 10th International Symposium on*, pages 254–261. IEEE, 2012.

[42] Jagadish Ghimire and Catherine Rosenberg. Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach. *IEEE Trans. Wireless Commun.*, 12(3):1340–1351, 2013.

[43] Jagadish Ghimire, Catherine Rosenberg, and Shalini Periyalwar. Why are relays not always good for you? performance of different relay deployment configurations in a heterogeneous network. In *2014 10th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, pages 254–261. IEEE, 2014.

[44] A. Gjendemsj, D. Gesbert, G.E. Oien, and S.G. Kiani. Binary power control for sum rate maximization over multiple interfering links. *Wireless Communications, IEEE Transactions on*, 7(8):3164–3173, 2008.

[45] J. Gora and S. Redana. In-band and out-band relaying configurations for dual-carrier LTE-Advanced system. In *Personal Indoor and Mobile Radio Communications (PIMRC), 2011 IEEE 22nd International Symposium on*, pages 1820–1824, 2011.

[46] I. Guvenc, Moo-Ryong Jeong, F. Watanabe, and H. Inamura. A hybrid frequency assignment for femtocells and coverage area analysis for co-channel operation. *Communications Letters, IEEE*, 12(12):880 –882, december 2008.

[47] S.V. Hanly. An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity. *Selected Areas in Communications, IEEE Journal on*, 13(7):1332 –1340, sep 1995.

[48] S. Haykin. Cognitive radio: brain-empowered wireless communications. *Selected Areas in Communications, IEEE Journal on*, 23(2):201–220, 2005.

[49] John D. Hobby and Holger Claussen. Deployment options for femtocells and their impact on existing macrocellular networks. *Bell Labs Technical Journal*, 13(4):145–160, Winter 2009.

[50] Ekram Hossain, Long Bao Le, and Dusit Niyato. *Radio resource management in multi-tier cellular wireless networks*. John Wiley & Sons, 2013.

[51] Huwawei. The second phase of LTE-Advanced. `http://www.huawei.com/ilink/en/download/HW_259010`.

[52] Han-Shin Jo, Ping Xia, and J.G. Andrews. Downlink femtocell networks: Open or closed? In *Communications (ICC), 2011 IEEE International Conference on*, pages 1 –5, june 2011.

[53] V. Jungnickel, K. Manolakis, S. Jaeckel, M. Lossow, P. Farkas, M. Schlosser, and V. Braun. Backhaul requirements for inter-site cooperation in heterogeneous lte-advanced networks. In *2013 IEEE International Conference on Communications Workshops (ICC)*, pages 905–910, June 2013.

[54] A. Karnik, A. Iyer, and C. Rosenberg. Throughput-optimal configuration of fixed wireless networks. *IEEE/ACM Trans. Netw.*, 16(5):1161–1174, 2008.

[55] Frank Kelly. Charging and rate control for elastic traffic. *European transactions on Telecommunications*, 8(1):33–37, 1997.

[56] Jeff L. Kennington. A survey of linear cost multicommodity network flows. *Operations Research*, 26(2):pp. 209–236, 1978.

[57] A. Khandekar, N. Bhushan, J. Tingfang, and V. Vanghi. LTE-Advanced: Heterogeneous networks. In *Wireless Conference (EW), 2010 European*, pages 978–982. IEEE, 2010.

[58] Daewon Lee, Hanbyul Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana. Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges. *Communications Magazine, IEEE*, 50(2):148–155, February 2012.

[59] L. Lei, C. Lin, J. Cai, and X. Shen. Flow-level performance of opportunistic OFDM-TDMA and OFDMA networks. *"IEEE Trans. Wireless Commun."*, 7(12):5461–5472, 2008.

[60] Wan Lei, Wu Hai, Yu Yinghui, and Zesong Fei. Heterogeneous network in LTE-advanced system. In *Communication Systems (ICCS), 2010 IEEE International Conference on*, pages 156 –160, nov. 2010.

[61] L. Li, M. Pal, and Y.R. Yang. Proportional fairness in multi-rate wireless LANs. In *Proc. IEEE INFOCOM 2008*, pages 1004–1012. IEEE, 2008.

[62] D. Lopez-Perez, Xiaoli Chu, and I. Guvenc. On the expanded region of picocells in heterogeneous networks. *Selected Topics in Signal Processing, IEEE Journal of*, 6(3):281–294, June 2012.

[63] D. Lopez-Perez and H. Claussen. Duty cycles and load balancing in hetnets with eicic almost blank subframes. In *Personal, Indoor and Mobile Radio Communications (PIMRC Workshops), 2013 IEEE 24th International Symposium on*, pages 173–178, Sept 2013.

[64] D. López-Pérez, G. de la Roche, A. Valcarce, A. Juttner, and J. Zhang. Interference avoidance and dynamic frequency planning for WiMAX femtocells networks. In *Communication Systems, 2008. ICCS 2008. 11th IEEE Singapore International Conference on*, pages 1579–1584. IEEE, 2008.

[65] D. Lopez-Perez, A. Ladanyi, A. Juttner, and Jie Zhang. Ofdma femtocells: A self-organizing approach for frequency assignment. In *Personal, Indoor and Mobile Radio Communications, 2009 IEEE 20th International Symposium on*, pages 2202–2207, Sept 2009.

[66] J. Luo, C. Rosenberg, and A. Girard. Engineering wireless mesh networks: Joint scheduling, routing, power control, and rate adaptation. *IEEE/ACM Trans. Netw.*, 18(5):1387–1400, 2010.

[67] Zhangchao Ma, Wei Xiang, Hang Long, and Wenbo Wang. Proportional fair-based in-cell routing for relay-enhanced cellular networks. In *Wireless Communications and Networking Conference (WCNC), 2011 IEEE*, pages 381 –385, March 2011.

[68] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and Tingfang Ji. Cell association and interference coordination in heterogeneous LTE-A cellular networks. *IEEE J. Sel. Areas Commun.*, 28(9):1479 –1489, december 2010.

[69] P. Marsch and G. Fettweis. Static clustering for cooperative multi-point (comp) in mobile communications. In *Communications (ICC), 2011 IEEE International Conference on*, pages 1–6, June 2011.

[70] P. Milgrom and I. Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

[71] Jeonghoon Mo and Jean Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking (ToN)*, 8(5):556–567, 2000.

[72] Daniel Pérez Palomar and Mung Chiang. A tutorial on decomposition methods for network utility maximization. *IEEE J. Sel. Areas Commun.*, 24(8):1439–1451, 2006.

[73] Patrick Donegan. Small Cell Backhaul: What, Why and How? (white paper). [Online], Website, July 2012. `http://www.tellabs.com/resources/papers/tlab_smallcellbackhaul_wp.pdf`.

[74] S.W. Peters, A.Y. Panah, K.T. Truong, and R.W. Heath. Relay architectures for 3GPP LTE-Advanced. *EURASIP Journal on Wireless Communications and Networking*, 2009:1, 2009.

[75] Tongwei Qu, Dengkun Xiao, and Dongkai Yang. A novel cell selection method in heterogeneous LTE-advanced systems. In *Broadband Network and Multimedia Technology (IC-BNMT), 2010 3rd IEEE International Conference on*, pages 510 –513, oct. 2010.

[76] S. Rajagopal, S. Abu-Surra, Zhouyue Pi, and F. Khan. Antenna array design for multi-gbps mmwave mobile broadband communication. In *Global Telecommunications Conference (GLOBECOM), 2011 IEEE*, pages 1–6.

[77] T.S. Rappaport, Shu Sun, R. Mayzus, Hang Zhao, Y. Azar, K. Wang, G.N. Wong, J.K. Schulz, M. Samimi, and F. Gutierrez. Millimeter wave mobile communications for 5G cellular: It will work! *Access, IEEE*, 1:335–349, 2013.

[78] R.T. Rockafellar. *Convex analysis*, volume 28. Princeton Univ Pr, 1997.

[79] C. Rosenberg, J. Luo, and A. Girard. Engineering wireless mesh networks. In *Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium on*, pages 1–6. IEEE, 2008.

[80] Mustafa E. Sahin, Ismail Guvenc, and Huseyin Arslan. An iterative interference cancellation method for co-channel multicarrier and narrowband systems. *Physical Communication*, 4(1):13 – 25, 2011.

[81] A.H. Sakr and E. Hossain. Location-aware cross-tier coordinated multipoint transmission in two-tier cellular networks. *Wireless Communications, IEEE Transactions on*, 13(11):6311–6325, Nov 2014.

[82] Abdallah Bou Saleh, Simone Redana, Jyri Hämäläinen, and Bernhard Raaf. Comparison of relay and pico eNB deployments in LTE-Advanced. In *VTC Fall*, 2009.

[83] Abdallah Bou Saleh, Simone Redana, Jyri Hämäläinen, and Bernhard Raaf. On the coverage extension and capacity enhancement of inband relay deployments in lte-advanced networks. *JECE*, 2010:4:1–4:10, Jan 2010.

[84] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, Young-Doo Kim, Wonjae Shin, and Eungsun Kim. Fairness-aware joint routing and scheduling in OFDMA-based cellular fixed relay networks. In *Communications, 2009. ICC '09. IEEE International Conference on*, pages 1 –6, june 2009.

[85] Rajasekhar Sappidi, Sajjad Mosharrafdehkordi, Catherine Rosenberg, and Patrick Mitran. Planning for small cells in a cellular network: why it is worth it. In *IEEE Wireless Commun. and Networking Conference (WCNC)*, April 2014.

[86] S. Shabdanov, P. Mitran, and C. Rosenberg. Cross-layer optimization using advanced physical layer techniques in wireless mesh networks. *IEEE Trans. Wireless Commun.*, 11(4):1622 –1631, april 2012.

[87] S. Shabdanov, C. Rosenberg, and P. Mitran. Joint routing, scheduling, and network coding for wireless multihop networks. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2011 International Symposium on*, pages 33 –40, may 2011.

[88] D.B. Shmoys and É. Tardos. An approximation algorithm for the generalized assignment problem. *Mathematical Programming*, 62(1):461–474, 1993.

[89] O. Somekh, O. Simeone, A. Sanderovich, B.M. Zaidel, and S. Shamai. On the impact of limited-capacity backhaul and inter-users links in cooperative multicell networks. In *42nd Annual Conference on Information Sciences and Systems, 2008. CISS 2008*, pages 776–780, March 2008.

[90] Rayadurgam Srikant. *The mathematics of Internet congestion control*. Springer, 2004.

[91] Karthikeyan Sundaresan and Sampath Rangarajan. Efficient resource management in OFDMA femto cells. In *Proceedings of the tenth ACM international symposium on Mobile ad hoc networking and computing*, MobiHoc '09, pages 33–42, New York, NY, USA, 2009. ACM.

[92] Karthikeyan Sundaresan and Sampath Rangarajan. Efficient resource management in OFDMA femto cells. MobiHoc '09, pages 33–42, New York, NY, USA, 2009. ACM.

[93] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. In *Decision and Control, 1990., Proceedings of the 29th IEEE Conference on*, pages 2130 –2132 vol.4, Dec 1990.

[94] Alvin Ting, David Chieng, Kae Hsiang Kwong, Ivan Andonovic, and K.D. Wong. Dynamic backhaul sensitive network selection scheme in LTE-WiFi wireless hetnet. In *2013 IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, pages 3061–3065, Sept 2013.

[95] UMTS Forum. Mobile traffic forecasts, 2010-2020 report, report number 44, January 2011.

[96] H. Viswanathan and S. Mukherjee. Performance of cellular networks with relays and centralized scheduling. In *Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th*, volume 3, pages 1923 – 1928 Vol.3, Oct. 2003.

[97] Cheng-Xiang Wang, F. Haider, Xiqi Gao, Xiao-Hu You, Yang Yang, Dongfeng Yuan, H. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir. Cellular architecture and key

technologies for 5G wireless communication networks. *Communications Magazine, IEEE*, 52(2):122–130, February 2014.

[98] Ping Wang, Hai Jiang, Weihua Zhuang, and H.V. Poor. Redefinition of max-min fairness in multi-hop wireless networks. *Wireless Communications, IEEE Transactions on*, 7(12):4786 –4791, december 2008.

[99] Jeong woo Cho, Jeonghoon Mo, and Song Chong. Joint network-wide opportunistic scheduling and power control in multi-cell networks. *Wireless Communications, IEEE Transactions on*, 8(3):1520 –1531, march 2009.

[100] Ping Xia, V. Chandrasekhar, and J.G. Andrews. Open vs. closed access femtocells in the uplink. *Wireless Communications, IEEE Transactions on*, 9(12):3798 –3809, december 2010.

[101] Hao Xu and Pinyi Ren. Joint user scheduling and power control for cell-edge performance improvement in backhaul-constrained network mimo. In *2013 IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, pages 1342–1346, Sept 2013.

[102] Y. Yang, H. Hu, J. Xu, and G. Mao. Relay technologies for WiMAX and LTE-advanced mobile systems. *Communications Magazine, IEEE*, 47(10):100–105, 2009.

[103] R.D. Yates and Ching-Yao Huang. Integrated power control and base station assignment. *Vehicular Technology, IEEE Transactions on*, 44(3):638 –644, aug 1995.

[104] K. Yonezawa, K. Yamazaki, and T. Inoue. Performance evaluation of centralized control algorithm for channel allocation in pico-cell system. In *Vehicular Technology Conference, 2007. VTC-2007 Fall. 2007 IEEE 66th*, pages 1659 –1663, 30 2007-oct. 3 2007.

[105] Wei Yu, Taesoo Kwon, and Changyong Shin. Joint scheduling and dynamic power spectrum optimization for wireless multicell networks. In *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2010.