ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2016-ENST-0082

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité Communication et Electronique

présentée et soutenue publiquement par

Nikolaos SAPOUNTZIS

le 14 decembre 2016

Optimisation au niveau réseau dans le cadre des réseaux hétérogènes nouvelle génération

Directeur de thèse: **Thrasyvoulos SPYROPOULOS**

Jury

M. Christian BONNET, Professeur, EURECOM	Président du jury
M. Albert BANCHS , Professeur, Universidad Carlos III de Madrid	Rapporteur
M. Christos VERIKOUKIS, Professeur, Barcelona University	Rapporteur
M. Umer SALIM, Ingénieur de Recherche, INTEL	Examinateur
M. Navid NIKAEIN, Maître de Conférences HDR, EURECOM	Examinateur
M. Walid DABBOUS, Directeur de Recherches, INRIA	Examinateur
M. Tommy SVENSSON, Maître de Conférences, Chalmers University of Technology	Examinateur

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

T H È S E

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

EURECOM
Sophia Antipolis

DISSERTATION

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
from Telecom ParisTech

Specialization

Communication and Electronics

École doctorale Informatique, Télécommunications et Électronique (Paris)

Presented by

Nikolaos SAPOUNTZIS

Network Layer Optimization for Next Generation Heterogeneous Networks

Defense scheduled on December 14^{th} 2016

before a committee composed of:

Prof. Christian Bonnet	President of the Jury
Prof. Christos Verikoukis	Reporter
Prof. Albert Banchs	Reporter
Dr. Umer Salim	Examiner
Prof. Tommy Svensson	Examiner
Prof. Walid Dabbous	Examiner
Prof. Navid Nikaein	Examiner
Prof. Thrasyvoulos Spyropoulos	Thesis Supervisor

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

EURECOM
Sophia Antipolis

THÈSE

présentée pour obtenir le grade de
Docteur de
Telecom ParisTech

Spécialité

Communication et Electronique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Nikolaos SAPOUNTZIS

Optimisation au niveau réseau dans le cadre des réseaux hétérogènes nouvelle génération

Soutenance de thèse prévue le 14 decembre 2016

devant le jury composé de:

Prof. Christian Bonnet	Président du jury
Prof. Christos Verikoukis	Rapporteur
Prof. Albert Banchs	Rapporteur
Dr. Umer Salim	Examinateur
Prof. Tommy Svensson	Examinateur
Prof. Walid Dabbous	Examinateur
Prof. Navid Nikaein	Examinateur
Prof. Thrasyvoulos Spyropoulos	Directeur de thèse

Abstract

By 2016, it is no secret to the global networking community: there are not adequate bandwidth resources to go around anymore. The growing traffic demand rate is expected to overpass the 10 exabytes per month soon, driven by not only the 10 billion user equipments (UE) but also the emerging Machine Type Communications (MTC) applications that are anticipated to bring online 50 billion other devices. Such differentiated types of communications, offering a plethora of differentiated services, sharpen the traffic demand heterogeneity adding tough constraints in terms of latency, capacity, jitter, etc. Operators struggling with such traffic increase bet on aggressively dense deployments, overlaying the conventional macro cell, where low-power small cells dominate. Such *heterogeneous network* (HetNet) deployments will allow to spatially reuse the given spectrum and provide additional capacity (e.g., in areas with dense usage such as malls, airports, stadiums, etc.) as well as considerably improve spectral efficiency.

Nevertheless, the higher the deployment density the higher the chance that these networks will suffer from intense spatio-temporal variations. Such fluctuations can create serious problems in terms of performance, if not well studied. For instance, the arising load imbalance of such deployments shall drive some Base Stations (BS) to congestion while leaving some other BSs idle. Obviously, the former BSs shall offer rather subpar Quality of Service (QoS) to their users as well as deteriorate (overall) system performance. On the other hand, the power wastage of the latter BSs that serve little or no traffic while ON, is an issue under scrutiny for system energy efficiency.

Additionally, the aggressive small cell densification, followed by a tremendous capacity crunch, threatens the capacities of the corresponding backhaul links that provide connectivity to the core. In other words, next generation networks are expected to be mostly dominated from backhaul links that are *under-provisioned*, i.e. incapable of meeting the tough capacity requirements that radio access network pose. Usually, in such deployments multiple BS might have to share the capacity of a single backhaul link due to, e.g, point-to-multipoint (PMP) or multi-hop mesh topologies to the aggregation nodes. A complete end-to-end backhaul path might also consist of links with different backhaul technologies (e.g., wired or wireless) and thus different capabilities and requirements (e.g., in terms of capacity or delay). To that end, backhaul network emerges as a complex performance bottleneck, and schemes that do not carefully consider it might lead to poor performance.

The goal of this dissertation is twofold: (i) consider multiple dimensions in the traffic demand entity to better reflect the heterogeneity of the services offered to users and the involved requirements, as well as, (ii) optimization of various networking problems by jointly considering the radio access and backhaul networks. This will allow us to achieve a good multi-dimension tradeoff between different performance metrics (e.g., spectral efficiency versus load balancing, or user QoS versus energy efficiency), preempt congestion issues, as well as reveal interesting

dependencies and bottlenecks for future HetNets. Towards this direction, we will perform appropriate modeling, performance analysis, optimization for a family of objectives, using tools mostly coming from (non) convex optimization, probability and queueing theory.

In particular, in Chapter 1 we provide a brief introduction to next generation heterogeneous networks and the motivation of our work.

In Section 2, we start by investigating the popular *user association* problem by solely focusing on the radio access network. We adopt an α-fair family of objective functions widely considered for user association, that directs the optimal solution towards different goals (e.g. throughput optimal, delay-optimal, load balancing, etc.), and extend it considerably to better capture the impact of traffic differentiation on radio access network performance. More precisely, we formulate a convex optimization problem that studies (i) traffic differentiation between two traffic classes (elastic and non-elastic flows), and (ii) uplink and downlink traffic performance, a controversy to most of the related works in this context that usually investigate solely DL elastic traffic demand. We then analytically prove different "device centric" user association rules that end up maximizing either an arithmetic or a weighted harmonic mean of the achieved performance along different dimensions, depending on whether uplink and downlink traffic of the same user can be "split" to different BSs or not. Finally, we portray that our derived formulas can be generalized and allow the inclusion of even more traffic dimensions in the problem setup and flexible derivation of the corresponding optimal user association rules without any analytical calculations.

In Chapter 3, we extend our framework for α-fair user association to extensively consider the backhaul network limitations along with the radio access. In particular, we will try to shed some light on the impact of (i) backhaul link capacities, and (ii) backhaul topology in different under-provisioned scenarios. To do so, we include into our optimization problem appropriate backhaul constraints, to ensure backhaul congestion avoidance, that eventually re-direct the optimal point towards a shrinked feasible set. Our challenge is to keep the user association rules "device centric", even in scenarios where backhaul topologies are rather mesh. To that end, using penalty functions for the emerging backhaul constraints we analytically derive novel backhaul-aware user association rules that promise scalability irrespectively of the backhaul topology. We also highlight that our rules still allow for the arithmetic or harmonic mean formula usage even in under-provisioned backhaul scenarios.

In Chapters 2 and 3 we focused on various tradeoffs of the access and backhaul networks, by assuming fixed bandwidth resource allocation between uplink and downlink (e.g., 50-50) on the Medium Access Control (MAC) scheduler. However, the (i) asymmetric transmit powers of UEs and different BSs leading to different physical data rates, along with the (ii) asymmetric traffic applications (e.g., social media or video game applications), necessitate the need of matching the uplink and downlink resources to the actual demand. To this end, in Chapter 4 we investigate the opportunity of flexible Time Division Duplex (TDD) schemes, where e.g., each BS (or, each backhaul link) having a limited amount of time resources can allocate them between the downlink and uplink dimension based on the system/traffic dynamics. We then show that optimizations of such TDD allocation shall interact jointly with UL/DL user association, since separate optimization can lead to unnecessary performance deterioration. To that end, we develop an α-fair optimization framework that tackles the interplay of (i) user-association, (ii) radio resource allocation, and (iii) backhaul resource allocation of TDD resources. We propose an algorithm that reduces the complexity of this non-convex problem by decomposing it into three optimization subproblems, each potentially solved by a different network element and at

different timescales. We finally show that under some certain circumstances such optimization converges to the global optimum, by offering up to 3× higher performance in certain scenarios.

Eventually, in Chapter 5, we focus on energy efficiency, by investigating the trade off between different user QoS criteria and energy savings. Specifically, we consider a novel sleep mode scheme where BSs are switched-off to minimize energy, subject to three (BCD) QoS constraints: Blocking probability, Coverage failure probability and Delay. The duration of the switching-off period plays a key-role in the modeling of the various QoS constraints. We claim that while short sleeping periods for small cells can result in a more complex analysis for the user QoS (since the lack of the system convergence can invalidate the stationary formulas usually considered, e.g. the Erlang-B for blocking probability), they can promise high energy savings when properly considered. Towards this direction we strike a tradeoff between realistically capturing some features of next generation cellular systems, while maintaining a certain analytical tractability to provide insights into the user QoS vs. Energy savings.

In chapter 6 we conclude the thesis and discuss about future research directions.

Acknowledgements

First and foremost I would like to express my gratitude to my supervisor Thrasyvoulos Spyropoulos. He is a great professor, not only because of his ability to provide prudent and brilliant guidance throughout tough research problems, but also due to his talent of maintaining a rather interacting, inspiring, and yet enjoyable atmosphere within work, as well as his charisma to make you think out-of-the-box. Additionally, I would like to thank my co-supervisor Navid Nikaein for his continuous guidance as well as excellent assistance on the various technical and challenging problems we went through during these years. This dissertation was only made through their patient support and trustiness.

The completion of Ph.D., a highly complex problem, is challenging and colleagues define its feasibility region. I want to thank the people of my group: Panos, Pavlos, Luigi and Delia that were always willing to discuss about research problems and offer me valuable advice.

Nevertheless, the life of a Ph.D. student is not hard constrained within such professional, and usually flat, bounds. Family relaxes them by driving it an ascent way. While my real family is kilometers away in Crete, I was lucky enough to construct from scratch a new one here, in Nice. First and foremost, I would like to thank Katerina: having known her for 10 years, she was always there, in every kind of problem I had to deal with; health-related, working or personal. The tradeoff between yolo and soberness is difficult to balance, and sometimes family battles appear in this context: thanks to Dorine and Panos I learnt to achieve a great equilibrium between them (or, to slope down to the left). All of our shared moments (e.g., enjoying amazing memories either locally or through (inter)national travels as well as struggling with tough situations in so many different domains) were of utmost value and importance for me.

Yet, life, even Ph.D. life, needs some sort of sprightliness and fluctuation, to avoid getting stuck in monotonicity. My dissertation would not be interesting without the various groups of friends I made, and the outstanding moments I have had with them. Starting with my lovely Tapaloca-group, I would like to thank: Valia, Maroui, Akis, Odile, Salvatore, Carolina, Miguel, Frank, San Jose and many others for helping me relax during so many Saturday nights through the uncountable meaningful (and, meaningless) discussions. Special thanks to Pagourakis and the great as well as unforgettable moments we had on our stamping grounds. Finally, I would like to express my thankfulness to the rest of my great friends including Loukas and Catt with whom I had so nice and simultaneously "battling" moments in France and California, as well as Stelman, Ntinos, Pavlos, Carol, Arvanitakis, Monir, Stelios, Katsalis, Giannis, TTT, Sygkounas, Stefanos, Eftuxia, Ne, Miltos, Iraklis, Alekos, Lefteris, Angelita, Delia, Dimitris and Dora.

Last but not least, I would like to thank as well as dedicate this thesis to my lovely parents Dimitra and Konstantinos, as well as my beloved sister Maria, brother Stelios and aunt Nitsa, for everything they have offered me in my life. In particular, their constant support, love and boundless encouragement, significantly helped me towards completing this work.

Contents

List of Figures

List of Tables

Acronyms

3GPP	3rd Generation Partnership Project.
4G	Fourth Generation.
5G	Fifth Generation.
ABS	Almost Blank Subframe.
ACS	Alternative Convex Search.
AU	Active Users.
BCD	Block Cordinate Descent.
BH	Backhaul.
BS	Base Station.
C-RAN	Cloud-Radio Access Network.
CAPEX	Capital Expenses.
CoMP	Coordinated MultiPoint.
CTMC	Continuous Time Markov Chain.
CU	Connected Users.
DL	Downlink.
DSL	Digital Subscriber Line.
DTMC	Discrete Time Markov Chain.
DU	Disconnected Users.
eNB	evolved Node B.
eICIC	enhanced Inter-Cell Interference Coordination.
eIMTA	enhanced Interference Mitigation and Traffic Adaptation.
FDD	Frequency Division Duplexing.

GBR Guaranteed Bit Rate.

HetNet Heterogeneous Networks.

HFC Hybrid Fiber-Coax.

HO Handover.

MAC Medium Access Control.

MBS Macro Base Station.

MC Markov Chain.

LB Load Balancing.

LoS Line-of-Sight.

LTE Long Term Evolution.

LTE-A Long Term Evolution-Advanced.

MBS Macro Base Station.

non-GBR non-Guaranteed Bit Rate.

OPEX Operational Expenses.

P2P Point to Point.

P2MP Point to MultiPoint.

PASTA Poisson Arrivals See Time Averages.

PHY Physical Layer.

PS Processor Sharing.

QCI Quality of service Class Identifier.

QoS Quality of Service.

RAN Radio Access Network.

SLA Service Level Agreements.

SC Small Cell.

SE Spectral Efficiency.

SINR Signal-to-Interference-plus-Noise Ratio.

SNR Signal-to-Noise Ratio.

SC Small Cell.

TDD Time Division Duplexing.

UL Uplink.

UE User Equipment.

VoIP Voice over IP.

Chapter 1

Introduction.

1.1 Heterogeneous Networks

Wireless cellular networks typically consist of a set of User Equipments (UEs) and a collection of Base Stations (BSs) that connect to the core network through a set of backhaul (BH) links. In the traditional networks, traffic intensity and demand across different UEs usually remain similar. Moreover, the BSs have alike transmit power levels, antenna patterns as well as backhaul connectivity to the core.

Nowadays, the exponentially *growing* and *varying* traffic demand for UE services arising from different applications is an emerging reality [1]. In particular, the "bandwidth hungry" augmented-reality, social-networking as well as the various Machine Type Communication (MTC) (e.g., monitoring healthcare or energy systems) applications pose tough capacity and latency requirements. Operators struggling to cope with this traffic increase tend to build more dense network deployments to improve spatial reuse. Specifically, they build additional low-transmit power small cells (SC) along with the already existing high-transmit power macro BSs (MBS). Let us point out that the power levels of the MBS is usually between 5 - 40 W, the ones for SCs is only $0.25 - 2$ W. Networks composed of a mixture of different BSs with different power levels (and thus different cell sizes) are called *Heterogeneous Networks* (HetNets).

Such HetNets have attracted the interest of Long Term Evolution (LTE) and Long Term Evolution-Advanced (LTE-A) systems. From the earliest steps, LTE started studying such deployments and referred to the different BSs as macro-, micro-, pico- and femto-cells; listed in decreasing transmit power order. Later, additional SC types introduced with different abilities and functionalities. For instance, in LTE Release 9 the Home eNB (HeNB) concept was introduced. HeNB is mainly used to provide indoors coverage, for Closed Subscriber Groups, e.g., in office premises without major macro-network coordination. In LTE Release 10 the Relay Node (RN) introduced as another low-power SC. RNs are BSs that offer enhanced coverage and capacity at hot-spot areas can also be used to connect to remote areas without fiber connection.

While a dense HetNet topology is a promising opportunity to meet the growing traffic demand, it requires a careful backhaul planning and support to serve the large number of SCs. To that end, current research appears to study the new backhaul requirements in terms of Capital Expenditure (CAPEX) and Operational Expenditure (OPEX), coverage, capacity, security, delay, synchronization, physical design and management compared to the traditional ones posed from macrocells. Next Generation Mobile Networks Alliance [2] and Small Cell Forum [3] provide

Wireless	Wired
Millimetre 70 - 80 GHz	Direct fibre
Millimetre 60 GHz	Digital subscriber line (xDSL)
Microwave point-to-point	FTTx
Microwave point-to-multipoint	Hybrid fibre-coax (HFC) and DOCSIS 3.0
Sub 6GHz Licensed	-
TVWS	-
Satellite	-

Table 1.1: Backhaul Technologies.

various crucial contributions to address them.

One of the most important parameters for the backhaul planning and the related system capabilities is undoubtedly the backhaul technology of the link(s). Two backhaul network families are widely considered in the related literature: the *wired* and *wireless* backhaul link technologies, as listed in Table 1.1 [3]. Their rather unsymmetrical characteristics make each of them preferable in different scenarios. For instance, wired is usually an expensive backhaul option and often impossible to deploy in rural areas with sparse BSs, hence making wireless a more viable solution. There is also competition in intra-family technologies. E.g., in the wireless family, the ability to provide Line-of-Sight (LoS) propagation, the duplexing mode, the licensing arrangement, the ability for point-to-point (P2P) or point-to-multipoint (P2MP) connectivity are some of the characteristics that suggest different wireless technologies to be better than others in different scenarios. To that end, different operators shall choose different backhauling strategies, and the complete backhauling end-to-end paths are expected to be dominated from multiple technologies.

1.2 HetNet challenges and related work.

Radio access network, implementing the radio technologies at the BS level, have played a major concern in HetNets during the last decade for both industrial and academic communities. The latters have tried to tackle a plethora of different problems arising from different levels, where typically the issue under scrutiny has been to improve the overall system performance. Such performance improvements revolve around the following efficiency metrics:

- *Area Spectral Efficiency*, capturing how efficiently a limited spectrum is utilized per area unit. Typically, HetNets promise significant enhancements since (i) the short-coverage area SCs are allowed to reuse the spectrum locally, and (ii) the UEs are brought closer to the BSs by enhancing the link quality.

- *Load Balancing Efficiency*, reflecting how well the traffic loads are distributed among different BSs can be improved since the high number of available SCs can offload the (usually congested) macro BSs so that the total load is distributed more evenly.

- *Energy Efficiency*, portraying the potential energy redundancy when the network runs, can be improved by switching off the under-utilized SCs.

Nevertheless, these efficiency metrics are not necessarily aligned and there are some non-trivial conflicts between them. To that end, a great amount of works that attempt to achieve a good tradeoff between them have appeared recently. For instance, since power and bandwidth constrict the achievable gains under differentiated manners, many works have appeared, trying to achieve a good tradeoff between energy and spectral efficiency [4, 5]. Load balancing and energy efficiency coupling is investigated in [6–9]. The issue under scrutiny upon this tradeoff is usually whether an under-utilized BS should be switched-off (to improve energy efficiency) or stay switched-on and carry more traffic from neighboring over-utilized BSs (to improve load-balancing). In [10] the tradeoff between spectral and load balancing efficiency is well studied, where the authors explain that different weights of such a tradeoff are able to improve different performance metrics.

Things get more complicated when user Quality of Service (QoS) enters into the picture.

- *User QoS* is the performance seen by the users, and is often considered in terms of service delay, error rates, blocking probability, throughput, jitter, etc.

User QoS has an explicit relationship with spectral, load balancing and energy efficiency. As an illustration, there has been an inconclusive debate about whether spectral or load balancing efficiency plays the key role to optimize user QoS. Believing that "spectral efficiency is the main driver of user QoS", as most of the people tend to think, is an old fashion myth as stated from Andrews et al., in [11]. There, as well as, in other related works [10, 12–16], it was shown that user QoS heavily depends also on the BS loads. To put it differently, when the system is highly loaded user QoS is usually killed. Additionally, another great amount of work tries to achieve a good tradeoff between the contradicting notions of energy efficiency and user QoS [5, 6, 17, 18]: switching-off BSs to save energy reduces the resources given to the users and harms their QoS. This tradeoff is a major concern for network greening and its importance is highlighted for next generation systems.

Typically, achieving a good tradeoff between such conflicting performance metrics (e.g., user QoS, energy efficiency, etc.) is usually a rather challenging task. Current literature on this abounds on examples that lie at the heart of different involved problems. We turn now our attention on such problems and discuss in detail how current systems (e.g., LTE or LTE-A) treat them to tradeoff different metrics. Such problems and involved tradeoffs for next generation networks, will be one of the major concerns of this dissertation.

1.2.1 Problem 1: User Association in HetNets.

User association is the problem of associating users with BSs.

In conventional cellular systems, this problem was tackled by associating each user to the BS with maximum SINR: such association rule was the base up to LTE-release 8. While this rule also maximizes the instantaneous rate of a user (i.e., the best modulation and coding scheme - MCS - supported), it reflects user QoS only when the BS is lightly loaded [11]. For example, user performance, in terms of *per flow delay* or *throughput*, may be severely affected if the BS offering the best SINR is congested [14,19]. Thus, current user association algorithms in HetNets with intense traffic fluctuations suggest to explicitly consider the two conflicting concerns of: (i) maximizing the spectral efficiency, and (ii) ensuring that the load across BSs is balanced to improve load balancing and preempt congestion events.

A number of research works have studied the problem of user association in HetNets, optimizing user rates [20–22], balancing BS loads [12, 23–27], or pursuing a weighted tradeoff of them [10, 12, 28] to better reflect different user QoS degrees. Cell range Expansion (CRE) techniques, where the SINR of lightly loaded BSs is biased to make them more attractive to the users are also popular for the latter tradeoff in user association [29–34]. Interference-aware and cooperative communication techniques such as coordinated beamforming [35, 36], coordinated multi-point (CoMP) [37], and Device-to-device cooperation [38] were also proposed in this context. Other related studies include the popular distributed user association algorithm proposed in [39], where the global outage probability and the long term rate maximization are jointly considered in the context of load balancing. The authors in [40] propose a framework that studies the interplay of user association and resource allocation in future HetNets, by formulating a non-convex optimization problem and deriving useful performance upper bounds.

Finally, a user association framework that has received much attention is [10]. This framework jointly considers a family of objective functions, each of which directs the optimal solution towards different goals using an iterative algorithm. This framework has a similar form with the α-fair utility function [41]. Specifically, by changing the α parameter one can flexibly optimize different performance metrics, such as spectral or load balancing efficiency, delay, throughput etc. [6,42,43] extend this framework to further include energy management, e.g., by switching off under-loaded BSs. There, a key input parameter weights the importance of power savings versus the original objective and the optimal solution is flexibly re-directed under the new feasible set.

1.2.2 Problem 2: Flexible TDD schemes in HetNets.

Typically, in wireless cellular systems, each BS is given an amount of bandwidth resources to utilize for both DL and UL traffic by duplexing on the frequency (Frequency Division Duplex-*FDD*) or the time (Time Division Duplex-*TDD*) domain[1].

While in traditional networks FDD was the most popular choice, fixed and flexible TDD started gaining more ground in LTE and LTE-A systems. Having been sketched initially for traditional full coverage MBS deployments, fixed TDD was rather rigid with scarce flexibilities, i.e., a pre-determined and fixed TDD pattern was expected to be selected for the complete network in the longterm [45]. While such static TDD systems prohibited the resource adaptation on potential traffic fluctuations, this was not a major problem in such primary systems because macro cells usually aggregated a large number of users without significant traffic variations.

Nonetheless, in current HetNet deployments where SCs dominate, each BS is usually associated with a small number of concurrent UEs. Thus, intense traffic fluctuations emerge the *dynamic TDD adaptation that match the UL and DL resources to the actual demand* as a key. As an example, focus on a SC that only serves one user. Obviously, system performance (in terms of e.g., user throughput) can be drastically enhanced by TDD adaption depending on whether he is doing uplink or downlink. Motivated by this, LTE Release 12 standardizes a related key-enabler for future 5G networks, namely "enhanced Interference Mitigation and Traffic Adaptation" (eIMTA) [46].

Recently, a few works have appeared that try to match the UL and DL resources to the actual demand in this context. For instance [47] propose that MBSs should be mostly scheduled for DL due to their high transmit power, and SCs for UL to minimize the path loss. Such a

[1]The impossibility for radios to transmit and receive on the same frequency band simultaneously, establishes this fundamental necessity [44].

scheme shall improve spectral efficiency due to the higher achieved SINRs in both directions. Other additional allocation schemes have further been proposed to better allocate the given resources and improve user QoS [48] [49] [50]. In these frameworks, the allocation is realized in a more sophisticated way by weighting the actual traffic demand in both directions, to avoid resource wastage. Performance evaluation shows that user QoS can be significantly improved when resources are flexibly adjusted on the current traffic fluctuations.

1.2.3 Problem 3: BS greening schemes in HetNets.

Minimizing energy consumption is of utmost importance for wireless cellular networks, not only because the latters are a major energy killer worldwide [51,52] and the fact that electricity is the main contributor for their high OPEX [53], but also because environmental protection becomes a global inevitable trend.

In particular, there is overwhelming evidence corroborating the notion that radio access network contribute 60-70% of the total energy consumption in wireless cellular systems [54]. Thus, most of the related works usually focus on decreasing the energy consumption of BSs e.g., see [6,55–61]. In particular, "cell breathing", known also as "cell zooming" techniques are rather popular in this context. These usually include sophisticated BS power management methods where multiple cells can coordinate together to adjust their transmit power according to network or traffic situation [58,60,62].

Additionally, under a certain under-utilization level e.g., during the night hours where traffic load is usually light or negligible, some SCs could annihilate their transmit power i.e. completely switch-off their functionality. There, some MBSs could remain ON, acting as umbrella-BSs, to serve the few UE that remain active. Nowadays, such *sleep mode* techniques, that switch-off BSs under various criteria, have played a major role for network greening. Various frameworks have been proposed in this context, e.g., [6,57,61,63,64], most of them by trying to address the complex tradeoff between power/energy consumption and user QoS.

1.3 Motivation and Contributions of the Thesis

Based on the previous discussion, HetNets indeed offer significant opportunities to improve various performance metrics or achieve a good tradeoff between them in next-generation systems.

1.3.1 Motivation.

Most of current works in this context, following conventionalities of earlier networks usually (i) maintain many traditional assumptions in the flow level and traffic modeling, (ii) ignore the emerging bottlenecks in the backhaul and fronthaul network, as well as (iii) try separate optimizations in coupling networking problems. Such weaknesses indeed can invalidate the offered insights, results, and improvements as well as question their effectiveness by calling for more modern and sophisticated related algorithms. In the rest of the section, we further investigate such weaknesses and omissions of current standards by claiming that they can lead to either wastage of bandwidth resources and system performance degradation. Later, in Section 1.3.2 we will show how one can efficiently address them through major problem revisions.

LTE QCI	**Resource Type**	Example Service
1	GBR	Conversational voice
2	GBR	Live streaming of conversational voice
3	GBR	Real time gaming
4	GBR	Non conversational video(Buffered streaming)
5	Non-GBR	IMS signaling
6	Non-GBR	Video (buffered streaming),TCP based applications
7	Non-GBR	Voice, video (live streaming), interactive gaming
8	Non-GBR	Video (Buffered streaming), TCP based applications
9	Non-GBR	Video (Buffered streaming), TCP based applications

Table 1.2: Flow types along with their QoS Class Identifier (QCI) proposed by LTE.

Traffic differentiation. Most of the current works studying the above-mentioned problems solely assume homogeneous traffic profiles, and usually only focus on the DL direction.

For example, [6, 42, 43, 65–68] assume that all flows generated by a UE are "best-effort" (i.e. elastic). However, modern and future networks will have to deal with high traffic differentiation, with certain flows being able to require specific, *dedicated* (i.e., non-elastic) resources [69]. Such dedicated flows do not share BS resources like best-effort ones, are subject to admission control, and sensitive to different performance metrics. Thus, in our work we propose that one shall explicitly consider the following traffic classes

- *dedicated* flows, where dedicated bearers are allocated for Guaranteed Bit Rate (GBR) type of traffic to meet the required bit rate or latency constraints. In LTE and LTE-A systems, these are differentiated by their QoS class of identifier (QCI) ranging from 1 to 4 (see also Table 1.2),
- *best-effort* flows, related to non-GBR traffic, and QCI from 5 to 9 (see also Table 1.2).

In addition, the majority of related studies only consider downlink (DL) traffic [6, 19, 22, 65, 69–76]. Uplink (UL) traffic is becoming important, due to symmetric (e.g., social networking) applications, MTC, augmented reality games (e.g., PokemonGo) etc. Yet, due to the asymmetric transmit powers of UEs and BSs, leading to different physical data rates, the BS which is optimal for DL traffic might lead to severely degraded performance for UL traffic. Thus, we also differentiate between:

- *downlink* flows, with direction from the BS to the UE, and
- *uplink* flows, with direction from the UE to the BS.

In this context, there are many questions arising and that we attempt to tackle. How shall such traffic differentiation affect the above mentioned problems? For example, shall dedicated flows that are obviously associated with different (e.g., latency or capacity) requirements affect similarly the BS loads and overall performance as the best-effort ones? What is the optimal way that one can deal with the asymmetric UL and DL flow dynamics?

Additionally, in current systems, depending on the operator capabilities and desires one can encounter two techniques for uplink and downlink traffic offloading

- *Split UL/DL*, where a UE is allowed to be associated with two different BSs: one for its UL and one for its DL traffic,
- *Joint UL/DL*, where each UE is required to be connected to a single BS for both UL and DL traffic.

Split UL/DL was standardized in LTE Release 12 [77]. While it is obvious that Split UL/DL can simultaneously optimize both downlink and uplink performance since it allows for two distinct associations, the corresponding improvements are not rather clear yet. Some initial heuristics trying to sketch a few introductory insights on the improvements of Split have appeared recently e.g., [18, 47, 78]. However, the maximum and precise enhancements in different performance metrics are clouded yet and call for more analytical frameworks that investigate them in depth.

Impact of backhaul limitations. Most related works focus on the radio access network e.g., considering the physical user data rate on the access radio interface or BS load, ignoring the backhaul network [4, 6, 9, 23, 31, 37, 40, 47–49, 61, 66, 67, 73, 75, 78–83].

While such an assumption might be reasonable for legacy cellular networks, given that the macrocell backhaul is often over-provisioned (e.g., fiber), this might be quite suboptimal for next generation deployments. There, the considerably higher number of SCs, and related CAPEX and OPEX[2] suggest that backhaul links will mostly be inexpensive wired or wireless (in licensed or unlicensed bands), and under-provisioned [3]. Multiple BS might also have to share the capacity of a single backhaul link due to, e.g, point-to-multipoint (PMP) or multi-hop mesh topologies to the aggregation node(s) [85]. Additionally, the various BS-coordinated schemes that discussed to better exploit the given spectrum (e.g., enhanced Inter-Cell Interference Coordination (eICIC) [82] and Coordinated Multi-Point (CoMP) transmission [37]) are expected to further stress the backhaul network capacities.

Such an emerging bottleneck gets even more difficult to be tackled if one thinks that backhaul network nowadays consists of heterogeneous links with different (e.g., in terms of capacity or latency) capabilities and requirements. Thus, a flow might follow an end-to-end path consisting of links with heterogeneous technologies.

Summing up, as the radio access network technologies (e.g., as analyzed in Chapter 1.2) are constantly improving, it is argued that the backhaul network will emerge as a major performance bottleneck. Thus, the algorithms that ignore the backhaul load, topology and technology will be usually leading to poor performance [70]. This calls for the extensive consideration of the backhaul network and the impact of its limitations.

Joint access and backhaul optimizations. As discussed there are plenty of algorithms that consider various access network functionalities such as user association and dynamic TDD allocation. However, as showed most of them do so in separation. Additionally, while there are some primal works that try to optimize backhaul network functionalities, most of them do so separately with the access network. Thus, the value of such works is questionable since there are strong dependencies between different intra-network functionalities (e.g., user association and TDD allocation) and inter-network ones (e.g., acess and backhaul).

[2]The dense deployments of SCs with low number of users suggest that the *cost* of their backhauling becomes a significant part of the total CAPEX/OPEX, and in some cases could exceed the cost of their equipment [84].

To make this more clear, assume that a UE is connected to a macro BS in the DL (from which it receives the highest signal level), and to an SC in the UL (where the pathloss is lower), as suggested in [47]. If the DL resources of the macro BS, or the UL resources of the SC are not sufficient, this approach can lead to *unnecessary* congestion or under-utilization in either direction. This suggests that, joint optimization of the two different access optimizations for user association and dynamic TDD is key. Similarly, if the corresponding backhaul resources are not well adapted to the requirements that radio access network pose, performance degradation will be inevitable. Thus, backhaul resource allocation policies shall interact with the different radio access policies, in order to satisfy the demands and the requirements that the latter generate. It is not clear yet how such a synchronization can be established, and the quantitative improvements it promises.

Short timescale sleep modes. As discussed previously a large research effort has been initiated recently in the area of "green" networks.

Nevertheless, most past studies are performed not only in the context of simple QoS constraints related to homogeneous traffic profiles (e.g., signal quality as in [86], or traditional blocking probabilities as in [17]), but also under the large time-scale sleeping mode (e.g., turning off BSs during the night or for some hours [6, 8, 17, 23, 28, 51, 52, 83, 87]). In modern and future HetNets, dealing with energy consumption issues becomes more challenging. Significantly more opportunities arise for switching off SCs in smaller time scales (e.g., in the order of some minutes), due to the spatio-temporal load variations as well as their rapid and (mostly) power-free transitions between ON and OFF states. Exploiting such opportunities in that short timescale durations has not been investigated, yet. As a result, a number of interesting questions arise in the context of HetNet that remain unexplored: Should the duration of switching-off period, affect our decision, and if so, how? Which types of users and BSs should one consider when making such a power management decision?

1.3.2 Contributions and Outline.

The focus of this thesis is on trying to provide answers to the above-mentioned questions by revising Problems 1-3. To do so, we use novel objective functions and constraints, more realistic assumptions as well as appropriate modeling to better reflect the arising requirements, bottlenecks and trends for next generation networks. Using tools mostly coming from queueing, probability as well as convex and non-convex optimization theory we provide the optimal solutions for the various considered problems and perform evaluation analysis and comparison with existing work. Throughout our results, we provide various insights, both quantitatively and qualitatively, for a number of different problems and the involved tradeoffs, and we explicitly show the impact of the arisen bottlenecks.

Specifically, the chapters of the thesis, and the main contributions in each one of them, are organized as following:

Chapter 2 - Traffic steering user-association optimizations. As stressed earlier, traffic differentiation is a key limitation in the works related to the *user association* problem.

In this chapter we consider (i) the dedicated flow types along with the best-effort ones as well as (ii) uplink traffic performance. We propose new scheduling disciplines for the dedicated flows at the BS level. We then analytically show how they affect the cell loads and the user QoS under

an α-fairness family of objective functions that optimize different metrics for both split and non-split scenarios. Note that playing with α one can flexibly tradeoff different performance metrics such as: spectral efficiency, load balancing, user throughput, flow delay, etc. Using convex optimization theory we analytically prove a set of optimal user-association rules that end up maximizing either an arithmetic or a weighted harmonic mean of the achieved performance along different dimensions (e.g. UL and DL performance or dedicated and best-effort performance). We underline that our rules are "device-centric" by allowing for distributed implementations. These rules are: scalable (constant amount of the BS broadcast messages irrespective of the number of users), simple (constant complexity of the rule with respect to the number of BSs), and offer flexible performance (defined from α values).

The work related to this chapter is

- *N. Sapountzis, T. Spyropoulos, N. Nikaein, U. Salim, An analytical framework for optimal downlink-uplink user association in HetNets with traffic differentiation, in Proc. IEEE Global Communications (GLOBECOM) Conference, San Diego, CA, USA, 2015.*

The following novel handover algorithm for HetNets, that considers jointly the user- and network- performance, was inspired by our proposed user association algorithm

- *K. Alexandris, N. Sapountzis, N. Nikaein and T. Spyropoulos. "Load-aware Handover Decision Algorithm in Next-Generation HetNets", IEEE WCNC, Doha, Qatar, 2016.*

Chapter 3 - Backhaul aware user-association optimizations. As stressed earlier, backhaul network is an emerging performance bottleneck and most of the existing *user association* frameworks ignore it. As a matter of fact, their performance is questionable.

In this chapter our focus is on backhaul network limitations and their impact on system performance while associating users with BSs. Specifically, we extend the work introduced in Chapter 2 by including (i) various backhaul link capacity constraints, as well as (ii) the backhaul topology. We thus derive novel user association rules that jointly consider radio access and backhaul performance, and end up optimizing the (same) α-fairness family of objective functions under the new feasible set. Finally, we show that our novel backhaul aware user association rules still satisfy the criteria of (i) scalability (constant amount of the BS broadcast messages irrespective of the number of users and backhaul topology), (ii) simplicity, (iii) fairness.

The work related to this chapter is

- *N. Sapountzis, T. Spyropoulos, N. Nikaein, U. Salim, Optimal Downlink and Uplink User Association in Backhaul-limited HetNets, in Proc. IEEE International Conference on Computer Communications (INFOCOM), San Francisco, CA, USA, 2016.*
  - *Best Presentation Award in Heterogeneous Networks Session.*
- *N. Sapountzis, T. Spyropoulos, N. Nikaein, U. Salim, User Association in HetNets: Impact of Traffic Differentiation and Backhaul Limitations, pending major revision, IEEE/ ACM Transactions on Networking (ToN), May 2016.*

Chapter 4 - Flexible TDD Allocation for Access and Backhaul Networks on top of User Association optimizations. As stressed earlier, dynamic TDD allocation at the

BS level is usually tackled separately from user association by threatening the success of future systems. Thus, joint optimization of them along with backhaul network optimizations is key.

In this chapter we develop a framework that tackles the optimal interplay of (i) user-association, (ii) radio resource allocation, and (iii) backhaul resource allocation of TDD resources for the α-fairness family of objective functions. We extend the α-objective function to capture the various fairness degrees upon the new resource allocation parameters. This problem is associated with three optimization variables and it is non-convex on them. We propose an iterative algorithm that reduces the complexity of this problem by decomposing it into three optimization subproblems, each potentially solved by a different network element and at different timescales. We prove convergence to the global optimum, and provide simulation results that demonstrate the performance benefits of our approach.

The work related to this chapter is

- *N. Sapountzis, T. Spyropoulos, N. Nikaein, U. Salim, HoP: Hierarchize then optimize: A distributed framework for user association and flexible TDD allocation for access and backhaul networks, Tech-Report RR-16-328, Eurecom, 2016.*

Chapter 5 - Energy Optimizations subject to user QoS constraints. As stressed earlier, while there are many studies investigating the potential energy savings upon various *sleeping mode* scenarios for BSs, most of them are performed under (i) homogeneous traffic profiles, and (ii) the large timescale assumption. As explained such frameworks fail to capture future HetNet necessities as well as opportunities.

Towards tackling these shortcomings, in this chapter we identify three QoS constraints, related to different ways that the performance of a UE could deteriorate. We then derive analytically the probability of violating each of them, as a function of user and network parameters. Our goal in this direction is to strike a tradeoff between realistically capturing some features of new, data-centric cellular systems, while maintaining a certain analytical tractability to provide insights into the user QoS vs. Energy savings. The main novelty of our methodology is that we can select even a small time-interval, for the sleeping period, and evaluate the energy-QoE tradeoff by switching to transient analysis (rather than stationary analysis) of the stochastic model in hand. Based on these QoE constraints and the time duration, we perform a preliminary study and show that significant energy savings can be achieved even for switching-off periods of the order of some minutes.

More specifically, the user QoS constraints considered include:

- "Blocking" probabilities, i.e. the probability that a flow that requires a certain amount of (dedicated) bandwidth, is blocked due to the lack of the available resources.
- "Coverage Failure probabilities", i.e. the probability that a random UE experiences poor signal quality when it needs to use the network (e.g. making a call, or sending a web request).
- "Delay" for regular "best-effort" flows, i.e. the ongoing delay for the flows that are multiplexed and have to compete for resources.

The works related to this chapter are

- *N. Sapountzis, T. Spyropoulos, N. Nikaein, U. Salim, Reducing the energy consumption of small cell networks subject to QoE constraints, in Proc. IEEE Global Communications (GLOBECOM) Conference, Austin, TX, USA, 2014.*
- *D. Wang, E. Karathanaras, A. Quddus, N. Sapountzis, L. Cominardi, F. Kuo, P. Rost, C.J. Bernardos, I. Berberana, SDN-based Joint Backhaul and Access design for Efficient Network Layer Operations, in Proc. IEEE European Conference on Networks and Communications (EuCNC), Paris, France, 2015.*

Chapter 2

Traffic-steering User Association Optimizations.

2.1 Introduction

As explicitly discussed in Chapter 1, future HetNet deployments are expected to experience high spatio-temporal load fluctuations. This suggests that conventional *user association* schemes, e.g., associating a UE with the BS with the highest SINR, shall offer subpar performance by calling for more sophisticated *user association* algorithms.

There are two, often conflicting, concerns when assigning UEs to a BS in the modern association schemes: (i) maximizing the spectral efficiency, and (ii) ensuring that the load across BSs is balanced to improve the utilization efficiency and preempt congestion events. While there are many works that try to pursue a good tradeoff between them, most of them are relatively simplified, not taking into account key features of future networks.

Firstly, most existing studies only consider homogeneous traffic profiles. For example, [6, 10, 65] assume that all flows generated by a UE are "best-effort" (i.e., elastic). However, as highlighted in Chapter 1, modern and future networks will have to deal with high traffic differentiation, with certain flows being able to require specific, *dedicated* resources (i.e., dedicated). Such dedicated flows do not share BS resources like best-effort ones, are subject to admission control, and sensitive to different performance metrics by suggesting revision of current user association algorithms. Secondly, while the majority of related studies only consider downlink traffic, uplink traffic is becoming of utmost importance too. The asymmetric transmit powers between the UEs and different BSs differentiate the DL and UL physical data rates significantly. This suggests that associating a UE with the BS that offers the highest DL SINR, may lead to subpar UL performance or require high UE transmission power and thus high energy wastage. What is more, the traffic load on the DL and UL may vary significantly, due to the asymmetric traffic applications. For instance, when a user is browsing he consumes resources mostly from the downlink, when uploading a video from the uplink, or when playing an online interactive video game from both downlink and uplink.

Additionally, most of the related work in the literature towards developing various user association rules usually require some centralized knowledge [18, 88–90]. These need a controller entity that governs the BSs and the UEs with access to all the necessary information. However, depending on the operator capabilities such an implementation may not be applicable.

Additionally, even when it is applicable, it may (a) require excessive message overhead and computational complexity that increase exponentially in the network size, as well as (b) allow only for slow adaptation on the queuing statistics at relatively long timescales, since such a controller is usually implemented in a server deep in the core network. Thus, to avoid relying on a centralized controller, current systems aim on distributed implementations by highlighting the importance of "device centric" user association rules.

To this end, in this Chapter we revisit the problem of user association in a more complex setup. We adopt the basic α-fair objective function and methodology proposed in [10] as our starting point, and extend the framework considerably to include the above-mentioned challenges. Specifically, our contributions can be summarized as follows:

1) We introduce (i) dedicated flows along with a different scheduling discipline, (ii) the asymmetric UL traffic performance, (iii) the ability for both Split/Joint UL/DL association into the system model.

2) We prove that a "device centric" user association rule can still be derived for the complete framework. Interestingly, this rule when considering multiple objectives resembles a (weighted) harmonic or arithmetic mean of the individual association rules, depending on the whether Split or Joint association is applied.

3) We further investigate the complex tradeoffs involved *quantitatively* to provide some initial insights and guidelines about user-association policies in future HetNets.

The remainder of the Chapter is organized as follows: Section 2.2 outlines our system model along with the considered scheduling disciplines. The proposed framework for the optimal user-association is described in Section 2.3. Section 2.4 presents some simulation results, and Section 2.5 concludes the Chapter.

2.2 System Model and Assumptions

In the following, we describe our assumptions related to the traffic arrival model (Assumptions A.1-A.3) and the radio access network (Assumptions B.1-B.10).

We use a similar problem setup as the one used in a number of related works [6,10,42,67], and extend it accordingly. To keep notation consistent, for all variables considered a first superscript "D" and "U" refers to downlink (DL) and uplink (UL) traffic, respectively. A second superscript "b" or "d" refers to best-effort and dedicated traffic, respectively. For brevity, in the following *we present most notation and assumptions in terms of downlink traffic only, assuming that the uplink case and notation is symmetric.* Specific differences will be elaborated, where necessary. In Table 2.1, we summarize some useful notation we use throughout the chapter as well as throughout the dissertation.

(A.1 - Traffic arrival rates) Traffic at location $x \in \mathcal{L}$ consists of file (or more generally *flow*) requests arriving according to an inhomogeneous Poisson point process with arrival rate per unit area $\lambda(x)$[1]. This inhomogeneity facilitates the creation of "hotspot" areas. Each new arriving request is for a *downlink (DL)* flow, with probability z^D, or *uplink (UL)* flow with probability $z^U = 1 - z^D$. Each DL (or UL) flow can furher be a *best-effort* flow (e.g., file

[1]Without loss of generality, we do not distinguish between users at location x, as we assume that all users/flows related to location x are treated similarly.

Table 2.1: Notation

Variable	Best-Effort Flows		Dedicated Flows	
	Downlink	Uplink	Downlink	Uplink
Flow type superscript	D,b	U,b	D,d	U,d
Flow type probability	$z^D \cdot z^b$	$z^U \cdot z^b$	$z^D \cdot z^d$	$z^U \cdot z^d$
Devoted bandwidth for BS i	$w_i \cdot \zeta_i \cdot \xi_i^D$	$w_i(1-\zeta_i) \cdot \xi_i^U$	$w_i \cdot \zeta_i(1-\xi_i^D)$	$w_i(1-\zeta_i)(1-\xi_i^U)$
Traffic arrival rate at x	$\lambda^{D,b}(x)$	$\lambda^{U,b}(x)$	$\lambda^{D,d}(x)$	$\lambda^{U,d}(x)$
Max. rate | servers at x of BS i	$c_i^{D,b}(x)$	$c_i^{U,b}(x)$	$k_i^D(x)$	$k_i^U(x)$
System load at x of BS i	$\rho^{D,b}(x)$	$\rho^{U,b}(x)$	$\rho^{D,d}(x)$	$\rho^{U,d}(x)$
LB degree parameter $\in [0,\infty)$	$\alpha^{D,b}$	$\alpha^{U,b}$	$\alpha^{D,d}$	$\alpha^{U,d}$
Total load of the i-th BS	$\rho^{D,b}$	$\rho^{U,b}$	$\rho^{D,d}$	$\rho^{U,d}$
Chance that a flow at x associate i	$p_i^{D,b}(x)$	$p_i^{U,b}(x)$	$p_i^{D,d}(x)$	$p_i^{U,d}(x)$
Flow size (bits) | duration (sec) a	$1/\mu^{D,b}$	$1/\mu^{U,b}$	$1/\mu^{D,d}$	$1/\mu^{U,d}$
Flow demand (bps)	-	-	B^D	B^U
Capacity of BH link j	$C_h^D(j)$	$C_h^U(j)$	-	-
Congestion indicator at BH link j	$\mathcal{I}^D(j)$	$\mathcal{I}^U(j)$	-	-

download) with probability z^b, or *dedicated* flow (e.g., a VoIP call), with probability $z^d = 1 - z^b$. z^D and z^b are input parameters that depend on the traffic mix.

Using a Poisson splitting argument [14], it follows that the above gives rise to 4 independent, Poisson flow arrival processes with respective rates

$$\lambda^{D,b}(x) = z^D \cdot z^b \cdot \lambda(x), \quad \lambda^{D,d}(x) = z^D \cdot z^d \cdot \lambda(x) \tag{2.1}$$

$$\lambda^{U,b}(x) = z^U \cdot z^b \cdot \lambda(x), \quad \lambda^{U,d}(x) = z^U \cdot z^d \cdot \lambda(x), \tag{2.2}$$

($\lambda^{D,b}(x)$ for the downlink best-effort flows, $\lambda^{U,b}(x)$ for the uplink best-effort flows, etc.).

(A.2 - Best effort flow characteristics) Each *best-effort* flow is associated with a *flow-size* (in bits) drawn from a generic distribution with mean $1/\mu^{D,b}$. [2]

(A.3 - Dedicated flow characteristics) Each *dedicated* flow has a *required data-rate* (in bits per second) that is drawn from a generic distribution with mean B^D. This rate must be guaranteed by the network throughout the flow's duration. This duration (in seconds) is another, independent random variable with mean $1/\mu^{D,d}$.

We now turn our attention to the radio access network.

(B.1 - Access network topology) We assume an area $\mathcal{L} \subset \mathbb{R}^2$ served by a set of base stations $\mathcal{B}$, that are either MBSs or SCs. These together constitute the radio access network.

(B.2 - DL resources) Each BS $i \in \mathcal{B}$ is associated with a total bandwidth w_i. Out of the total bandwidth we perform two splits. We firstly introduce the parameter $0 < \zeta_i < 1$. $w_i \cdot \zeta_i$ is the bandwidth resources allocated to DL traffic and the rest $w_i \cdot (1-\zeta_i)$ to the UL. Likewise, the parameter $0 < \xi_i^D < 1$ further splits the DL resources. Specifically, $w_i \cdot \zeta_i \cdot \xi_i^D$ are the bandwidth

[2] Note that one can model heterogeneous flow characteristics across locations by considering different $1/\mu^{D,b}(x)$ at different x. This will come at the cost of the BS broadcast message increase (the number of broadcast messages, as discussed in Theorem 3.1, will increase from 2 to 4).

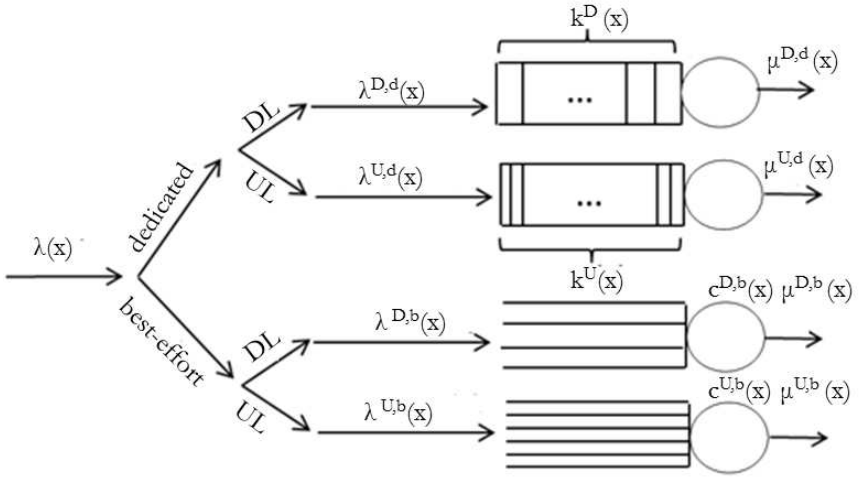


Figure 2.1: Processor sharing and k-loss queuing systems for (i) downlink dedicated, (ii) downlink best-effort, (iii) uplink dedicated, and (iv) uplink best-effort flows.

resources for DL best effort traffic whereas $w_i \cdot \zeta_i \cdot (1 - \xi_i^D)$ for the DL dedicated traffic. To simplify discussion, in this Chapter we assume that these splits are fixed. However, later in Chapter 4 we analytically show how one can optimize such bandwidth resource allocations.

(B.3 - DL physical data rate) BS i can deliver a *maximum* physical data transmission rate of $c_i^D(x)$ to a user at location x, in absence of any other flows served, which is given by the Shannon capacity[3]

$$c_i^D(x) = (\zeta_i \cdot w_i) \cdot \log_2(1 + \text{SINR}_i(x)), \tag{2.3}$$

where $\text{SINR}_i(x) = \frac{G_i(x)P_i}{\sum_{j \neq i} G_j(x)P_j + N_0}$ and P_i the transmit power of the BS. N_0 is the noise power, and $G_i(x)$ represents the path loss and shadowing effects between the i-th BS and the UE located at x (as well as antenna and coding gains, etc.)[4]. We assume that effects of fast fading are filtered out. Our model assumes that the total intercell interference at location x is static, and considered as another noise source, as is previously considered in most aforementioned works [6, 10].

Note that, due to B.2 the effective capacity for DL best effort flows will only be $c_i^{D,b}(x) = \xi_i^D \cdot c_i^D(x)$ whereas for DL dedicated only $c_i^{D,d}(x) = (1 - \xi_i^D) \cdot c_i^D(x)$.

The next 4 points (B.4-B.7) describe the scheduling and performance model for best effort traffic only. We return to dedicated traffic in (B.8-B.9).

(B.4 - Best effort load density) We introduce the *load density* for best effort flows, at different locations x,

$$\rho_i^{D,b}(x) = \frac{\lambda^{D,b}(x)}{\mu^{D,b} c_i^{D,b}(x)}, \tag{2.4}$$

[3]We use Shannon capacity for clarity of presentation. However, our approach could be easily adapted to include modulation and coding schemes (MCS). Furthermore, capacity improving technologies, e.g., the use of MIMO, and modifications to this capacity formula are othogonal to our framework.

[4]In the case of UL, we assume that the Tx power of each user is P^{UE}, and slightly abuse notation for SINR, G, etc., as these don't play a major role in the remaining discussion.

which is the contribution of location x to the total load of a BS i, when location x is associated to BS i.

(B.5 - Best effort load) Each location x is associated with routing probabilities $p_i^{D,b}(x) \in [0,1]$, which are the probabilities that best effort DL flows generated for users at location x get associated with (i.e., are served by) BS i. We can thus define the *total best effort load* $\rho_i^{D,b}$ for BS i as

$$\rho_i{}^{D,b} = \int_{\mathcal{L}} p_i^{D,b}(x)\rho_i^{D,b}(x)dx. \tag{2.5}$$

Similarly to [10, 19], we are interested in the *flow-level dynamics* of this system, and model the service of DL best-effort flows at each BS as a queueing system with load $\rho_i^{D,b}$ shown in Fig. 2.1. Finally, since we are interested in the aggregation of all flows at BS level (i.e., all flows from all locations x assosicated to BS i), even if flow arrivals at each location are not Poisson (as in A.1), the Palm-Khintchine theorem [14] suggests that Poisson assumption could be a good approximation for the input traffic to a BS.

(B.6 - Best effort scheduling) Proportionally fair scheduling is often implemented in modern networks for best-effort flows, due to its good fairness and spectral efficiency properties [69]. This can be modeled as an M/G/1 multi-class processor sharing (PS) system (see, e.g., [6, 10, 19]). It is multi-class, because each flow might get different rates for similarly allocated resources, due to different channel quality and MCS at x. Channel-based scheduling could also be included in the model and can be accounted for using a multiplicative factor in the average service rate [91].

(B.7 - Performance for best effort flows) The stationary number of flows in BS i is equal to $E[N_i] = \frac{\rho_i^{D,b}}{1-\rho_i^{D,b}}$ [14]. Hence, minimizing $\rho_i^{D,b}$ minimizes $E[N_i]$, and by Little's law it also minimizes the per-flow delay for that base station [14]. Also, the throughput for a flow at location x is $c_i^{D,b}(x)(1-\rho_i^{D,b})$. This observation is important to understand how the user's physical data rate $c_i^{D,b}(x)$ (related to users at location x only) and the BS load $\rho_i^{D,b}$ (related to *all* users associated with BS i) affect the optimal association rule.

(B.8 - Dedicated traffic load density) Unlike best-effort flows which are elastic, dedicated flows are subject to admission control, since they require some resources for exclusive usage in order to be accepted in the system. Specifically, let $c_i{}^{D,d}(x)$ denote the maximum offered rate to users at location x corresponding to dedicated flows only (referred to $(1-\xi_i^D)$ - see B.2 above). If each flow at x demands, on average, a rate of B^D (see A.3), then at most $k_i^D(x) = \frac{c_i{}^{D,d}(x)}{B^D}$ dedicated flows from x could be served in parallel by BS i (assuming again *no other flows in the system*), and any additional flows would be rejected[5]. Similarly to the best effort case (B.4), we can define a system load density for dedicated traffic at x

$$\rho_i{}^{D,d}(x) = \frac{\lambda^{D,d}(x)}{\mu^{D,d}k_i^D(x)} = \frac{\lambda^{D,d}(x)\cdot B^D}{\mu^{D,d}\cdot c_i{}^{D,d}(x)}. \tag{2.6}$$

(B.9 - Dedicated traffic performance) Given the above heterogeneous blocking model for dedicated flows, we can approximate the allocation of BS i dedicated resources with an M/G/k/k

[5] In fact, since the rate requirement for each flow is a random variable, using its mean B^D in the denominator yields a lower bound for $k_i^D(x)$ (by Jensen's inequality), which can be used as a conservative estimate.

(or k-loss) system, where the total load $\rho_i^{D,d}$ can be calculated as in (B.5) and Eq. (2.5), using the density of Eq. (2.6) and corresponding routing probability $p_i^{D,d}(x)$ for dedicated flows (see also Fig. 2.1). It is known that for M/G/k/k systems, minimizing $\rho_i^{D,d}$ is equivalent to minimizing the blocking probability, given from the Erlang-B formula, for new flows [14]. This observation is important to understand that a similar tradeoff (as in B.7) exists between choosing a BS at x that maximizes $k_i^D(x)$ (related only to flow and channel characteristics at x) and choosing a BS whose *total* load $\rho_i^{D,d}$ (related to *all* users attached to BS i).

(B.10 - UL/DL association split) We investigate two scenarios, depending on the whether a UE is allowed to be attached to different BSs for its DL and UL traffic [77]:

Split UL/DL: Each UE can be associated to different BSs for its DL and UL traffic. This allows one to optimize UL and DL performance independently.

Joint UL/DL: Each UE must be associated with the same BS for both UL and DL traffic. This is the standard practice in current networks.

2.3 Traffic-Steering User Association: Problem and Optimal Rules.

We remind to the reader that based on our system model, the association policy consists in finding appropriate values for the routing probabilities $p_i^{l,t}(x),\ l \in \{D,U\},\ t \in \{b,d\}$, for DL and UL, best-effort and dedicated traffic, respectively (defined earlier in assumption B.5 and B.9). That is, for each location x, we would like to optimally choose to which BS i to route different flow types generated from (UL) or destined at (DL) users in x[6]. As it will turn out later, these probabilities can take the same or different values, depending on the scenario.

Our goal for this association problem is threefold: (i) ensure that the capacity of no BS is exceeded; (ii) achieve a good tradeoff between spectral efficiency, user QoS and load balancing, (iii) investigate how UL/DL association split impacts the optimal rule derivation and the performance benefits of split UL/DL.

To that end, in Section 2.3.1 we explicitly sketch an extended version of the convex α-fair objective function defined on a convex feasible set and then, we formulate the arising Optimization Problem 1. In Sections 2.3.2, 2.3.3 we separately solve this problem upon Split and Joint UL/DL association and illustrate our novel user association rules. Section 2.3.4 completes our framework by illustrating our proposed distributed implementation.

2.3.1 Feasible Set, Objective Function and Optimization Problem 1

We define the feasible region for the aforementioned routing probabilities, by requiring that no BS capacity being exceeded.

Definition 1. *(Feasibility) Let $l \in \{U,D\}, t \in \{b,d\}$, and let ϵ be an arbitrarily small positive*

[6]The use of a probabilistic association rule simplifies solving the problem. As it will turn out, the optimal values will be either 0 or 1 (deterministic).

constant. The set $f^{l,t}$ *of feasible BS loads* $\rho^{\mathbf{l,t}} = (\rho_1^{l,t}, \rho_2^{l,t}, \ldots, \rho_{\|\mathcal{B}\|}^{l,t})$ *is*

$$\begin{aligned} f^{l,t} = \Big\{ \rho^{\mathbf{l,t}} \mid \rho_i^{l,t} = \int_{\mathcal{L}} p_i^{l,t}(x)\rho_i^{l,t}(x)dx, \\ 0 \leq \rho_i^{l,t} \leq 1-\epsilon, \\ \sum_{i\in\mathcal{B}} p_i^{l,t}(x) = 1, \\ 0 \leq p_i^{l,t}(x) \leq 1, \forall i \in \mathcal{B}, \forall x \in \mathcal{L} \Big\}. \end{aligned} \tag{2.7}$$

Lemma 1. *The feasible sets* $f^{D,b}, f^{D,d}, f^{U,b}, f^{U,d}$ *as well as the* $[f^{D,b}; f^{D,d}]$, $[f^{U,b}; f^{U,d}]$, $[f^{D,b}; f^{U,b}]$, $\mathcal{F} = [f^{D,b}; f^{D,d}; f^{U,b}; f^{U,d}]$, *are convex.*

Proof. The proof for the feasible set $f^{D,b}$ is presented in [10]. It can be easily adapted for the other cases, too. □

Following [10] we extend the proposed objective to also include the DL dedicated traffic (see A.1, A.3). We introduce parameter $\theta \in [0,1]$ that helps the operator weigh the importance of DL best effort versus DL dedicated traffic performance. $\alpha^{D,b}$ controls the amount of load balancing desired in the DL best-effort resources, and $\alpha^{D,d}$ in the DL dedicated. Lets denote $\alpha^{\mathbf{D}} = [\alpha^{D,b}; \alpha^{D,d}]$ and $\rho^{\mathbf{D}} = [\rho^{D,b}; \rho^{D,d}]$.

Definition 2. *(Objective function for DL) Our objective is*

$$\phi_{\alpha^{\mathbf{D}}}(\rho^{\mathbf{D}}) = \sum_{i\in\mathcal{B}} \theta \cdot \frac{(1-\rho_i^{D,b})^{1-\alpha^{D,b}}}{\alpha^{D,b}-1} + (1-\theta)\cdot \frac{(1-\rho_i^{D,d})^{1-\alpha^{D,d}}}{\alpha^{D,d}-1}, \text{ if } \alpha^{D,d}, \alpha^{U} \neq 1. \tag{2.8}$$

If $\alpha^{D,b}$ *(or,* $\alpha^{D,d}$*) is equal to* 1*, the respective fraction must be replaced with* $\log(1-\rho_i^{D,b})^{-1}$ *(or,* $\log(1-\rho_i^{D,d})^{-1}$*).*

As a final step, we want to further extend this objective to also capture the UL traffic performance and thus, we introduce $0 \leq \tau \leq 1$ to trade it off with DL. Specifically, as $\tau \to 0$ the weight turns to DL traffic performance whereas as $\tau \to 0$ to UL. Lets assume that $\alpha = [\alpha^{D,b}; \alpha^{D,d}; \alpha^{U,b}; \alpha^{U,d}]$ and $\rho = [\rho^{D,b}; \rho^{D,d}; \rho^{U,b}; \rho^{U,d}]$.

Definition 3. *(Objective function for DL and UL) The objective that jointly considers DL and UL performance follows*

$$\phi_{\alpha}(\rho) = \tau \cdot \phi_{\alpha^{\mathbf{D}}}(\rho^{\mathbf{D}}) + (1-\tau)\cdot \phi_{\alpha^{\mathbf{U}}}(\rho^{\mathbf{U}}). \tag{2.9}$$

Lemma 2. *The objective function* $\phi_\alpha(\rho)$ *is convex.*

Proof. Since $\phi_\alpha(\rho)$ is a weighted sum of four convex functions [10], convexity is preseved [92]. □

Definition 4. *(Optimization Problem 1) The UL/DL user association problem can be expressed as*

$$\begin{aligned} &\underset{\rho}{\text{minimize}}\{\phi_\alpha(\rho) | \rho \in \mathcal{F}\}, \\ &\textit{subject to } p_i^D(x) = p_i^{D,b}(x) = p_i^{D,d}(x) \textit{ and } p_i^U(x) = p_i^{U,b}(x) = p_i^{U,d}(x), \\ &\textit{(dependent) subject to } p_i^D(x) = p_i^U(x) \textit{ iff joint UL/DL association}, \forall x \in \mathcal{L}. \end{aligned} \tag{2.10}$$

The objective function to minimize is the α-fair cost function introduced in Eq. 2.9. The first constraint ensures that all DL best-effort flows as well as all DL dedicated shall be originated from the same BS, as most current cellular networks require. Similarly for UL. (This constraint can be easily relaxed to reflect additional flexibilities, as we will show later.) On the other hand, the DL and UL traffic is allowed to be associated with different BSs only for the split scenario without additional constraints, as the second (dependent) constraint suggest. This constraint, for joint uplink and downlink association, shall only be required in joint UL/DL association scenarios (see B.10).

Lemma 3. *Optimization Problem 1 is a convex minimization problem since its objective function is convex in the convex set $\mathcal{F}$.*

In the next sections we discuss how one can derive "device-centric" user association rules, by minimizing Optimization Problem 1.

2.3.2 Optimal Rules for Split UL/DL

We start with the simpler *Split UL/DL association* scenario where the *dependent* constraint defined in Optimization Problem 1 is relaxed. Thus, in this scenario a UE can be associated with different BSs for its DL and UL traffic (see B.10), i.e., $p_i^D(x)$ and $p_i^U(x)$ are allowed to take different values.

It is obvious that in this case the UL and the DL problem *decouple*. Specifically, one can separately optimize $\phi_{\alpha^{\mathbf{D}}}(\rho^{\mathbf{D}})$ and $\phi_{\alpha^{\mathbf{U}}}(\rho^{\mathbf{U}})$. In the reminder of the section we focus on the DL (the UL case is symmetric) and to simplify notation we drop the "D" superscript.

Theorem 3.1. *(Split UL/DL User Association Rule) If $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ denotes the optimal load vector, the optimal "device-centric" user association rule at x is*

$$i(x) = \arg\max_{i \in \mathcal{B}} \left(\underbrace{c_i(x)}_{\text{user knowledge}} \cdot \overbrace{P_i}^{\text{BS broadcast message}} \right) \tag{2.11}$$

where each BS $i \in \mathcal{B}$ shall broadcast the following weighted harmonic mean (of individual rules) formula

$$P_i = \frac{\left(1-\rho_i^{*b}\right)^{\alpha^b} \cdot \left(1-\rho_i^{*d}\right)^{\alpha^d}}{e^b \cdot \left(1-\rho_i^{*d}\right)^{\alpha^d} + e^d \cdot \left(1-\rho_i^{*b}\right)^{\alpha^b}}.$$

Note that, $e^b = \frac{\theta z^D z^b}{\mu^b \zeta_i \xi_i^D}$ as well as $e^d = \frac{(1-\theta) z^D z^d B^D}{\mu^d \zeta_i (1-\xi_i^D)}$ optimally weight the corresponding individual association rules depending on the traffic statistics.

Proof. We prove that the above user association rule Eq. 2.11 indeed minimizes the objective defined in Eq. 2.8, that is mainly the DL part of Eq. 2.9. As discussed there this is a convex optimization problem. Hence, it is adequate to check the following condition for optimality

$$\langle \nabla \phi(\rho^*), \Delta\rho^* \rangle \geq 0 \tag{2.12}$$

for all $\rho \in f$, where $\Delta\rho^* = \rho - \rho^*$. Let $p(x)$ and $p^*(x)$ be the associated routing probability vectors for ρ and ρ^*, respectively. Using the deterministic cell coverage generated by Eq. 2.11,

the optimal association rule is given by:

$$p_i^*(x) = \mathbf{1}\left\{ i = \arg\max_{i\in\mathcal{B}} c_i^D(x) \frac{\left(1-\rho_i^{*b}\right)^{\alpha^b} \cdot \left(1-\rho_i^{*d}\right)^{\alpha^d}}{e^b \cdot \left(1-\rho_i^{*d}\right)^{\alpha^d} + e^d \cdot \left(1-\rho_i^{*b}\right)^{\alpha^b}} \right\}. \tag{2.13}$$

Then the inner product in Eq. 2.12 can be written as:

$$\begin{aligned}
&\langle \nabla\phi\left(\rho^*\right), \Delta\rho^* \rangle = \sum_{z=\{b,d\}} \frac{\partial\phi}{\partial\rho_z}\left(\rho^*\right)\left(\rho_z - \rho_z^*\right) \\
&= \frac{\partial\phi}{\partial\rho^b}(\rho^*)(\rho^b - \rho^{*b}) + \frac{\partial\phi}{\partial\rho^d}(\rho^*)(\rho^d - \rho^{*d}) \\
&= \theta \sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^b)^{\alpha^b}}(\rho_i^b - \rho_i^{*b}) + (1-\theta)\sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^d)^{\alpha^d}}(\rho_i^d - \rho_i^{d*}) \\
&= \sum_{i\in\mathcal{B}} \frac{\theta\int_L \rho_i^b(x)(p_i(x) - p_i^*(x))dx}{(1-\rho_i^b)^{\alpha^b}} + \frac{(1-\theta)\int_L \rho_i^d(x)(p_i(x) - p_i^*(x))dx}{(1-\rho_i^d)^{\alpha^d}} \\
&= \int_L \lambda(x) \sum_{i\in\mathcal{B}} (p_i(x) - p_i^*(x)) \left(\frac{e^b(1-\rho_i^{*d})^{\alpha^d} + e^d(1-\rho_i^{*b})^{\alpha^b}}{c_i(x)\cdot(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}} \right) dx.
\end{aligned} \tag{2.14}$$

Note that,

$$\sum_{i\in\mathcal{B}} p_i(x) \frac{e^b(1-\rho_i^{*d})^{\alpha^d} + e^d(1-\rho_i^{*b})^{\alpha^b}}{c_i(x)\cdot(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}} \geq \sum_{i\in\mathcal{B}} p_i^*(x) \frac{e^b(1-\rho_i^{*d})^{\alpha^d} + e^d(1-\rho_i^{*b})^{\alpha^b}}{c_i(x)\cdot(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}} \tag{2.15}$$

holds because $p^*(x)$ in Eq. 2.13 is an indicator for the minimizer of $\frac{e^b(1-\rho_i^{*d})^{\alpha^d} + e^d(1-\rho_i^{*b})^{\alpha^b}}{c_i(x)\cdot(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}}$. Hence, the inner product condition defined in Eq. 2.12 holds. □

While θ linearly weights the best effort versus dedicated flow performance, the impact of $\alpha^{D,b}, \alpha^{D,d}$ is not obvious. We now discuss their impact on the system performance and refer to [10], [93] for the respective proofs.

- *Spectral Efficiency Optimization:* $\alpha^{D,d} = 0$ maximizes the average physical rate for best-effort flows (defined in B.3), whereas $\alpha^{D,d} = 0$ maximizes the average dedicated servers for dedicated flows (defined in B.8). Obviously, these optimize the user $SINR$ and spectral efficiency.

- *Optimizing related QoS metrics:* when $\alpha^{D,b} = 1$ the derived rules tend to maximize the average user throughput. If $\alpha^{D,b} = 2$ the per-flow delay is minimized since the objective for best effort flows corresponds to the delay of an M/G/1/PS system. If $\alpha^{D,d} = 1$ the corresponding optimal rule becomes equivalent to the average *idle* dedicated servers in a M/G/k/k system, and the actual blocking probability is minimized.

- *Load-Balancing Efficiency Optimization:* as $\alpha^{D,b} \to \infty$, we minimize the maximum BS utilization, i.e., load balancing between the $\rho^{D,b}$ is achieved. Similar for $\alpha^{D,d}$ and $\rho^{D,d}$'s. Note that, the point of $\alpha^{D,b}$ that all BS best-effort utilizations are equalized might be different from the one for dedicated, depending on the respective traffic statistics.

In the case of split UL/DL association, the above analysis can be applied *separately* on UL and DL traffic, and optimize UL and DL associations independently.

2.3.3 Optimal Rules for Joint UL/DL

Traditional cellular networks suggest that a UE should be connected to a single BS for both UL and DL traffic. In this section we derive the optimal rules for such a joint UL/DL association scenario, where the *dependent* constraint $p_i^D(x) = p_i^U(x)$ discussed in Optimization Problem 1 shall be satisfied. The unique BS a user shall associeated with in such a scenario is portrayed from the following theorem.

Theorem 3.2. *[Joint UL/DL User Association Rule] The optimal user-association rule at x for joint UL/DL association turns out to be*

$$i(x) = \arg\max_{i\in\mathcal{B}} \frac{1}{\frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)}} \tag{2.16}$$

where each BS $i \in \mathcal{B}$ shall broadcast

$$P_i^D = \frac{\sum\limits_{t\in\{b,d\}} e^{D,t} \prod\limits_{c\in\Omega\neq(D,t)} ((1-\rho^{*c})^{\alpha^c})}{\prod\limits_{c\in\Omega}((1-\rho^{*c})^{\alpha^c})} \textit{(similar in UL)},$$

and $\Omega \in \{(D,d),(D,b),(U,d),(U,b)\}$. This formula ends up maximizing the weighted harmonic mean formula of the individual rules, with corresponding weighting factors $e^{D,b} = \tau\frac{\theta^D z^D z^b}{\mu^{D,b}\zeta\xi^D}$, $e^{D,d} = \tau\frac{(1-\theta^D)z^D z^d B^D}{\mu^{D,d}\zeta(1-\xi^D)}$, $e^{U,b} = (1-\tau)\frac{\theta^U z^U z^b B^U}{\mu^{U,b}(1-\zeta)\xi^U}$ and $e^{U,d} = (1-\tau)\frac{(1-\theta^U)z^U z^d B^U}{\mu^{U,d}(1-\zeta)(1-\xi^U)}$.

Proof. The proof follows similar steps with the one for Split UL/DL scenario, where now one has to require also $p_i^D(x) = p_i^U(x)$. Then, the inner product becomes:

$$\begin{aligned}
&\langle \nabla\phi(\rho^*), \Delta\rho^*\rangle = \sum_{z=\{b,d\}} \frac{\partial\phi}{\partial\rho_z}(\rho^*)(\rho_z - \rho_z^*) \\
&= \theta^D \sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^{D,b})^{\alpha^{D,b}}}(\rho_i^{D,b} - \rho_i^{*D,b}) + (1-\theta^D)\sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^{D,d})^{\alpha^{D,d}}}(\rho_i^{D,d} - \rho_i^{D,d*}) \\
&+ \theta^U \sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^{U,b})^{\alpha^{U,b}}}(\rho_i^{U,b} - \rho_i^{*U,b}) + (1-\theta^U)\sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^{U,d})^{\alpha^{U,d}}}(\rho_i^{U,b} - \rho_i^{U,d*}) \\
&= \int_L \lambda(x) \sum_{i\in\mathcal{B}} (p_i(x) - p_i^*(x)) \left(\frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)}\right) dx \geq 0,
\end{aligned} \tag{2.17}$$

due to the corresponding minimizer $p_i(x)$ derived from Eq. 2.16. □

Remark 1. All the derived rules are "device centric" i.e., the user is able to select where to associate *based on own measurements (e.g. SINR) and BS-transmitted information.* Such broadcast quantities can be easily integrated through the newly proposed Access Network Discovery and Selection Function (ANDSF) mechanism or in the absolute/dedicated priority list mechanisms of LTE [94]. In both Split and Joint association scenarios, we proved that each BS only needs to broadcast two different messages: one value related to its DL dynamics (P_i^D), and another one related to the UL (P_i^U). This is inline with user association in current LTE systems, where user association depends on "device centric" information (e.g. SINR measurements) but also BS-transmitted information (e.g. priority lists of BSs to monitor). These rules:

- are *scalable* (constant amount of the BS broadcast messages irrespective of the number of users),
- are *simple* (constant complexity of the rule with respect to the number of BSs), and
- offer *flexible performance* (defined from α values).

Remark 2. The optimal rule derived in Eq. 2.16 suggests that in the *joint UL/DL* scenario associated with objectives that potentially conflict with each other (due to the different flow type performances), it is optimal to associate a user with the BS that maximizes a weighted version of the *harmonic mean* of the individual association rules when considering each objective alone. To better understand this, we focus on a simple scenario with only DL and UL best-effort traffic. And assume the following BS options for a user: (BS A) offers 50Mbps DL and only 1Mbps UL; (BS B) 200Mbps DL and 0.5Mbps UL; (BS C) 20Mbps DL and 5Mbps UL. If we care about UL and DL performance equally (i.e. $\tau = 0.5$), one might assume that the BS that maximizes the arithmetic mean (or arithmetic sum) of rates would be a fair choice (i.e. BS B). However, this would lead to rather poor UL performance. Maximizing the harmonic mean would lead to choosing (BS C) instead[7]. Additionally, note that in the case of *split UL/DL*, covered in Section 2.3.2, where each user is free to be associated with two different BSs for the DL and UL traffic offloading, DL traffic would be associated with (BS B), and UL traffic with (BS C) by maximizing the arithmetic mean (or, sum) of their throughputs[8]. These simple examples intuitively explain how split UL/DL impacts the user association policies, by allowing to independently optimize each objective. This also demonstrates why UL/DL split may perform considerably better than the joint association. We will further explore this in the Simulations.

Remark 3. We finally underline that, the "formula" of harmonic or arithmetic mean maximization further allows to add more dimensions in our setup and *flexibly* derive the optimal rules without any analytical calculations. For instance, consider a more modern offloading technique, where different downlink, or uplink, flow types are able to be offloaded to different BSs (e.g., per flow/QCI offloading) with conflicting aims. Using our model we can consider an additional respective α-function for each flow type, and either analytically or flexibly, optimize the complete objective as showed earlier. As an example, one can relax the constraints $p_i^{D,b}(x) = p_i^{D,d}(x)$ and $p_i^{U,b}(x) = p_i^{U,d}(x)$ to allow a UE offload its DL (or, UL) flows to different BSs. Then we can directly derive the optimal association rules using the arithmetic formula of the individual rules.

2.3.4 Distributed Implementation Framework

In Theorems 3.1 and 3.2, we derived different "device centric" user association rules for different scenarios, where each UE is able to decide where to associate by itself in both downlink and uplink. We now elaborate on our proposed implementation framework where such rules apply, and direct the network towards the optimal BS loads.

Following [10], we sketch a distributed implementation that is applied iteratively, adapts to spatial traffic loads, and mainly involves two parts: the *user* and *base station* tier.

[7]While this simple example captures the main principle, the actual rule is more complex, as it weighs each objective with the complex factor e^l.

[8]The usage of harmonic mean and arithmetic mean/sum appears in a number of physical examples, such as in the calculation of the total resistance in circuits where all resistances are set in series or in parallel.

Table 2.2: Simulation Parameters

Parameter	Variable	Value
Transm. Power of eNB/ SC/ UE	$P_{eNB}/P_{SC}/P_{UE}$	43/24/12 dBm
BS Bandwidth for DL, UL	w/W	10/10 MHz
Noise Power Density	N_0	-174 dBm/Hz
Splitting parameter for DL, UL	ζ_i^D, ζ_i^U	0.5/0.5
Average DL/UL flow sizes	$\frac{1}{\mu^{D,b}}/\frac{1}{\mu^{U,b}}$	100/20 Kbytes
Average DL/UL flow demands	$B^D(x)/B^U(x)$	512, 128 kbps
Different flow ratios	z^b, z^D	0.3,0.6

At the k-th period, each user at some location x receives from each BS in range the two corresponding values P_i^D, P_i^U in order to apply the derived association rule (e.g., in Eq. 2.11 the DL association is based on P_i^D and in UL on P_i^U), e.g., through broadcast control messages. Then each new flow request simply selects the BS i that maximizes the corresponding quantity. Also, at each k iteration, BSs measure their average utilizations $\rho^{(k)}$ after some required period of time (e.g., see Eq. (2.5)). Then, based on the previous BS loads $\tilde{\rho}^{(k)}$, the new BS vector $\tilde{\rho}^{(k+1)}$ needed for the broadcast control message in the next iteration would be

$$\tilde{\rho}^{(k+1)} = \beta^{(k)} \cdot \rho^{(k)} + (1 - \beta^{(k)}) \cdot \tilde{\rho}^{(k)}, \tag{2.18}$$

where $\beta^{(k)} \in [0, 1)$ is an exponential-averaging parameter. Note that, in the Split UL/DL scenario, the UL and DL loads can be independently updated, whereas in the Joint UL/DL should be updated jointly using the same $\beta^{(k)}$. This iteration converges to the globally optimal point ρ^*, requiring a simple modification to the proof found [10].

Alternatively, note that our framework could also be implemented in an SDN framework, using a centralized or hierarchical implementation, where a controller derives the optimal associations and directly sends them through the network to the UEs [93] [95].

2.4 Simulations

In this section we briefly present some numerical results and discuss related insights.

We consider a 2×2 km^2 area. Figure 2.2 shows a color-coded map of the heterogeneous traffic demand $\lambda(x)$ ($flows/hour$ per unit area) (blue implying low traffic and red high), with 2 hotspots. We assume that this area is covered by two macro BSs and eight SCs. The macro BSs that are shown with asterisks are numbered from 1-2, and the SCs that are shown with triangles are numbered from 3-10, as we can see in Fig. 2.3, Fig. 2.5. We also consider standard parameters as adopted in 3GPP [96], listed in Table 2.2[9]. If not explicitly mentioned, we assume $\theta^D = \theta^U = \tau = 0.5$, and the split UL/DL scenario as default.

We will present the impact of our proposed association rules via coverage snapshots to show how users associate in the considered network. Additionally, we will also provide values for related performance metrics that complete our study numerically.

[9] As for (i) the sizes and ratios of different flows, (ii) splitting parameters, we can use different values in order to capture different simulation scenarios, and derive similar results.

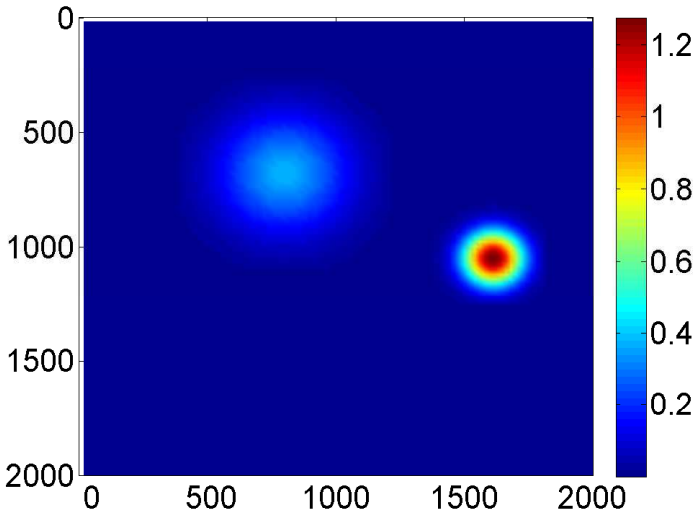
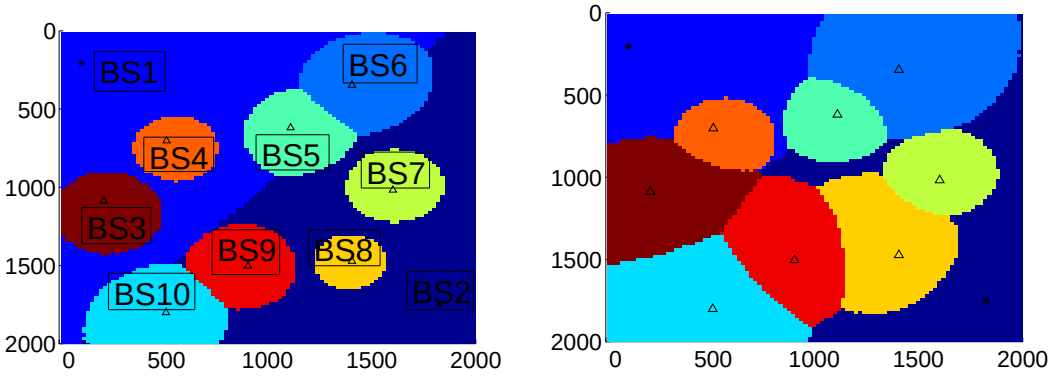


Figure 2.2: Traffic Arrival Rate (blue colour implying low traffic and red colour implying high traffic demand).

Coverage snapshots: Spectral vs. Load balancing Efficiency. Figure 2.3(a) outlines the optimal DL user-associations if $\alpha^{D,b} = \alpha^{D,d} = 0$, i.e., when *(area) spectral efficiency* is maximized. Thus, each UE at x is attached to the BS that offers the *highest DL SINR* and promises higher DL physical rate for best effort flows $c_i^{D,b}(x)$, and more "dedicated" servers $k_i^D(x)$; i.e. most of UEs are attached to macro BSs due to their high power transmission, and fewer to SCs, forming small circles around them. Consequently, macrocells are overloaded and load imbalance within the cells is sharpened (decreased $1 - MSE^{D,b}$, $1 - MSE^{D,d}$; see line 1 of Table 4.2). Note that Load Balancing (LB) efficiency is considered in terms of the mean square error (MSE) between different BS loads (normalized). However, in Fig. 2.3(b) we emphasize the *load-balancing* efficiency and set $\alpha^{D,b} = \alpha^{D,b} = 10$. Now, most SCs vastly increase their coverage area in order to offload the overloaded macro BSs (e.g., BSs 6, 8, 10); "heavily" loaded (due to the hotspots) BSs, roughly maintain the same coverage (BS 4 and 7). Thus load balancing is improved, at the cost of $E[c^{D,b}], E[k^D]$ (see line 2 of Table 4.2). For further implications of α parameters we refer the reader to [10].

Best-effort versus dedicated traffic performance. Although in the previous scenarios the best-effort- and dedicated- related traffic rules (represented from $\alpha^{D,b}, \alpha^{D,d}$) are aligned, one could ask how would two conflicting optimization objectives affect our network? The answer lays in the usage of θ, that judges which objective carries more importance. E.g., an operator has two main goals: (i) to maximize the average number of servers for "dedicated" traffic captured by $E[k^D]$ (set $\alpha^{D,d} = 0$), (ii) to better balance the utilization of best-effort resources between BSs (set $\alpha^{D,b} = 10$). As shown in Fig. 2.4, if $\theta \to 0$ $E[k^D]$ is maximized, whereas as $\theta \to 1$, $1 - MSE^{D,b}$ (DL best-effort load balancing) is optimized, and each objective comes at the price



(a) Spectral Efficiency Optimization $\alpha^{D,b} = \alpha^{D,d} = 0$.

(b) Enhanced Load Balancing $\alpha^{D,b} = \alpha^{D,d} = 10$.

Figure 2.3: Optimal user-associations (considering the tradeoff between spectral efficiency and load balancing efficiency.)

of the other.

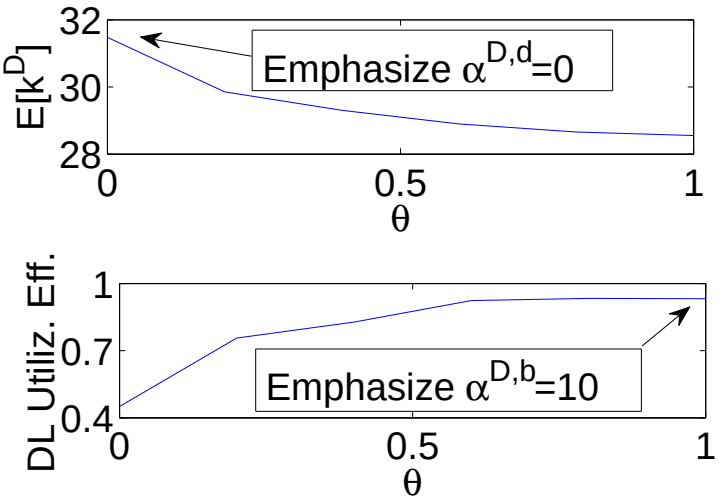


Figure 2.4: θ versus Number of Servers for dedicated flows ($E[k^D]$) and Load Balancing (or, utilization) efficiency for best-effort traffic (tradeoff between best effort vs. dedicated traffic performance).

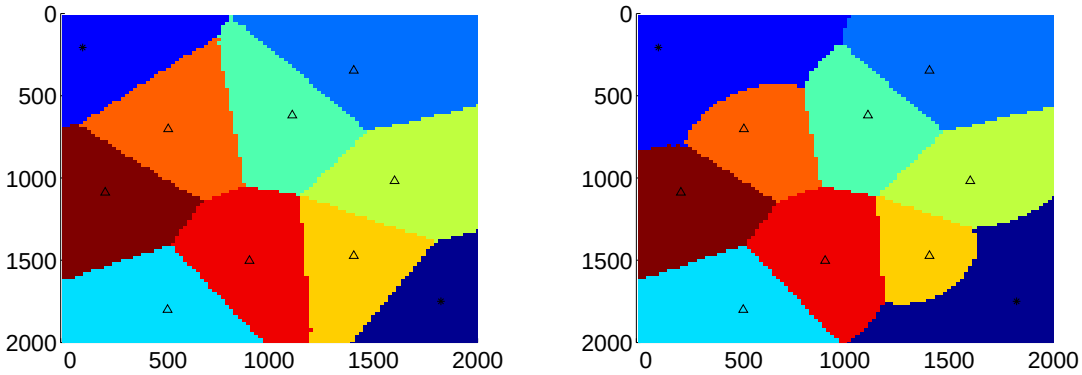
DL vs. UL traffic performance is considered in Figure 2.3(a), 2.5(a)-2.5(b), with respective numerical performance metrics in Table 2.4. The first two figures depict the DL and UL optimal associations, in case of split UL/DL, for each user at x. However, if split is not available from the operator point of view, we have to weight whether the DL or UL performance is more important while selecting a *single* BS for joint UL/DL association, using parameter τ. To that end, Figure 2.3(a) (also) outlines the optimal associations in the joint UL/DL case if the whole emphasis is on the *DL performance* ($\tau = 1$): this hurts the UL performance due to

Table 2.3: Numerical values for Physical Data Rates and Load Balancing (Figure 2.3).

	Rates and Servers		**Load Balancing**	
	$E[c^{D,b}]$ (Mbps)	$E[k^D]$	1-$MSE^{D,b}$	1-$MSE^{D,d}$
Fig. 2.3(a)	16.3	32	0.77	0.78
Fig. 2.3(b)	14.3	27	0.96	0.995

the asymmetric transmission powers of the UEs and BSs (see line 1 of Table 2.4). In Fig. 2.5(a) the emphasis is moved on the *UL performance* ($\tau = 0$), and each UE is attached to the nearest BS, in order to minimize the path loss [47] and enhance the UL performance; this hurts its DL performance though (see line 3 of Table 2.4). Finally, Fig. 2.5(b) shows the optimal coverage areas when one assigns equal importance to the UL and DL performance (i.e. $\tau = 0.5$): this moderates both DL and UL performance (line 2 of Table 2.4).

Split vs. Joint UL/DL Association. We saw that within joint UL/DL association it is impossible to achieve optimal UL/DL performance *simultaneously*. Using τ, we can trade-off which carries more importance while selecting the single BS for association, though. On the other hand, according to the Split UL/DL each UE is attached to two BSs: one that maximizes its DL, and one that maximizes its UL performance, as shown in Fig. 2.3(a), 2.5(a), Table 2.4, and implied in Remark 2.



(a) UL performance optimization $\tau = 0$.

(b) Equal emphasis of DL vs. UL $\tau = 0.5$.

Figure 2.5: Optimal user associations (considering the tradeoff between downlink (DL) versus uplink (UL) traffic performance)

Table 2.4: Numerical values for downlink and uplink performance (Figure 2.5).

	DL performance		**UL performance**	
	$E[c^{D,b}]$ (Mbps)	$E[k^D]$	$E[c^{U,b}]$ (Mbps)	$E[k^U]$
Fig. 2.3(a)	16.3	32	2.3	18
Fig. 2.5(b)	14.7	28	3	24
Fig. 2.5(a)	13.3	26	3.6	28

2.5 Conclusion

In this Chapter, we considered the user-association problem for future dense HetNets. We made a first step towards addressing various key issues that future user association schemes shall be aware of, such as: (i) traffic consisting of both best-effort and dedicated flows, and (ii) UL and DL performance, (iii) differentiation between Split and Joint UL/DL associations. To that end, novel "device centric" association rules were analytically derived within our framework, for different scenarios. We also showed that such rules end up maximizing different kinds of mean, depending on the association. Finally, initial simulation results are presented that corroborate the correctness of the framework, and reveal interesting tradeoffs.

Chapter 3

Backhaul-aware User Association Optimizations.

3.1 Introduction

In the previous Chapter we considered the user association problem, by focusing on the radio access network. More precisely, we revisited the popular α-fair objective for user association. The issue under scrutiny was to derive novel "device centric" user association rules that achieve a good tradeoff between different efficiencies by capturing the traffic differentiation and Split/ Joint UL/DL association schemes.

Nevertheless, while our proposed rules surely lead to BS loads that are supported from the access network, they perhaps will not be supported from the backhaul link (or the corresponding backhaul link path) for that BS, since they ignore potential backhaul limitations. As highlighted in Chapter 1, the chance that the backhaul link capacities will not sufficient to support the demand of the access network, is relatively high, in future HetNets. The main reasons for that include (i) the constant improvement of access network functionalities, and (ii) the fact that multiple BS might also have to share the capacity of a single backhaul link; both threatening the backhaul capacities. Due to these reasons, as well as the heterogeneous technologies that are expected to dominate in future backhaul systems (heterogeneous in terms of capacity, latency, etc.), there is an emerging call for more sophisticated, backhaul-aware, user association schemes.

To this end, in this Chapter we revisit the user association problem, by jointly considering not only the radio access but also the backhaul network performance. Our aim is to derive novel *backhaul-aware user association rules*. Specifically, our main contributions can be summarized as it follows

1) We further extend the popular α-fair objective function to include (i) (heterogeneous) backhaul capacity constraints, and (ii) backhaul topology limitations.

2) We analytically derive novel backhaul-aware association rules. We use appropriate tools mostly coming from convex optimization theory, so that (i) our rules are "device centric" (applicable in distributed implementations); (ii) maintain the desired properties required by future systems (e.g., scalability).

3) We use our framework to investigate the various tradeoffs arising in this complex association problem, and provide some initial insights and guidelines about the impact of backhaul

limitations in optimal user-association policies for future HetNets.

3.2 System Model and Assumptions

The system model in terms of traffic and access network modeling stays the same as in (A.1-A.3) and (B.1-B.10). However, we need to make some explicit assumptions regarding the backhaul network (C.1-C.4) as it follows later in the Chapter.

In order to better elucidate the considered problem at hand and without loss of generality, we focus on a simple scenario with only best-effort traffic. This will allow us, to better understand the impact of backhaul limitations. So, in the remainder of the section we drop the corresponding superscripts "b", "d" to simplify notation. We remind to the reader that to keep notation consistent, for all variables considered a first superscript "D" and "U" refers to downlink (DL) and uplink (UL) traffic, respectively and that, for brevity, in the following we present most notation and assumptions in terms of downlink traffic only, assuming that the uplink case and notation is symmetric (Table 2.1 summarizes some useful notation). (Specific differences will be elaborated, where necessary.)

Our assumptions for the backhaul network follow.

(C.1 - Backhaul network topology) Each access network node (either MBS or SC) is connected to the core network through an eNB aggregation gateway via a certain number of backhaul links that constitute the backhaul network. This connection can be either direct ("star" topology) or through one or more SC aggregation gateways ("tree" or "mesh" topology).

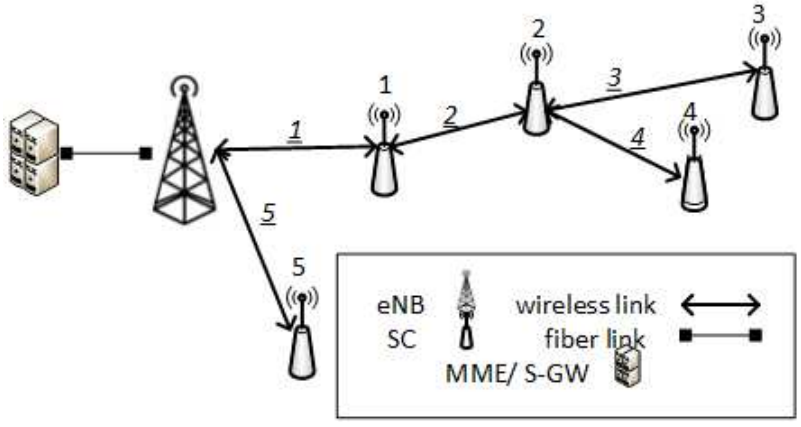


Figure 3.1: Future Backhaul topology of a HetNet.

Without loss of generality, we assume that there is a fiber link from the eNB to the core network, and focus on the set of capacity-limited backhaul links (wired or wireless) connecting SCs to the eNB, denoted as $\mathcal{B}_h$. We denote as routing path $\mathcal{B}_h(i)$ the set of all backhaul links $j \in \mathcal{B}_h$ along which traffic is routed from BS i to an eNB aggregation point, and we assume that it is *given* (e.g., calculated in practice as a Layer 2 (L2) spanning tree). For example, in Fig. 3.1, $\mathcal{B}_h(1) = \{1\}$, and $\mathcal{B}_h(3) = \{1, 2, 3\}$. We further denote as $\mathcal{B}(j)$ the set of all BS $i \in \mathcal{B}$ whose traffic is routed over backhaul link j. E.g., $\mathcal{B}(1) = \{1, 2, 3, 4\}$ and $\mathcal{B}(2) = \{2, 3, 4\}$ in Fig. 3.1.

(C.2 - Backhaul Resource Allocation Policy) Each $j \in \mathcal{B}_h$ backhaul link is associated

with a total capacity $C_h(j)$. While traditional backhaul links are multiplexed using FDD, nowadays TDD gains more ground due to the performance improvements it promises [97]. So, in the context of TDD, we introduce the backhaul resource allocation parameter $0 < Z(j) < 1$, that splits the backhaul capacity of the j link between DL ($Z(j) \to 1$) and UL ($Z(j) \to 0$). To simplify discussion, throughout the chapter we assume that $Z(j)$ is pre-determined and remains fixed. However, in Chapter 4 we show how one can optimally derive the the various $Z(j)$ given the current traffic conditions.

Note that, backhaul links usually don't implement any particular scheduling algorithm, so they can be seen as a data "pipe".

(C.3 - Backhaul load) The DL load on a backhaul link j consists of the sum of DL loads of all BSs using that link ($i \in \mathcal{B}(j)$), divided by its offered backhaul capacity [95]

$$\sum_{i \in \mathcal{B}(j)} \frac{\frac{\rho_i^D}{\zeta_i} \cdot (\zeta_i \cdot \tilde{c}_i^D)}{Z(j) \cdot C_h(j)} = \sum_{i \in \mathcal{B}(j)} \frac{\rho_i^D \cdot \tilde{c}_i^D}{Z(j) \cdot C_h(j)}. \tag{3.1}$$

where $\tilde{c}_i^D$ is a parameter use to "dimension" the BH link and corresponds to an estimate of the maximum DL total rate that BS i might request the backhaul to transport. A BS is characterized by its "peak" rate (often upper bounded by the maximum MCS available), and a "busy" rate when this BS serves many users [3]. The latter is usually quite smaller than the former, since users near the edge of the cell tend to bring the average rate down. However, the use of channel-based scheduling and related multi-user diversity gains suggest that conservatively setting $\tilde{c}_i^D$ closer to its nominal peak value is safer. In practice, a BS can directly measure it.

(C.4 - Backhaul provisioning) Each BH link j is associated with a backhaul load (see C.3), that shall be maintained below 1 to prohibit backhaul congestion. As a result, each BH link is associated with a *backhaul constraint*:

$$\sum_{i \in \mathcal{B}(j)} \frac{\rho_i^D \tilde{c}_i^D}{Z(j) \cdot C_h^D(j)} < 1, \ \forall j \in \mathcal{B}_h \tag{3.2}$$

Throughout this dissertation, we assume that the backhaul network is either *under-provisioned* if the capacity of *at least* one backhaul link is exceeded, or *over-provisioned* otherwise.

3.3 Backhaul-Aware User Association: Problem and Optimal Rules.

Our aim remains to find appropriate values for the various association probabilities $p_i^D(x), p_i^U(x)$ under the underlying backhaul network. We include to our goals (i) that no backhaul link should be congested, (ii) the investigation of backhaul topology and capacity on key performance metrics, (iii) whether the harmonic/ arithmetic mean formula maximization and related insights for Split and Joint UL/DL association remain similar or not.

We follow the same presentation as in Chapter 2 (where the backhaul network was ignored or it was assumed to be over-provisioned). Thus, in Section 3.3.1 we formulate Optimization Problem 2 that portrays the α-fair user association problem in backhaul-limited networks. Investigating both star and tree backhaul topologies in Sections 3.3.2 and 3.3.3 we derive the optimal user association rules for both Split and Joint UL/DL association scenarios, respectively.

3.3.1 Feasible Set, Objective Function and Optimization Problem 2

While the feasible region of the aforementioned probabilities shall shrink (see Definition 1) by including the convex backhaul constraints defined in Eq. 3.2, we keep the same α-objective function that can capture various fairness degrees (see Definition 2).

The arising Optimization Problem 2, namely user association in HetNets with limited backhaul resources follows.

Definition 5. ***(Optimization Problem 2)*** *The UL/DL user association problem under backhaul constraints can be expressed*

$$\begin{aligned} &\underset{\rho}{\text{minimize}} \left\{ \phi_\alpha(\rho) | \rho \in \mathcal{F} \right\}, \\ &\textit{subject to} \sum_{i \in \mathcal{B}(j)} \frac{\rho_i^D \tilde{c}_i^D}{Z(j) \cdot C_h^D(j)} \prec 1, \ \forall j \in \mathcal{B}_h, \textit{(similar in UL)} \\ &\textit{(dependent) subject to } p_i^D(x) = p_i^U(x) \textit{ iff joint UL/DL association}, \forall x \in \mathcal{L}. \end{aligned} \tag{3.3}$$

The first constraint ensures that all backhaul links have utilization less that 1, and the second one requires that a UE shall be offloaded to the same BS for UL and DL in the joint UL/DL scenario.

3.3.2 Optimal Rules for Split UL/DL

We start out discussion, with the Split UL/DL case, where the UL and DL problem decouple from each other. Following the analysis of the previous Chapter, we focus on the DL (UL is symmetrical in terms of formulation). To better illustrate our approach, we first consider the simple star backhaul topology, and then generalize for a tree backhaul topology.

(Backhaul Scenario: Star Topology)

The first challenge we need to tackle for Optimization Problem 2 is the proper treatment of the backhaul constraints. While famous solvers for such convex problems shall try to tackle them through a centralized controller entity, e.g., through the Lagrangian dual function [92], we want to follow another direction so that our rules remain "device centric" and allow for distributed implementations. To that end, we chose to consider the backhaul constraints in the objective function as appropriate *penalty functions* [95]. This not only facilitates deriving a distributed implementation of the policy as it will turn out later, but also allows us to treat the backhaul constraint as a "soft" constraint that ends up being "hard" and satisfy convergence to a feasible solution [98].

To do so, we need to define a new indicator variable that helps us formulating the penalties. Specifically let $\mathcal{I}(i)$ show whether the i-th backhaul link is congested ($\mathcal{I}(i)$=1) or not ($\mathcal{I}(i)$=0). Precisely for a star topology (see C.2)

$$\mathcal{I}(i) = \begin{cases} 0, & \text{when } \frac{\rho_i \tilde{c}_i}{Z(j) C_h(i)} < 1 \\ 1, & \text{otherwise.} \end{cases} \tag{3.4}$$

Thus, the *constrained Optimization Problem 2* (ignoring the *dependent* constraint since we are in the Split UL/DL scenario) is equivalent to the following *unconstrained optimization problem*

$$\underset{\rho}{\text{minimize}} \left\{ \Phi_\alpha(\rho) = \phi_\alpha(\rho) + \gamma \sum_{i \in \mathcal{B}_h} \mathcal{I}(i) \left(\frac{\rho_i \tilde{c}_i}{Z(i) C_h(i)} - 1 \right)^2 | \rho \in \mathcal{F} \right\}, \tag{3.5}$$

$\phi_\alpha(\rho)$ is the standard α-cost function for each BS i, already analyzed in the previously. The second sum introduces a penalty for each backhaul link i whose capacity is exceeded ($\mathcal{I}(i) = 1$). Note that, in the star topology, there is a single backhaul link for each BS i, and $\|\mathcal{B}_h\| = \|\mathcal{B}\|$. So, we can use the same index i for both. This penalty function is quadratic on the amount of excess load. Quadratic penalty functions are often considered in convex optimization literature [98].

γ could be chosen as a large constant, introducing a "soft" constraint for the backhaul links (i.e., backhaul capacity could be slightly exceeded, if this really improves access performance), or be iteratively adapted using increasing values, so as to converge to a "hard" constraint. The latter is usually preferred in such optimization problems since it ensures that the algorithm converges and doesn't get stuck in steep valleys [98, 99].

Theorem 3.1. ***(Backhaul-aware association rule for star backhaul topology [Split scenario])*** *If $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ denotes the optimal load vector, the user association rule at location x is*

$$i(x) = \arg\max_{i \in \mathcal{B}} \left(\underbrace{c_i(x)}_{\text{user knowledge}} \cdot \overbrace{P_i}^{\text{BS broadcast message}} \right), \tag{3.6}$$

where,

$$P_i = \frac{(1-\rho_i^*)^\alpha}{1 + 2\gamma \cdot (1-\rho_i^*)^\alpha \cdot \tilde{c}_i \cdot \frac{\mathcal{I}(i)}{Z(i)C_h(i)} \cdot \left(\frac{\rho_i^* \tilde{c}_i}{Z(i)C_h(i)} - 1 \right)}.$$

Proof. We now prove that the above rule indeed minimizes Optimization Problem 2. This is a convex optimization problem. Its feasible set is convex, and the objective $\Phi_\alpha(\rho)$ is also convex due to the summation of two convex terms: the first is convex as discussed earlier, and the second due to the composition property of convexity [92]. Let ρ^* be the optimal solution of this minimization problem. Again, it is adequate to check for optimality if

$$\langle \nabla \Phi_\alpha(\rho^*), \Delta\rho^* \rangle \geq 0 \tag{3.7}$$

for all $\rho \in f$, where $\Delta\rho^* = \rho - \rho^*$. Let $p(x)$ and $p^*(x)$ be the associated routing probability vectors for ρ and ρ^*, respectively. Using the deterministic cell coverage generated by (3.6), the optimal association rule is given by:

$$p_i^*(x) = \mathbf{1}\left\{ i = \arg\max_{i \in \mathcal{B}} \frac{c_i(x)(1-\rho_i^*)^\alpha}{1 + 2\gamma \cdot (1-\rho_i^*)^\alpha \cdot \tilde{c}_i \cdot \frac{\mathcal{I}(i)}{Z(i)C_h(i)} \cdot \left(\frac{\rho_i^* \tilde{c}_i}{Z(i)C_h(i)} - 1 \right)} \right\}. \tag{3.8}$$

Before proceeding to the calculation of the inner product, we analytically calculate the derivative of the corresponding cost function $\Phi_\alpha(\rho)$, described in Eq. (3.5). The derivative is an i-th dimensional vector; the i-th element of which has value:

$$\nabla \Phi_\alpha(\rho_i) = \begin{cases} (1-\rho_i)^{-\alpha}, & \text{if } \frac{\rho_i \tilde{c}_i}{Z(i)C_h(i)} \leq 1 \\ (1-\rho_i)^{-\alpha} + \gamma \mathcal{I}(i) \frac{2\rho_i \tilde{c}_i^2 - 2\tilde{c}_i Z(i) C_h(i)}{Z(i) C_h(i)^2}, & \text{if } \frac{\rho_i \tilde{c}_i}{Z(i)C_h(i)} \geq 1. \end{cases} \tag{3.9}$$

When $\rho_i = \frac{Z(i)C_h(i)}{\tilde{c}_i}$, we work out explicitly from the definition to calculate the derivative. It is:

$$\lim_{\rho_i \to \frac{Z(i)C_h(i)}{\tilde{c}_i}^+} \nabla\Phi(\rho_i) = \lim_{\rho_i \to \frac{Z(i)C_h(i)}{\tilde{c}_i}^-} \nabla\Phi(\rho_i) = (1-\rho_i)^{-\alpha}. \tag{3.10}$$

Summarizing, the i-th element of the derivative of the considered function can be written:

$$\nabla\Phi(\rho_i) = (1-\rho_i)^{-\alpha} + \gamma\mathcal{I}(i)\frac{2\rho_i\tilde{c}_i^2 - 2\tilde{c}_i Z(i)C_h(i)}{Z(i)C_h(i)^2}. \tag{3.11}$$

To that end, the inner product defined in Eq. (3.7), becomes:

$$\begin{aligned}
\langle \nabla\phi(\rho^*), \Delta\rho^* \rangle &= \sum_{i\in\mathcal{B}} \left(\frac{1}{(1-\rho_i^*)^\alpha} + \gamma\mathcal{I}(i)\frac{2\rho_i^*\tilde{c}_i^2 - 2\tilde{c}_i Z(i)C_h(i)}{Z(i)C_h(i)^2} \right)(\rho_i - \rho_i^*) \\
&= \sum_{i\in\mathcal{B}} \frac{1 + 2\gamma\mathcal{I}(i)(1-\rho_i^*)^\alpha \frac{(\rho_i^*\tilde{c}_i^2 - \tilde{c}_i Z(i)C_h(i))}{Z(i)C_h(i)^2}}{(1-\rho_i^*)^\alpha} \int_{\mathcal{L}} \rho_i(x)\left(p_i(x) - p_i^*(x)\right)dx \\
&= \int_{\mathcal{L}} \frac{\lambda(x)}{\mu(x)} \sum_{i\in\mathcal{B}} \left(\frac{1 + 2\gamma(1-\rho_i^*)^\alpha \tilde{c}_i \frac{\mathcal{I}(i)}{Z(i)C_h(i)}\left(\frac{\rho_i^*\tilde{c}_i}{Z(i)C_h(i)} - 1\right)}{c_i(x)(1-\rho_i^*)^\alpha} \right)\left(p_i(x) - p_i^*(x)\right)dx.
\end{aligned}$$

Note that,

$$\begin{aligned}
&\sum_{i\in\mathcal{B}} p_i(x) \left(\frac{1 + 2\gamma(1-\rho_i^*)^\alpha \tilde{c}_i \frac{\mathcal{I}(i)}{Z(i)C_h(i)}\left(\frac{\rho_i^*\tilde{c}_i}{Z(i)C_h(i)} - 1\right)}{c_i(x)(1-\rho_i^*)^\alpha} \right) \geq \\
&\sum_{i\in\mathcal{B}} p_i^*(x) \left(\frac{1 + 2\gamma(1-\rho_i^*)^\alpha \tilde{c}_i \frac{\mathcal{I}(i)}{Z(i)C_h(i)}\left(\frac{\rho_i^*\tilde{c}_i}{Z(i)C_h(i)} - 1\right)}{c_i(x)(1-\rho_i^*)^\alpha} \right)
\end{aligned}$$

holds because $p_i^*(x)$ in Eq. 3.8 is an indicator for the minimizer of $\frac{1+2\gamma\cdot(1-\rho_i^*)^\alpha\cdot\tilde{c}_i\cdot\frac{\mathcal{I}(i)}{Z(i)C_h(i)}\cdot\left(\frac{\rho_i^*\tilde{c}_i}{Z(i)C_h(i)}-1\right)}{c_i(x)(1-\rho_i^*)^\alpha}$. Hence, Eq. 3.7 holds. □

Regarding the derived backhaul-aware association rule of Eq. 3.6, we note that when the capacity constraint for the backhaul link i is not active (i.e., $\mathcal{I}(i) = 0$, e.g., in provisioned backhaul networks), the above theorem states that the optimal association rule is the same as the one found in [10], or the one defined in Theorem 3.1 when $\theta \to 1$ (since, as discussed at the beginning of the section we only consider best effort flows in this Chapter). However, when the backhaul link of BS i gets congested, a second term is added in the denominator that penalizes that BS making it less preferable to UEs at location i, even if the offered radio access rate $c_i(x)$ is high, or the radio interface of i is not itself congested.

(Backhaul Scenario: Tree Topology)

We now consider a more complex backhaul scenario, where a single backhaul link might route traffic from multiple BSs, and the traffic of a single BS might be routed over multiple backhaul

links (multi-hop path) towards the eNB gateway. Now, the indicator variable $\mathcal{I}(j)$ turns out being (see also C.2)

$$\mathcal{I}(j) = \begin{cases} 0, & \text{when } \frac{\sum_{i\in\mathcal{B}(j)} \rho_i \tilde{c}_i}{Z(j)C_h(j)} < 1 \\ 1, & \text{otherwise.} \end{cases} \tag{3.12}$$

While the procedure of deriving the optimal rules follows similar steps as before, we highlight that now one shall consider the complete backhaul routing path to the eNB gateway (that might consist of multiple links with heterogeneous capacities). The corresponding unconstrained problem for a tree backhaul topology is

$$\underset{\rho}{\text{minimize}} \left\{ \Phi_\alpha(\rho) = \phi_\alpha(\rho) + \gamma \sum_{j\in\mathcal{B}_h} \mathcal{I}(j) \left(\frac{\sum\limits_{i\in\mathcal{B}(j)} \rho_i \tilde{c}_i}{Z(j)C_h(j)} - 1 \right)^2 \right\}, \tag{3.13}$$

Theorem 3.2. ***(Backhaul-aware association rule for tree backhaul topology [Split scenario])*** *The optimal user-association rule at location x is now*

$$i(x) = \arg\max_{i\in\mathcal{B}} \left(\underbrace{c_i(x)}_{\text{user knowledge}} \cdot \overbrace{P_i}^{\text{BS broadcast message}} \right), \tag{3.14}$$

where each BS $i \in \mathcal{B}$ shall broadcast

$$P_i = \frac{(1-\rho_i^*)^\alpha}{1 + 2\gamma \cdot (1-\rho_i^*)^\alpha \cdot \tilde{c}_i \sum\limits_{j\in\mathcal{B}_h(i)} \frac{\mathcal{I}(j)}{Z(j)C_h(j)} \cdot \left(\frac{\sum\limits_{k\in\mathcal{B}(j)} \rho_k^* \tilde{c}_k}{Z(j)C_h(j)} - 1 \right)}.$$

Proof. The steps of this proof are similar to the star case, so we present here directly the corresponding inner product.

$$\begin{aligned} &\langle \nabla\Phi_\alpha(\rho^*), \Delta\rho^* \rangle = \\ &= \sum_{i\in\mathcal{B}} \left(\frac{1}{(1-\rho_i^*)^\alpha} + 2\gamma \sum_{j\in\mathcal{B}_h(i)} \mathcal{I}(j) \left(\frac{\sum_{k\in\mathcal{B}(j)} \rho_k^* \tilde{c}_k}{Z(j)C_h(j)^2} \tilde{c}_i - \frac{\tilde{c}_i}{Z(j)C_h(j)} \right) \right) (\rho_i - \rho_i^*) \\ &\quad \cdot \int_{\mathcal{L}} \rho_i(x) \left(p_i(x) - p_i^*(x) \right) dx = \\ &= \int_L \frac{\lambda(x)}{\mu(x)} \sum_{i\in\mathcal{B}} \left(\frac{1 + 2\gamma(1-\rho_i^*)^\alpha \tilde{c}_i \sum\limits_{j\in\mathcal{B}_h(i)} \frac{\mathcal{I}(j)}{Z(j)C_h(j)} \cdot \left(\frac{\sum\limits_{k\in\mathcal{B}(j)} \rho_k^* \tilde{c}_k}{Z(j)C_h(j)} - 1 \right)}{c_i(x)(1-\rho_i^*)^\alpha} \right) \cdot \\ &\quad \cdot \left(p_i(x) - p_i^*(x) \right) dx \geq 0, \end{aligned} \tag{3.15}$$

due to the corresponding maximizer $p_i^*(x)$ derived from (3.14). □

As one can see, P_i, i.e., the value of the BS broadcast message differs compared to the star backhaul topology. First, the penalty term in the denominator now considers the whole backhaul

path $\mathcal{B}_h(i)$ that traffic from BS i traverses, and adds a penalty for *every* link along that path that is congested (outer sum in the denominator). This observation provides some support for the number of backhaul hops heuristic proposed in [18, 90]. However, our analysis also suggests that it can be suboptimal, as a path with few hops might still include one or more congested links, and provides the optimal way to weigh in the amount of congestion on each backhaul link.

Second, the actual congestion on each backhaul link j is now not only dependent on the load of the candidate BS i, but also on other BSs whose load is routed over j. Hence, a BS i which would otherwise be a good candidate for traffic at location x, might still be penalized and not selected, even if it does not impose itself a large load on a backhaul link j. This is because *other* BSs sharing the same backhaul link might be heavily loaded or congested.

In the case of split UL/DL traffic, the above analysis can be applied *separately* on UL and DL traffic, and optimize UL and DL associations independently. It thus suffices to consider the respective quantities with the appropriate superscript "D" or "U", as explained earlier.

Finally, although we have provided separate solutions for star and tree topologies, to better illustrate our approach, the optimal rule for the tree topology is generic, and includes star topologies as well.

3.3.3 Optimal Rules for Joint UL/DL

Here, we need to modify our framework accordingly, as we did in the previous Chapter, by considering the (dependent) constraint $p_i^D(x) = p_i^U(x)\ \forall i \in \mathcal{B}$. Additionally, we need to extend the penalty function to consider both uplink and downlink capacity being exceeded on the backhaul link.

The corresponding objective is

$$\Phi_\alpha(\rho) = \phi_\alpha(\rho) + \gamma \sum_{k\in\{D,U\}} \sum_{j\in\mathcal{B}_h} \mathcal{I}^k(j) \left(\frac{\sum\limits_{i\in\mathcal{B}(j)} \rho_i^k \tilde{c}_i^k}{C_h^k(j)} - 1 \right)^2 . \tag{3.16}$$

We present our results directly for the general case of tree backhaul topology, and we remind the reader that this is applicable to star backhaul topologies as well.

Theorem 3.3. ***(Backhaul-aware association rule for tree backhaul topology [Joint scenario])*** *The optimal user-association rule at location x is*

$$i(x) = \arg\max_{i\in\mathcal{B}} \frac{1}{\frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)}}, \tag{3.17}$$

where the BS broadcast messages are:

$$P_i^D = \frac{e^D \cdot (1-\rho^{U*})^{\alpha^U}}{(1-\rho^{*D})^{\alpha^D} \cdot (1-\rho^{U*})^{\alpha^U}},$$

$$P_i^U = \frac{e^U \cdot (1-\rho^{*D})^{\alpha^D}}{(1-\rho^{*D})^{\alpha^D} \cdot (1-\rho^{U*})^{\alpha^U}}.$$

If $g^D = \tau, g^U = 1 - \tau$, then the corresponding factors of the harmonic mean formula follow

$$e^l = \frac{z^l \left(g^l + 2\gamma \left(1 - \rho_i^{*l}\right)^{\alpha^l} \sum\limits_{j \in \mathcal{B}_h(i)} \frac{\mathcal{I}^l(j)}{Z(j) C_h^l(j)} \left(\frac{\sum\limits_{k \in \mathcal{B}(j)} \rho_k^{*l} \tilde{c}_k^l}{Z(j) C_h^l(j)} - 1 \right) \right)}{\mu^l(x)}, l \in \{D, U\}.$$

Proof. Since we depict immediately the tree topology, we provide the proof with analytical steps. Let ρ^* be the optimal solution of this optimization problem, that corresponds to the optimal association rule (3.17). In a similar way, it is adequate to check the following condition for optimality

$$\langle \nabla \Phi_\alpha(\rho^*), \Delta\rho^* \rangle \geq 0 \tag{3.18}$$

for all $\rho \in \mathcal{F}$, where $\Delta\rho^* = \rho - \rho^*$. Let $p(x)$ and $p^*(x)$ be the associated routing probability vectors for ρ and ρ^*, respectively. Using the deterministic cell coverage generated by (3.17), the optimal association rule is given by:

$$p_i^*(x) = \mathbf{1}\left\{ \arg\max_{i \in \mathcal{B}} \frac{1}{\frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)}} \right\}. \tag{3.19}$$

Similarly to the previous case, we calculate the inner product of Eq. (3.18). It is

$$\begin{aligned}
&\langle \nabla \Phi_\alpha \left(\rho^*\right), \Delta\rho^* \rangle = \\
&= \sum_{i \in \mathcal{B}} \left(\tau \cdot \frac{1}{(1 - \rho_i^{*D})^{\alpha^D}} + 2\gamma \sum_{l \in \mathcal{B}_h} \mathcal{I}^D(l) \left(\frac{\sum_{j \in \mathcal{B}(l)} \rho_j^D \tilde{c}_j^D}{Z(l) C_h(l)^2} \tilde{c}_i^D - \frac{\tilde{c}_i^D}{Z(l) C_h(l)} \right) \right) \rho_i^D - (\rho_i^{*D})) + \\
&+ \sum_{i \in \mathcal{B}} \left((1 - \tau) \cdot \frac{1}{(1 - \rho_i^{*U})^{\alpha^U}} + 2\gamma \sum_{l \in \mathcal{B}_h} \mathcal{I}^U(l) \left(\frac{\sum_{j \in \mathcal{B}(l)} \rho_j^U \tilde{c}_j^U}{Z(l) C_h(l)^2} \tilde{c}_i^U - \frac{\tilde{c}_i^U}{Z(l) C_h(l)} \right) \right) (\rho_i^U - (\rho_i^{U*})) = \\
&= \sum_{i \in \mathcal{B}} \left(\tau \cdot \frac{1}{(1 - \rho_i^{*D})^{\alpha^D}} + 2\gamma \sum_{l \in \mathcal{B}_h} \mathcal{I}^D(l) \left(\frac{\tilde{c}_i^D (\sum_{j \in \mathcal{B}(l)} \rho_j^D \tilde{c}_j^D - Z(l) C_h(l))}{Z(l) C_h(l)^2} \right) \right) \cdot \\
&\quad \cdot \int_{\mathcal{L}} \rho_i^D(x) \left(p_i(x) - p_i^*(x)\right) dx + \\
&+ \sum_{i \in \mathcal{B}} \left((1 - \tau) \cdot \frac{1}{(1 - \rho_i^{*D})^{\alpha^U}} + 2\gamma \sum_{l \in \mathcal{B}_h} \mathcal{I}^U(l) \left(\frac{\tilde{c}_i^U (\sum_{j \in \mathcal{B}(l)} \rho_j^U \tilde{c}_j^U - Z(l) C_h(l))}{Z(l) C_h(l)^2} \right) \right) \cdot \\
&\quad \cdot \int_{\mathcal{L}} \rho_i^U(x) \left(p_i(x) - p_i^*(x)\right) dx = \\
&= \int_L \lambda(x) \sum_{i \in \mathcal{B}} \begin{pmatrix} \frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)} \\ 1 \end{pmatrix} \cdot \left(p_i(x) - p_i^*(x)\right) dx.
\end{aligned} \tag{3.20}$$

Note that,

$$\sum_{i \in \mathcal{B}} p_i(x) \begin{Bmatrix} \frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)} \\ 1 \end{Bmatrix} \geq \sum_{i \in \mathcal{B}} p_i^*(x) \begin{Bmatrix} \frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)} \\ 1 \end{Bmatrix} \tag{3.21}$$

holds because $p_i^*(x)$ due to the minimizer defined in (3.19). Hence (3.18) holds.

□

Remark 1. We note that still in backhaul-limited networks, in case of Joint UL/DL association, if one jointly considers potentially conflicting objectives, it is optimal to associate a user with the BS that maximizes the *harmonic mean* of the individual association rules, when considering each objective alone. Note that now, the actual rule is more complex, as it weighs each objective with the factor e^D, e^U that is related to not only radio access performance but also backhaul penalties, and we provide the optimal way to weight them. Or, in case of Split UL/DL it is optimal to optimize their arithmetic mean (each objective independently). Thus, our results prove that we can still include more dimensions to our setup, and *flexibly* derive the optimal association rule in more complicated offloading scenarios that also include potential limitations in the backhaul network.

Remark 2. Our derived rules are still "device centric", and maintain the same desired properties. In particular, the rules are still *scalable* since there is a constant amount of the BS broadcast messages irrespective of the number of users and backhaul topology. This plays an interesting contribution since we showed that through our rules: *each BS is able to reflect its potential of serving more downlink users in terms of both access and backhaul resources by broadcasting only one value, even when the backhaul topology up to the eNB gateway is rather "mesh" with multiple hops, (e.g., this value is P_i in Theorem 3.1).* To that end, using the same distributed implementation scheme proposed in Section 2.3.4 these rules can be applied by the users iteratively, and the BS loads will eventually converge to the optimal values under the feasible set that the backhaul constraints sketch. Obviously, our rules remain *simple*, and offer *flexible performance*.

Note that our model allows that the backhaul links can have heterogeneous backhaul link technologies and capacities, by appropriately fixing the parameter $C_h(j)$ (see C.2).

3.4 Simulations

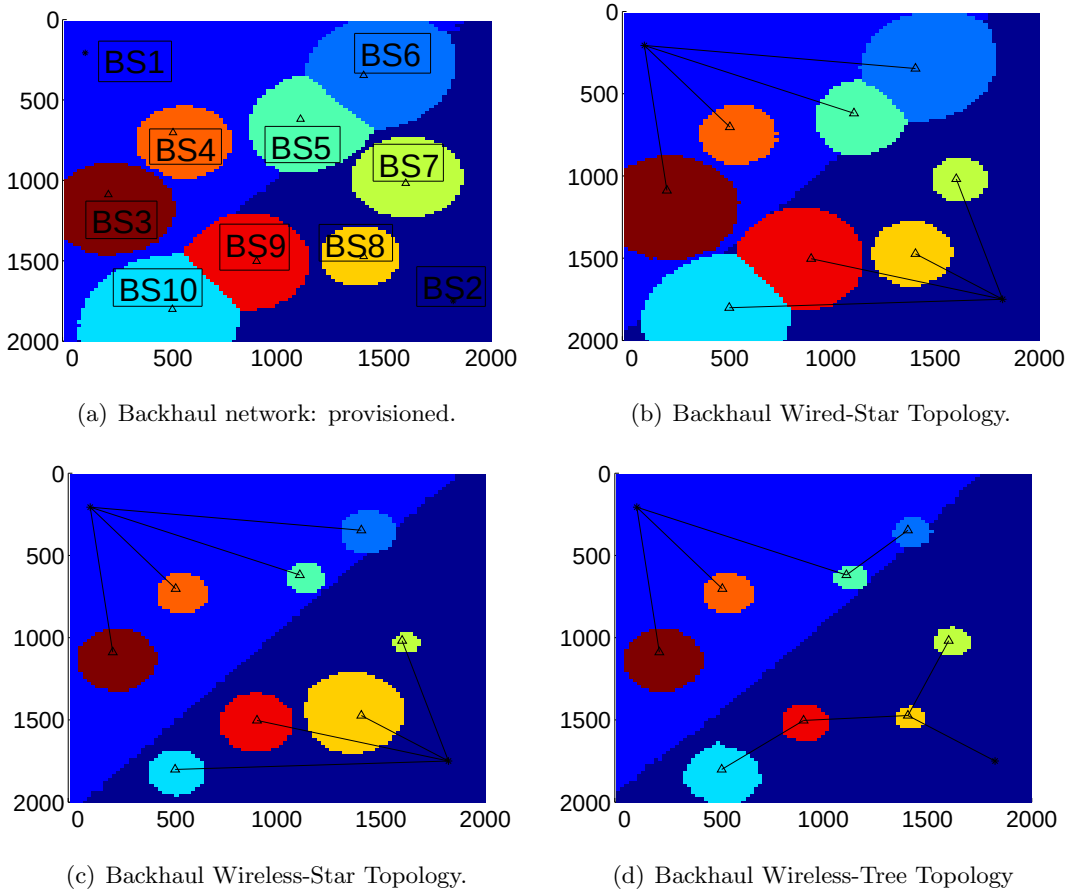
In this Section we focus on some backhaul-limited network scenarios, and evaluate their performance by applying our derived rules.

We again consider the same 2×2 km^2 topology as in Section 2.4. We assume that this area is covered by two macro BSs and eight SCs as considered previously, and the traffic demand rate $\lambda(x)$ ($flows/hour$ per unit area) as well as the other parameters remains the same (see Fig. 2.2). If not explicitly mentioned, we assume $\tau = 0.5$, and the Split UL/DL scenario as default.

We remind to the reader that our focus in this Chapter is on the backhaul links *between the macro cells and SCs* (for simplicity we assume provisioned links between the macro cells and core network). As already discussed in assumption C.1, we investigate two different backhaul topology families: (i) "star" topologies (single-hop paths), (ii) "tree" topologies (with multi-hop paths), along with two backhaul links types: *wired and wireless*[1]. Our aim is to evaluate the derived association rules for different *under-provisioned* scenarios, by fixing $\alpha^D = \alpha^U = 1$ (throughput optimal values). Also, we assume *fixed* backhaul routing paths, pre-established

[1]Note that copper and fiber access are the key technologies for wired backhaul links, and microWave and millimeter-wave P2P or P2MP access are the counterpart for the wireless backhaul links [100].

with traditional Layer 2 routing, that the BH capacities on the DL and UL are the same (i.e. $C_h^D(j) = C_h^U(j) = C_h$, $\forall j \in \mathcal{B}_h$), and if not explicitly mentioned we assume them to be equal to $400Mbps$. We maintain this assumption to facilitate our discussion, although our framework works for heterogeneous backhaul links and UL/DL capacities (see C.2).



(a) Backhaul network: provisioned.

(b) Backhaul Wired-Star Topology.

(c) Backhaul Wireless-Star Topology.

(d) Backhaul Wireless-Tree Topology

Figure 3.2: Optimal downlink user associations when (a) backhaul network is provisioned, or when backhaul network is under-provisioned and we consider (b) Wired-Star Topology, (c) Backhaul Wireless-Star Topology, (d) Backhaul Wireless-Tree Topology.

Before proceeding, we need to make an assumption about the backhaul link capacities. In case of *wired* backhaul links, we assume that the peak backhaul capacity C_h is always guaranteed. For *wireless* backhaul links we adopt a simple model associating peak backhaul capacity to distance: if the length of the i-th link is r_i, the peak capacity drops as:

$$d(r_i) = \begin{cases} 1, & r_i \leq r_0 \\ (\frac{r_0}{r_i})^n, & \text{otherwise,} \end{cases} \tag{3.22}$$

where r_0 is some threshold range within which the maximal rate is obtained (e.g. Line-of-Sight), and n is the attenuation factor. Hence, the available capacity drops to $d(r_i)C_h(j)$ $(\leq C_h(j))$. For our simulations, we assumed that $r_0 = 200m$, and $n = 3$. While the above model is perhaps

oversimplifying, our main goal is to simply include a generic model for the propagation related impact on wireless backhaul, compared to wired, without getting into the details of specific backhaul implementations. For detailed path loss models for different backhaul technologies, we refer the interested reader to [71].

Coverage Snapshots. In Fig. 3.2(a) we depict the optimal DL user-associations for provisioned backhaul network with respect to the traffic arrival rates shown in Fig. 2.2. Compared to the associations showed in Figure 2.3(a) where $\alpha^D = \alpha^U = 0$, we note that now some SCs have slightly increased coverage area, in order to improve the mean user throughput [10].

In the following, we focus on different *under-provisioned* backhaul scenarios, and study the DL associations. In Fig. 3.2(b) we adopt a *wired-star* backhaul topology, where SCs shrink their coverage areas, by handing-over users to other BSs, in order to offload the corresponding (under-provisioned) backhaul links; this phenomenon becomes more intense in the "hot-spot" areas (e.g., BS7 have vastly decreased their coverage areas) due to the higher traffic demand. Similarly, in Fig. 3.2(c), we assume a *wireless-star* backhaul topology, where SCs further decrease their coverage areas, due to the higher backhaul capacity loss caused from the long wireless links (see Eq.(3.22)).

In Fig. 3.2(d) we adopt a *wireless-tree* topology, where some SCs are required to carry also traffic of other SCs, and end up more congested. As a result, most SCs further decrease their coverage area, compared to the star-wireless topology. However, BS7 and BS10 enlarge their coverage areas, compared to the star case. This occurs because these SCs are far from the eNB, and multi-hop topology allows them to route their traffic over shorter wireless links with smaller capacity losses, compared to the star case (Fig. 3.2(c)). Hence, there are two main factors affecting the coverage areas in such wireless backhaul networks: (*topology*) each BS-load might traverse through multi-hop backhaul paths, by "wasting" resources from more than one backaul links (drawback for tree topologies); *(location)* the higher the η,r_0 the worse the capacity loss "wastage" over a dedicated direct backhaul link (drawback for star topologies that require longer links).

As backhaul networks become increasingly complex, e.g. "mesh" topologies, each BS has *multiple* possible routing paths to follow, beyond what is shown in the figures (we remind the reader that the above shown topologies are simply the given spanning routing trees). The above observations thus underline the shortcomings of predetermined, Layer 2 (L2) backhaul routing mechanisms, and call for a *joint* optimization of user-association on the radio access network along with dynamic, Layer 3 (L3) backhaul routing.

Under-provisioning impact on user performance. Figure 3.3, 3.4 depict the *average* DL and UL user throughputs, as a function of the backhaul capacity constraint C_h, on different scenarios. Generally, as C_h drops, the mean throughputs are decreased, since users are handed over to (potentially far-away) macro BSs, causing performance degradation. Interestingly, *the slope of the dropping rate* becomes more steep for lower values of C_h, due to the logarithmic capacity formula chosen in assumption (B.2). Also, as C_h increases, the average throughputs "converge" to the value corresponding to a provisioned backhaul network. Note that the average UL throughput convergences more quickly, compared to the DL. This happens due to the asymmetry between the DL and UL traffic demand on the radio access network: the UL one is much lower, mainly due to the asymmetry between the transmission powers of BSs and UEs, as well as different file sizes assumed in each direction. Beyond this point, the UL backhaul resources will be underutilized. This calls for a *flexible* TDD duplexing scheme, that will dynamically

distribute the backhaul resources accordingly, for example by giving more backhaul resources to DL when the UL demand is already satisfied. Finally, in the wired case, star topology is always slightly better than the tree, whereas in the wireless the opposite, as explained earlier.

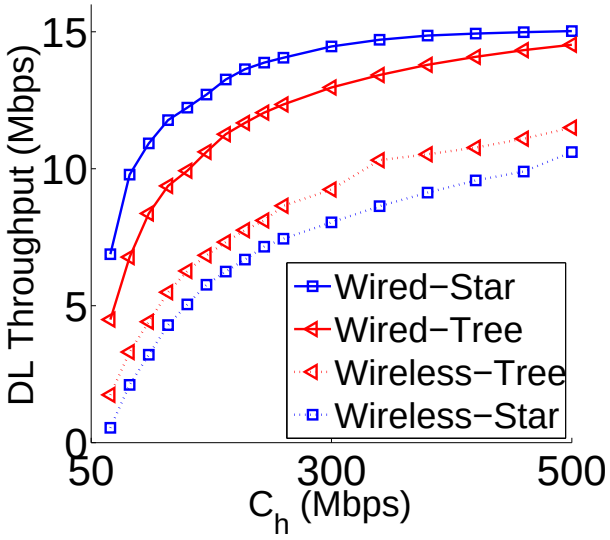


Figure 3.3: Downlink user throughput considering all the users for various backhaul topologies (in Mbps).

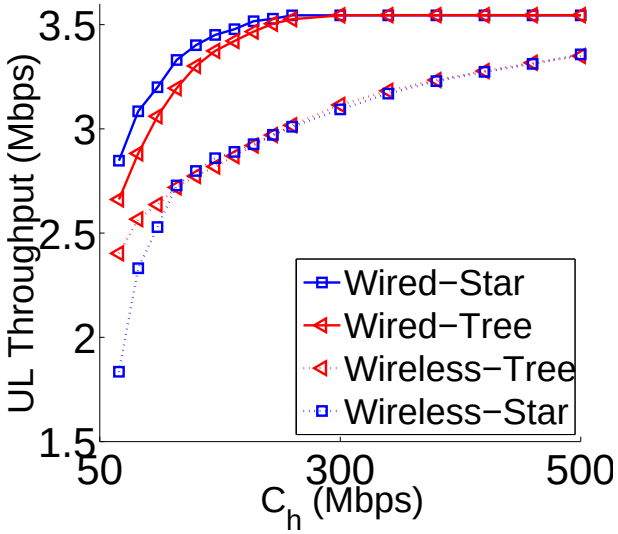


Figure 3.4: Uplink user throughput considering all the users for various backhaul topologies (in Mbps).

Table 3.1: Mean throughput for handed-over users (in Mbps).

Topology	$C_h = 50$	250	500 (Mbps)
DL / UL thr.: Star-Wired	1.1 / 0.2	3.1 / 1.6	4.1 / X
DL / UL thr.: Tree-Wired	0.6 / 0.1	2.4 / 0.7	3.2 / X
DL / UL thr.: Tree-Wirel.	0.2 / 0.03	1.7 / 0.07	2.1 / 0.15
DL / UL thr.: Star-Wirel.	0.1 / 0.001	1.4 / 0.05	1.7 / 0.02

One could notice that user throughputs drop slightly on the C_h constraint, e.g. in a wired-star topology if C_h drops 500 $\rightarrow$ 50 Mbps (10 times), the mean user throughput only drops 15 $\rightarrow$ 6 Mbps ($\sim$ 3 times). This is due to the fact that, under-provisioned backhaul links do not affect the whole network, but specific groups of users associated with the cells that suffer from low backhaul capacity. To better illustrate this, in Table 3.1 we show the average throughput of the *handed-over users*, as a function of C_h. Indeed, their performance is severely affected: for the same scenario, their DL throughput drops all the way to 1.1 Mbps ($\sim$ 15 times). (In scenarios with no handovers, we mark the respective table entry with an X.)

Under-provisioning impact on Network Performance. Turning our attention to network-related performance, Fig. 3.5 considers spectral efficiency ($bit/s/Hz$), *normalized* by the *maximum* corresponding value when the network is provisioned. Load-balancing ("utilization") efficiency is further considered in Fig. 3.6 in terms of the MSE metric, described earlier. Both efficiencies converge to 1 as the network gets provisioned. Low C_h values will push users to handover to far-away BSs, and this will potentially decrease their $SINR$ (spectral efficiency decrease), and create steep differences between BSs loads, e.g. by congesting macro BSs and under-utilizing the SCs (load balancing decrease). Note that, the joint degradation of these performances also impacts user performance negatively (e.g. user throughput), as explained in Section B.6. Regarding spectral efficiency, more specifically, although in the wired scenario, star topology is always better compared to the tree, in the wireless scenario this is not the case. For low values of C_h, the star topology is worse, due to the higher capacity loss of the long and direct links. However, as C_h is increased, and some links start becoming provisioned in the star topology, the capacity loss cost due to the long wireless links in the star topology, is dominated from the capacity loss cost due to multi-hop sharing links of the tree topology, by making tree a worse choice. We highlight that this trade-off can suggest different topologies as optimal in different under-provisioned scenarios, and can affect different performance metrics.

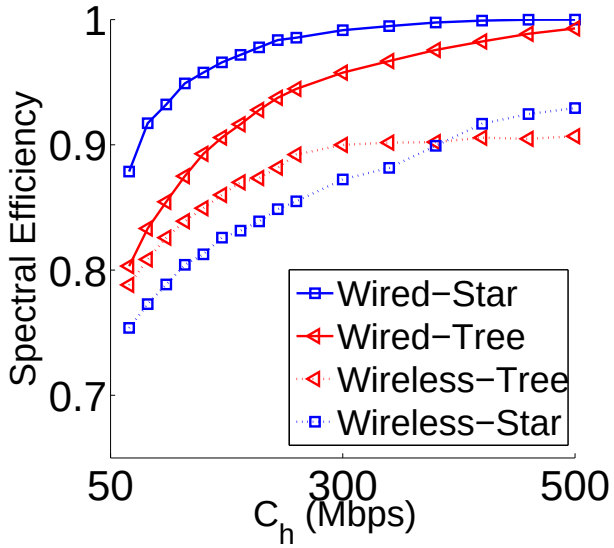


Figure 3.5: Downlink Spectral Efficiency (SE) for different backhaul topologies (normalized).

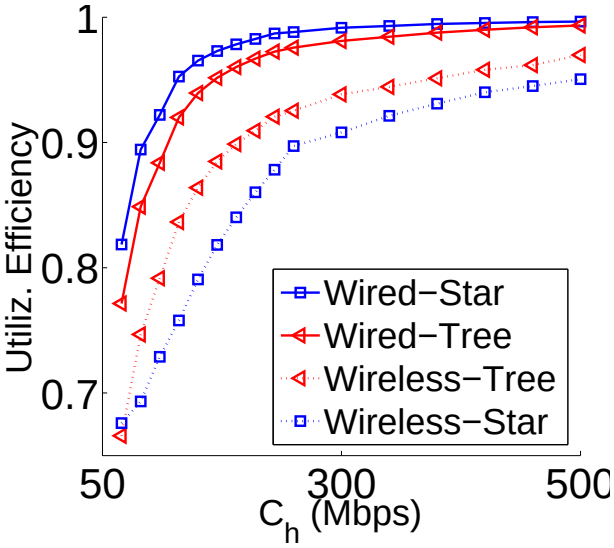


Figure 3.6: Downlink Load Balancing (or, Utilization) (LB) Efficiency for different backhaul topologies (normalized).

Split UL/DL impact. As discussed earlier, while split is able to optimize the DL and UL performance, *simultaneously*, joint UL/DL association is incapable of this parallel optimization and using $0 \leq \tau \leq 1$ we can trade-off which dimension carries more importance. Table 3.2 illustrates the *performance improvements* that split promises over the joint UL/DL association, in terms of various metrics, for various τ when backhaul is underprovisioned. We underline that split enhances the UL performance considerably, e.g. the average UL throughput is increased up to 37%. This is due to the *dependency* that joint UL/DL generates between the DL and UL associations in the access network, that often makes the DL the bottleneck in the backhaul (due to aforementioned asymmetry between the peak access rates). Thus, DL will often "preempt"

the backhaul constraint, and potentially (i) leave some UL resources unused, (ii) cause UL performance degradation.

Table 3.2: Split Vs. Joint UL/DL association Improvements

Performance	$\tau = 0$	$\tau = 0.5$	$\tau = 1$
DL / UL Throughput	6% / 32%	4% / 35%	0% / 37%
DL / UL Spectr. Eff.	4% / 29%	3% / 31%	0% / 33%
DL / UL Uiliz. Eff.	7% / 34%	4% / 38%	0% / 41%

3.5 Conclusion

In this Chapter, we considered a complete user-association framework for future HetNets that encompasses joint optimization of access and backhaul networks. We showed how different backhaul topologies and capacity limitations affect the user and network performances. Initial simulation results corroborate the correctness of our framework, and reveal interesting trade-offs for different network scenarios, as well as potential drawbacks of schemes operated in the backhaul, currently.

Chapter 4

Hierarchize then optimize (HoP): joint user association, access and backhaul TDD Allocation Optimizations.

4.1 Introduction

In the previous chapters we tried to shed some light on how (i) traffic differentiation (Chapter 2), as well as (ii) backhaul limitations (Chapter 3) shall affect user association in next-generation HetNets by trading-off performance. We thus provided both analytical insights as well as qualitative and qualitative performance evaluation to explain them.

Nevertheless, in our proposed frameworks we assumed that the bandwidth resources allocation between UL and DL at the BS level (see B.2) *remain static*. While such schemes indeed improve performance, they do so under a such defensive, fixed and pre-determined resource allocation. It is easy to understand that more flexible allocation policies that distribute the resources under a more sophisticated manner (e.g., rather than the fixed 50-50 for DL/UL) shall further enhance performance by adding additional degrees of freedom for the operator. For instance, as explained earlier, if most of the associated users of a BS are currently doing UL, this BS should turn its resource allocation to UL to avoid resource wastage on DL as well as improve performance on UL (i.e., by decreasing its ζ).

Such dynamic/flexible TDD schemes require additional considerations, in particular in asymmetric interference scenarios. As a typical example, if an SC is doing UL while a nearby MC is transmitting on the DL (with much higher power), the performance of the SC might be significantly degraded from this *cross-interference* (see e.g., Fig. 4.1). Enhanced Inter-Cell Interference Coordination (eICIC) schemes such as Almost Blank Subframes (ABS) could alleviate this but only to some extent [101, 102]. Large amounts of mismatch might lead to excessive usage of resources for eICIC, instead of user traffic, leading instead to considerable performance degradation. While a key-enabler for 5G networks, namely "enhanced Interference Mitigation and Traffic Adaptation" (eIMTA) has been standardized in LTE-A Release 12 [46] for this opportunity, it is not clear which allocation scheme is the best option under different scenarios and how it should interact with user association.

We underline that such dynamic allocation schemes shall not only be considered in the radio access TDD resources. Backhaul resource allocation policies (see C.2) should interact with the various (e.g., *user association* and *flexible TDD*) radio access policies, in order to satisfy the UL and DL traffic demands that the latter generate. E.g., as showed in the previous Chapter (depending on user associations and related dynamics), if the downlink traffic on the radio network level is lighter than the uplink (e.g., due to the asymmetry between (i) transmit power of BS-UE as well as (ii) corresponding file-sizes) the backhaul should distribute more resources to DL to improve performance.

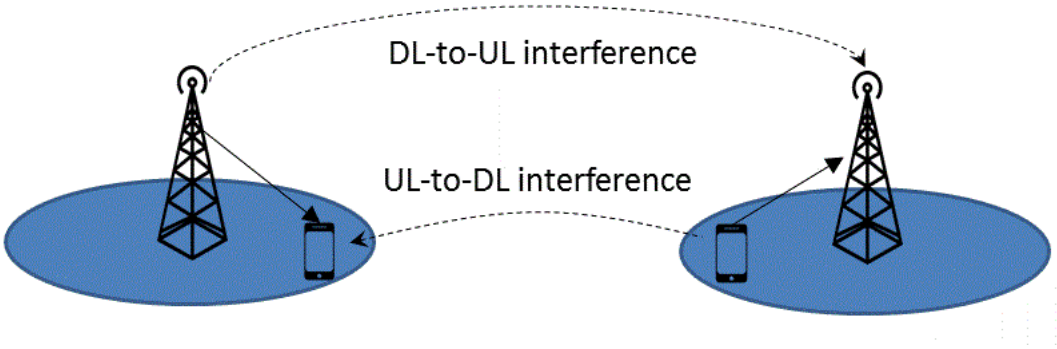


Figure 4.1: Different types of cross-interference: downlink-to-uplink (DL-to-UL) and uplink-to-downlink (UL-to-DL) in dynamic TDD systems.

By this token, in this section we stress that the success of future HetNets not only depends on a good enough access/backhaul TDD flexible scheme and decent user association rules, *but on their optimal interplay*. To that end, we propose an optimization framework that considers all these dimensions in one complex problem. To our best knowledge, this is the first work to attempt it. Our main contributions can be summarized as follows:

(1) We propose an analytical framework to study the interaction between (i) user association, (ii) radio access resource allocation with cross-interference management, and (iii) backhaul resource allocation, significantly extending Chapters 2 and 3.

(2) We show that the joint problem is non-convex, unlike variants studied in the past [10, 72, 93, 95], but possesses some "hidden" convexity properties that allows its decomposition into three subproblems. These subproblems can be solved through convex optimizers, at possibly different elements (e.g. UE, BS, backhaul link), and at different timescales, facilitating a hierarchical implementation.

(3) Using extensive simulations, we highlight complex trade-offs involved between the different subproblems, and show that significant performance improvements could be achieved compared to current standards.

The reminder of the Chapter is organized as follows. In Section 4.2 we provide a high-level tutorial about biconvex optimization problems and some related insights that we will need later. In Section 4.3 we provide the system model of our framework. In Section 4.4 and 4.5 we present the complete biconvex optimization problems for under- as well as over- provisioned backhaul networks, and attempt to tackle them using decomposition properties. Simulation results that promise significant improvements compared to schemes with fixed allocations are illustrated in

Section 4.6.

4.2 Bi-convexity

Many networking problems fall into the category of convex optimization problems. Such problems (as the ones studied in Chapters 2 and 3) can be efficiently and quickly tackled by standard convex solvers centrally or distributively (e.g., through Lagrange multipliers, Newton descent method or distributed-gradient techniques ,see [92]). However, more realistic and multi-variable problems usually create partially non-convexities and thus end up being non-convex. This prohibits the usage of convex solvers, and require more sophisticated solvers that usually cannot guarantee convergence in a reasonable amount of time (e.g., famous solvers for such NP-hard problem are pruning, branch and bound).

Nevertheless, many of such problems usually have some "hidden convexities" that when revealed, they can be solved efficiently. Biconvex optimization problems fall into this category. We are going to offer some insights for such optimization problems, and reveal interesting properties. For an analytical survey on such problems, we refer the interested reader to [103].

Let $X \in \mathcal{R}^n$ and $Y \in \mathcal{R}^n$ be two non-empty convex sets and let $\mathcal{F} \subseteq X \times Y$. We define $x-$ and $y-$ sections of $\mathcal{F}$ as follows

$$\mathcal{F}_x := \{y \in Y : (x, y) \in \mathcal{F}\}, \mathcal{F}_y := \{x \in Y : (x, y) \in \mathcal{F}\}. \tag{4.1}$$

Definition 6. *The set $\mathcal{F}$ is called a biconvex set, if $\mathcal{F}_x$ is convex for every $x \in X$, and if $\mathcal{F}_y$ is convex for every $y \in Y$.*

Definition 7. *A function $f : \mathcal{F} \to \mathcal{R}$ on a biconvex set $\mathcal{F}$ is called a biconvex function if*

$$f_x(\blacklozenge) := f(x, \blacklozenge) : \mathcal{F}_x \to \mathcal{R}$$

is a convex function on $\mathcal{F}_x$ for every fixed $x \in X$ and

$$f_y(\blacklozenge) := f(\blacklozenge, y) : \mathcal{F}_y \to \mathcal{R}$$

is a convex function on $\mathcal{F}_y$ for every fixed $y \in Y$.

Definition 8. *An optimization problem of the form*

$$\underset{x,y}{\text{minimize}}\{f(x, y) : (x, y) \in \mathcal{F}\} \tag{4.2}$$

is said to be biconvex optimization problem, if the feasible set $\mathcal{F}$ is biconvex and the objective function is biconvex on $\mathcal{F}$.

One of the most traditional ways to tackle such a problem is the Block-Coordinate Decent (BCD) multi-convex method (or, Alternate Convex Search (ACS)), where the feasible set and objective function are convex in each block of variables. Mainly, it minimizes the original objective cyclically over a certain block through a convex solver, by keeping the remaining ones fixed [104]. It guarantees convergence to a stationary point, that could be a saddle point, a local or global optimal [105]. Nevertheless, *strictly quasi-convexity*, *pseudo-convexity* [106] or *Geometric Programming (GP) transformation* [107] of the original cost function can guarantee

the uniqueness of global optimum. Note that, such a decomposition, dates back to 1950 [108] and is closely related to the Gauss-Seidel and SOR methods for linear equation systems.

Eventually, we stress the fact that that the analysis above can be extended to "multi-convex" sets, objective functions and optimization problems when we have to deal with a higher number of optimization (sets of) variables.

4.3 System Model and Assumptions

In this Chapter our focus is on the BS loads ρ (that direct user associations) as well as ζ and Z. We remind to the reader, that ζ_i is the *access resource allocation parameter* for BS i (see B.2). In particular, ζ_i reflects the amount of radio resources (e.g., time, frequency, space) available for downlink transmissions. Without loss of generality, we now focus on time resources, as e.g., in the context of the envisioned flexible TDD standard. Although traditional LTE systems only allow some fixed and predefined values for ζ_i (depending on the TDD configuration), we relax them to be more generally applicable. Likewise, $Z(j)$ reflects the corresponding portion of downlink resources for the j-th link (see C.2).

While in Chapters 2 and 3 we assumed these splits to be fixed, in this Chapter we want to consider their flexible adaptation (depending on the current traffic loads and system dynamics) as an additional degree of freedom. Such a suppleness shall affect our system model for the radio access network in a twofold manner.

(B.11 - BS load under flexible ζ) Our first observation is that the load of BS i is strongly coupled by the tuning of resources ζ_i. For instance, when a BS decreases by two the downlink resources (i.e., when $\zeta_i \to \frac{\zeta_i}{2}$), the corresponding downlink utilization shall double (i.e., $\rho_i^D \to 2 \cdot \rho_i^D$), given that the serving traffic loads remain unchanged (see B.1 - B.5). To make this relation explicit, we define our next assumption.

In the remaining of the chapter we will use the load variables ρ_i^D to portray the load corresponding to BS i when all resources are used for DL (similarly for UL). We are interested in the flow-level dynamics of this system, and model the service of DL flows at each BS as a queueing system with effective load (or utilization) $\frac{\rho_i}{\zeta_i}$.

(B.12 - UL/DL cross interference avoidance) Secondly, cross-interference can be generated in neighboring BSs with UL/DL overlapping slots. For instance, if for BS i, the $\zeta_i = 80\%$ of its time is scheduled for DL, while a neighbor BS j, $1 - \zeta_j = 70\%$ is scheduled for UL, we can clearly see that at least half (50%) of their scheduled slots will be cross-interfered. This threatens SINR (due to the increased interference, especially the DL-to-UL interference) and invalidates our assumption B.3. To make this constraint explicit, we define the next assumption.

Without loss of generality, we assume that each BS i cross interferes with a subset of other BSs $\mathcal{C}_i \subseteq \mathcal{B} \setminus \{i\}$. In practice, a distance based rule, or alternatively the cell cluster concept, can be used to determine these sets. If i is on the DL and a BS $j \in \mathcal{C}_i$ on the UL (or vice versa) then these BSs might cause severe interference to each other (that invalidates assumption B.3). We refer to this as *cross interference*. A sufficient condition to avoid cross-interference is

$$\rho_i^D + \rho_j^U \leq 1, \forall i \in \mathcal{B}, j \in \mathcal{C}_i. \tag{4.3}$$

We explain the above condition here. Consider two such BSs i and j. If $\zeta_i = \zeta_j$ then there is no cross-interference, because i and j can synchronize their DL (and UL) slots to avoid it. If

$\zeta_i \neq \zeta_j$, cross-interference might occur, but *it also depends on the effective loads.* ζ_i slots are *at most* used for DL. But out of these only $\frac{\rho_i^D}{\zeta_i} \cdot \zeta_i = \rho_i^D$ will be busy (since $\frac{\rho_i^D}{\zeta_i}$ is the utilization of the downlink resources, according to B.5-B.7). The rest of the DL slots $(1 - \frac{\rho_i^D}{\zeta_i}) \cdot \zeta_i = \zeta_i - \rho_i^D$ could be blanked with ABS frames (see also Fig. 4.2). Similarly, the percentage of slots that j will be *active* on the UL is $\frac{\rho_j^U}{1-\zeta_j} \cdot (1-\zeta_j) = \rho_j^U$ slots. Hence, if $\frac{\rho_i^D}{\zeta_i} \cdot \zeta_i + \frac{\rho_j^U}{1-\zeta_j} \cdot (1-\zeta_j) \leq 1$, there are enough different slots in a frame to schedule all DL and UL of i and j without any overlap. Taking care for all such links on the interference graph, gives us Eq.(4.3). Finally, we stress that this constraint applies to the long-term allocation policy of resources. The actual MAC scheduling may still allocate resources in those time slots to transmissions that are non-interfering.

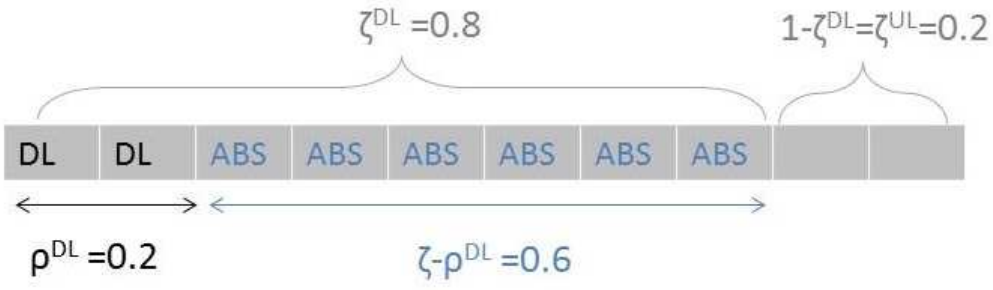


Figure 4.2: A frame example for a certain BS where the total sub-frames are allocated between downlink and uplink. The sub-frames on the DL (or, uplink) that are not busy, can be muted through ABS.

As for the backhaul network, we claim that cross interference is not a major issue, and add the following assumption.

(C.5 - Interference-free Backhaul) Modern backhaul architectures are developed using (highly) directional P2P or P2MP static architectures [109]. These are planned topologies and thus cross interference between BH links with asymmetric UL/DL schedules can be considered negligible. However, our one can apply similar assumptions and constraints in scenarios where backhaul cross interference is under scrutiny, as did in (B.12).

Since we have already seen and analyzed the tradeoff between Split and Joint UL/DL association, *in the following we only focus on the Split association scenario that is able to achieve the highest performance in both dimensions simultaneously.*

4.4 TDD Access Allocation and User Association Optimization

We start our discussion by ignoring the backhaul network (assuming it is provisioned), and attempt to solve the (i) *user association*, and (ii) *access resource allocation* problems, jointly. More specifically, we are interested in finding the optimal values for the variable ζ_i and $\rho_i^D, \rho_i^U, \forall i \in \mathcal{B}$. In Section 4.4.1 we define the corresponding optimization problem. In Section 4.4.2 we sketch a convergent algorithm that decomposes it in smaller problems that can be efficiently tackled as shown in Section 4.4.3.

4.4.1 Feasible set, Objective Function and Optimization Problem 3.

The feasible region for our problem can be delimited by the requirement that the effective load of no BS being exceeded (see B.5).

Definition 9. ***(Feasible set)*** *If ϵ is an arbitrarily small positive constant, the feasible region of $(\rho^D;\rho^U;\zeta) = ((\rho_1^D,\rho_2^D,\ldots,\rho_{\|\mathcal{B}\|}^D);(\rho_1^U,\rho_2^U,\ldots,\rho_{\|\mathcal{B}\|}^U);(\zeta_1,\zeta_2,\ldots,\zeta_{\|\mathcal{B}\|}))$ is*

$$\mathcal{F} = \Big\{(\rho^D,\rho^U,\zeta) \mid \rho_i^y = \int_{\mathcal{L}} p_i^y(x)\rho_i^y(x)dx, \tag{4.4a}$$

$$\sum_{i\in\mathcal{B}} p_i^y(x) = 1, \tag{4.4b}$$

$$0 \le p_i^y(x) \le 1, \ \ \forall x \in \mathcal{L}, \ \ y \in \{U,D\}, \tag{4.4c}$$

$$0+\epsilon \le \zeta_i \le 1-\epsilon, \tag{4.4d}$$

$$0 \le \frac{\rho_i^D}{\zeta_i}, \frac{\rho_i^U}{1-\zeta_i} \le 1-\epsilon,, \ \ \forall i \in \mathcal{B}, j \in \mathcal{C}_i\Big\} \tag{4.4e}$$

Lemma 4. *The feasible set $\mathcal{F}$ is convex.*

Proof. The proof for the feasible set $\mathcal{F}$ without the last two constraints can be found in [10]. Constraints (4.4d are linear, and constraint (4.4e) refers to the image of ρ under different perspectives. So they preserve convexity [92], and the complete feasible set remains convex. □

Following Chapter 2 we extend the proposed objective to also include the resource allocation variables $\zeta_i, \forall i \in \mathcal{B}$ (see B.2).

Definition 10. ***(Objective function)*** *Our objective is*

$$\phi_\alpha(\rho,\zeta) = \sum_{i\in\mathcal{B}} \tau \frac{(1-\frac{\rho_i^D}{\zeta_i})^{1-\alpha^D}}{\alpha^D-1} + (1-\tau)\frac{(1-\frac{\rho_i^U}{1-\zeta_i})^{1-\alpha^U}}{\alpha^U-1}, \text{if } \alpha^D,\alpha^U \neq 1. \tag{4.5}$$

If α^D is equal to 1, the respective fraction must be replaced with $\log(1-\frac{\rho_i^D}{\zeta_i})^{-1}$.

Lemma 5. *The objective function $\phi_\alpha(\rho,\zeta)$ is a biconvex function, i.e., it is convex in ρ for fixed ζ, and versa.*

Proof. The objective function is the sum of the basic α function $\frac{(1-\frac{\rho}{\zeta})^{1-\alpha}}{\alpha-1}$ over different BSs, with $(\rho,\zeta) \in \mathcal{F}$. When ζ is fixed this is the simplest form of the well known α-fair function which is clearly convex in ρ. And so is the corresponding sum over all BSs (sum preserves convexity). For fixed ρ, the basic α function is also convex in ζ (it has non-negative second derivative, namely $2\rho\zeta^{-3}(1-\rho/\zeta)^{-\alpha} + \alpha\rho^2\zeta^{-4}(1-\rho/\zeta)^{-\alpha-1} \ge 0$), and so does its sum. □

Definition 11. ***(Optimization Problem 3)*** *The joint user association and radio resource allocation problem can be expressed*

$$\begin{aligned} &\underset{\rho,\zeta}{\text{minimize}} \left\{\phi_\alpha(\rho,\zeta) | (\rho,\zeta) \in \mathcal{F}\right\}, \\ &\text{subject to } \rho_i^D + \rho_j^U \le 1, \forall i \in \mathcal{B}, j \in \mathcal{C}_i. \end{aligned} \tag{4.6}$$

Lemma 6. *Problem 3 is a biconvex minimization problem.*

Proof. This is a biconvex optimization problem since the objective function is biconvex on the (bi)convex feasible set $\mathcal{F}$, and the constraints are affine functions. □

Our aim is to minimize $\phi_\alpha(\rho, \zeta)$ i.e. our α-fair objective function tuned by the various resource allocation parameters ζ_i subject to the various cross interference constraints. Since we are interested in the Split DL/UL scenario, there is no need consider additional constraints.

4.4.2 Decomposition Algorithm for Optimization Problem 3.

Our nonconvex objective is block separable in ρ^D, ρ^U. Indeed, if we fix ζ, the problem decomposes in two simpler problems with variables ρ^D and ρ^U, that are coupled from constraint (4.3), and so we call ζ the *complicating* variable. Therefore, it makes sense to decompose the objective into two levels of optimization, following the *primal decomposition method* [110]. Specifically, at the lower level there are *two subproblems* that run in parallel, that aim to find the optimal values of ρ^{*D} and ρ^{*U}, namely $\rho = [\rho^{*D}; \rho^{*U}]$, upon a fixed ζ. At the higher level we encounter the *master problem*, where we attempt to update (and eventually optimize), the complicating variable ζ. Note that constraint (4.3) only depends on ρ and thus does not affect the master problem. Formally, the subproblems and the master problem are

$$\min_{\rho}\{\phi_\alpha(\rho, \zeta)\} \text{ subj. to Eq.(4.3)} \qquad \text{(sub-problems)} \tag{4.7}$$

$$\min_{\zeta}\{\phi_\alpha(\rho, \zeta)\} \qquad \text{(master problem)} \tag{4.8}$$

The above decomposed problems are convex since Optimization Problem 3 is biconvex (see Lemma 6). Thus, they can efficiently be tackled through convex optimizers.

Our proposed iterative algorithm is sketched in Alg. 1. At the (k) iteration step the master problem allocates the available resources by directly giving each subproblem the amount of resources that it can use ($\zeta^{(k)}$ for the DL and $(1 - \zeta^{(k)})$ for the UL traffic), as it is usually done in primary decomposition methods [Algorithm 1 line 3]. Then, we solve the two subproblems (derive ρ^{*D}, ρ^{*U}) based on their given resources and the coupling constraint [Algorithm 1 line 5-6]. In the next iteration $(k + 1)$, we update the complicating parameter (derive $\zeta^{(k+1)}$), and re-solve the two subproblems. We repeat the process until the stopping criterion [Algorithm 1 line 1] is satisfied, and $\zeta^{(k)}$ converges to a stationary point ζ^*. Usually, such a stopping criterion is of the form of $\|\zeta^{(k)} - \zeta^{(k-1)}\| < \epsilon$, where ϵ is an arbitrary small positive constant. Convergence and stability are guaranteed if the two subproblems are solved on a faster timescale than the higher level master problem, so that at each iteration of a master problem both subproblems at a lower level have already converged [110]. In Section 4.4.3.1 we show how one can derive the optimal values ρ^*, whereas in Section 4.4.3.2 the sequence $\zeta^{(k)}$.

Lemma 7. *Algorithm 1 converges to the global optimal point of Problem 3.*

Proof. Our proposed decomposition algorithm falls into the category of Alternate Convex Search (ACS) [103, 105], that is a special case of the popular Block Coordinate Decent method [104]. There, starting from an initial feasible point, one attempts to minimize the objective by cyclically iterating through the different optimization directions with respect to one coordinate direction at a time. Precisely, in our case at the end of the k iteration it is

$$\phi_\alpha(\rho, \zeta^{(k)}) \prec \phi_\alpha(\rho, \zeta^{(k-1)}).$$

Algorithm 1 Decomposition Sketch of Optimization Problem 3.

1: **Repeat** until $\|\zeta^{(k)} - \zeta^{(k-1)}\| < \epsilon$.
2: *(Update the master problem (Section (4.4.3.2)).)*
3: Resource allocation: $\zeta \rightarrow$ DL, $1 - \zeta \rightarrow$ UL.
4: *(Solve the two subproblems (Section (4.4.3.1)).)*
5: Derive ρ^{*D} given the available resources (ζ).
6: Derive ρ^{*U} given the available resources ($1 - \zeta$).

This will continue until convergence to a stationary point, where the gradient vanishes and the above inequality approaches equality. ACS algorithms in its simplest form suggest that the stationary point could be a saddle point, a local or global optimal [105]. However, Alg. 1 guarantees convergence to the global optimum due to the following two points.

(1) *Uniqueness of optimum point:* Optimization Problem 3 can be converted to a geometric programming (GP) problem, since both its objective and constraints can be written as a sum of posynomials terms composed of positive monomials, according to the transformation in [107]. Such problems have a single optimum. (The GP equivalent form of our problem is not convenient for decomposition, so we use this argument only to prove uniqueness, but not to solve the joint problem.)

(2) *Saddle point escape:* Our proposed algorithm can escape from potential saddle points, as discussed in Section 4.4.3.2. □

4.4.3 Subproblems and Master Problem.

We now analytically discuss how the decomposed (convex) problems defined in Eq. 4.7 and 4.8 can be tackled.

4.4.3.1 Subproblem Optimization (Eq. (4.7))

We present here the DL subproblem only. The UL problem is symmetric. An discussed, an efficient way to tackle the coupling constraints in a distributed implementation setup is to directly include the constraints in the objective as *penalty functions* that increase the objective when a cross-interference constraint is violated [92]. We can then solve the new *unconstrained* problem

$$\underset{\rho}{\text{minimize}} \left\{ \Phi(\rho, \zeta) = \phi_\alpha(\rho, \zeta) + \gamma \sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{C}_i} \mathcal{I}_{ij} (\rho_i^D + \rho_j^U - 1)^2 \right\}, \tag{4.9}$$

where $\mathcal{I}_{ij}$ is the indicator variable that reveals whether BS i cross interferes with BS j. Specifically (see also B.12)

$$\mathcal{I}_{ij} = \begin{cases} 1, & \text{when } \rho_i^D + \rho_j^U > 1 \\ 0, & \text{otherwise.} \end{cases} \tag{4.10}$$

Parameter γ can be again chosen as a large constant, introducing a "soft" constraint (i.e., cross-interference could be slightly exceeded, if this really improves our main objective), or be increased progressively, so as to converge to a "hard" constraint [98].

Theorem 4.1. *If $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ denotes the optimal load vector, the optimal DL association rule for location x is*

$$i^D(x) = \arg\max_{i\in\mathcal{B}} \left(\underbrace{c_i^D(x)}_{\text{user knowledge}} \cdot \overbrace{P_i^D}^{\text{BS broadcast message}} \right) \tag{4.11}$$

where

$$P_i^D = \frac{\zeta_i \cdot \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D}}{1 + 2\gamma \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D} \sum\limits_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{*D} + \rho_j^{*U} - 1)}.$$

Proof. Problem (4.9) is convex. Let ρ^* be its optimal solution. A sufficient condition for optimality is if $\langle \nabla\Phi(\rho^*), \Delta\rho^* \rangle \geq 0$ for all $\rho \in \mathcal{F}$, where $\Delta\rho^* = \rho - \rho^*$. To write the remaining of the proof compactly with respect to the coupling constraints, we denote (only within the proof) $\zeta^D = \zeta, \zeta^U = 1 - \zeta$, $I(D) = \mathcal{I}_{ij}, I(U) = \mathcal{I}_{ji}$ and assume that L is either D or U ($L \in \{D, U\}$) with complementary value $\tilde{L}$. Let $p(x)$ and $p^*(x)$ be the associated routing probability vectors for ρ and ρ^*, respectively. Using the deterministic DL and UL cell coverage generated by (4.11) the respective optimal rules are $p_i^{*L}(x) = \mathbf{1}\left\{ i = i^L(x) \right\}$.

Then, the inner product $\langle \nabla\phi\left(\rho^*\right), \Delta\rho^* \rangle$ is equal to

$$\sum_L \sum_{i\in\mathcal{B}} \left(\frac{1}{\zeta_i^L \left(1 - \frac{\rho_i^{*L}}{\zeta_i^L}\right)^{\alpha^L}} + 2\gamma \sum_{j\in\mathcal{C}_i} I(L)(\rho_i^{*L} + \rho_j^{*\tilde{L}} - 1) \right) (\rho_i^L - \rho_i^{*L}) =$$

$$\sum_L \sum_{i\in\mathcal{B}} \left(\frac{1 + 2\gamma \left(1 - \frac{\rho_i^{*L}}{\zeta_i^L}\right)^{\alpha^L} \sum\limits_{j\in\mathcal{C}_i} I(L)(\rho_i^{*L} + \rho_j^{*\tilde{L}} - 1)}{\zeta_i^L \left(1 - \frac{\rho_i^{*L}}{\zeta_i^L}\right)^{\alpha^L}} \right) \cdot \int_{\mathcal{L}} \rho_i^L(x) \left(p_i^L(x) - p_i^{*L}(x) \right) dx =$$

$$= \int_{\mathcal{L}} \sum_L \frac{\lambda^L(x)}{\mu^L(x)} \sum_{i\in\mathcal{B}} \left(\frac{1 + 2\gamma \left(1 - \frac{\rho_i^{*L}}{\zeta_i^L}\right)^{\alpha^L} \sum\limits_{j\in\mathcal{C}_i} I(L)(\rho_i^{*L} + \rho_j^{*\tilde{L}} - 1)}{\zeta_i^L c_i^L(x) \left(1 - \frac{\rho_i^{*L}}{\zeta_i^L}\right)^{\alpha^L}} \right) \cdot \left(p_i^L(x) - p_i^{L*}(x) \right) dx.$$

Note that in the DL i.e. $L = D$ (similarly in UL)

$$\sum_{i\in\mathcal{B}} p_i^D(x) \left(\frac{1 + 2\gamma \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D} \sum\limits_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{*D} + \rho_j^{*U} - 1)}{\zeta_i c_i^D(x) \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D}} \right) \geq$$

$$\sum_{i\in\mathcal{B}} p_i^{D*}(x) \left(\frac{1 + 2\gamma \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D} \sum\limits_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{*D} + \rho_j^{*U} - 1)}{\zeta_i c_i^D(x) \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D}} \right)$$

holds because $p_i^{*D}(x)$ is an indicator for the minimizer of $\frac{1 + 2\gamma \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^L} \sum\limits_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{*D} + \rho_j^{*U} - 1)}{\zeta_i c_i^L(x) \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D}}$. So, $\langle \nabla\Phi(\rho^*), \Delta\rho^* \rangle \geq 0$. □

When the interference constraints for the BS i are not violated (i.e., $\mathcal{I}_{ij} = 0, \forall j \in \mathcal{C}_i$), the above rules state that the optimal downlink associations are the same as the one in [10]. However, when the BS i cross interferes with another BS, an additional term is added in the denominator that penalizes BS i making it less preferable to users at location x. Note that the amount of penalization depends on the amount of *total* cross interference (sum term) from nearby BSs. This penalization makes sense, since additional users to that BS would increase its effective load as well as its DL busy slots, and thus increase the cross interference. Eventually, as γ increases penalties will be monotonically decreased until they totally vanish, and we end up to a feasible point (as explained in Chapter 3 for backhaul constraints).

4.4.3.2 Master Problem Update (Eq.(4.8))

Descent methods suggest:

$$\zeta^{(k+1)} = \zeta^{(k)} + t^{(k)} \Delta\zeta^{(k)}, \tag{4.12}$$

such that $\phi(\rho^*, \zeta^{(k+1)}) < \phi(\rho^*, \zeta^{(k)})$, where $\Delta\zeta^{(k)}$ is a *descent direction*, and $t^{(k)}$ a *step size*. The master step update for ζ could be performed centrally (e.g. at an SDN controller), or distributed (e.g., each BS calculates their optimal update independently from each other).

Nevertheless, since our objective is differentiable, we chose to apply the *Newton method* that provides the steepest descent direction in local Hessian norm, in order to speed up convergence. We also apply *backtracking line search* that determines the maximum amount to move along the search direction [92]. We start with a relatively large estimate of the step size, and iteratively shrink it until a decrease of the objective function is observed that adequately corresponds to the decrease that is expected, based on the local gradient. Finally, when stationarity is reached, we ensure that this is not a saddle point through a "noisy" gradient criterion: a noise vector with mean 0 is added to the gradient direction of stationary points that provably pushes them away from saddle points [111].

4.5 TDD Access/ Backhaul Allocation and User Association Optimization

Here, we investigate the complete problem setting of optimizing the (i) *user association*, (ii) *access* and (iii) *backhaul resource allocation* jointly. Specifically, we include to our aims the optimal derivation for the variables $Z(j), \forall j \in \mathcal{B}_h$, by adding to our feasible set $\mathcal{F}$ the convex constraints discussed in C.4, and $0 < Z(j) < 1$. Since the algorithmic sketch is the same as the one in Section 4.4, we skip the details here. And we immediately present the Optimization Problem 4, and the corresponding convergent decomposition algorithm.

Definition 12. ***(Optimization Problem 4)*** *User association and Access and Backhaul Resource Allocation can be expressed as*

$$\begin{aligned} &\underset{\rho,\zeta,Z}{\text{minimize}} \left\{ \phi_\alpha(\rho,\zeta) \quad | (\rho,\zeta,Z) \in \mathcal{F} \right\}, \\ &\text{subject to} \qquad \rho_i^D + \rho_j^U \leq 1, \forall i \in \mathcal{B}, j \in \mathcal{C}_i, \\ &\text{subject to} \qquad \sum_{i \in \mathcal{B}(j)} \frac{\rho_i^D \tilde{c}_i^D}{Z(j) \cdot C_h^D(j)} < 1, \ \forall j \in \mathcal{B}_h. \end{aligned} \tag{4.13}$$

Lemma 8. *Problem 4 is multi-convex optimization problem.*

Proof. This is a multiconvex optimization problem since the objective function and the constraints are multiconvex on the (multi)convex feasible set $\mathcal{F}$. □

To keep the complexity decreased, decomposition optimization theory suggests to hierarchically decouple these three problems (related to ρ, ζ, Z) in three different optimization levels [110], as depicted in Algorithm 2. In the two lower levels [Algorithm 2 lines 5], we encounter Algorithm 1. There, as discussed previously the coupled (i) access resource allocation and (ii) user association problems (referred as secondary master problem and subproblems, respectively) are solved. In the highest level [Algorithm 2 line 3], we encounter the master problem of (iii) backhaul resource allocation, where we update Z, upon the convergence of the two lower level problems until the stopping criterion, that now considers Z, is satisfied. In such multi-level optimizations convergence and stability are guaranteed if the lower level master problem is solved on a faster timescale than the higher level master problem, so that at each iteration of a master problem all the problems at a lower level have already converged. Note that, one can alternate the decomposition order and tackle the three problems in different timescales, based on his freedom degrees.

Algorithm 2 Decomposition Sketch of Optimization Problem 4.

1: **Repeat** until $\|Z^{(l)} - Z^{(l-1)}\| < \epsilon$.
2: *(Update the master problem and increase γ (Section 4.5.1).)*
3: Backhaul Resource allocation: $Z \rightarrow$ DL, $(1 - Z) \rightarrow$ UL.
4: *(Update the secondary master problem and solve the subproblems (Section 4.5.1, 4.5.2).)*
5: Run Algorithm 1.

Similarly to Algorithm 1, Algorithm 2 converges to the global optimal point of the Optimization Problem 4. Now, it remains to show whether the additional backhaul constraints change the different structural subalgorithms, and if so how.

Our first observation is that these backhaul constraints are function of Z, our third optimization variable. Thus, in order to (i) keep being amenable on distributed user association implementations and (ii) be able to sketch a gradient step for Z update, we still want to tackle these constraints by including them in the objective through the *penalty function* method. And then, to solve the corresponding unconstrained problem(s).

4.5.1 Secondary master and master problem

The update of access (ζ) and backhaul resource allocation Z can be realized using the same intuition as in Section 4.4.3.2. However, note that for mesh backhaul tree topologies, the update

of Z can only be achieved either (i) centralized in an SDN controller, or (ii) using a distributed SDN controller environment upon allowance for coordination with the links of the same backhaul path.

4.5.2 Subproblem optimization

The unconstrained minimization objective is:

$$\underset{\rho,\zeta,Z}{\text{minimize}} \Bigg\{ \phi_\alpha(\rho,\zeta,Z) + \gamma \sum_{i\in\mathcal{B}} \sum_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^D + \rho_j^U - 1)^2 + \gamma \sum_{k\in\mathcal{B}_h} \mathcal{J}^D(k) \left(\frac{\sum\limits_{i\in\mathcal{B}(k)} \rho_i^D \tilde{c}_i^D}{Z(k)\cdot C_h^D(k)} - 1 \right)^2 + \mathcal{J}^U(k) \left(\frac{\sum\limits_{i\in\mathcal{B}(k)} \rho_i^U \tilde{c}_i^U}{(1-Z(k))\cdot C_h^U(k)} - 1 \right)^2 \Bigg\} \tag{4.14}$$

where $\phi_\alpha(\rho,\zeta)$ is the basic α-fair objective function (see Definition 5) and the second term illustrates the penalties for the interference constraints (defined in Eq. 4.3). The last term refers to the penalties for the backhaul constraints (defined in Eq. 3.2). $\mathcal{J}^D(k)$ is the indicator variable that reveals whether the k-th backhaul link is congested ($\mathcal{J}^D(k)$=1) or not ($\mathcal{J}^D(k)$=0). Precisely,

$$\mathcal{J}^D(k) = \begin{cases} 0, & \text{when } \frac{\sum_{i\in\mathcal{B}(k)} \rho_i \tilde{c}_i}{Z(k) C_h(k)} < 1 \\ 1, & \text{otherwise.} \end{cases} \tag{4.15}$$

The optimal user association rules that update ρ follow.

Theorem 5.1. *If $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ denotes the optimal load vector, the optimal DL association rule for location x is*

$$i^D(x) = \arg\max_{i\in\mathcal{B}} \left(\underbrace{c_i^D(x)}_{\text{user knowledge}} \cdot \overbrace{P_i^D}^{\text{BS broadcast message}} \right) \tag{4.16}$$

where

$$P_i^D = \frac{\zeta_i \cdot \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D}}{1 + 2\gamma \left(\sum\limits_{k\in\mathcal{B}_h(i)} \tilde{c}_i^D \frac{\mathcal{J}^D(k)}{Z(k)\cdot C_h(k)} \left(\frac{\sum\limits_{l\in\mathcal{B}(k)} \rho_l^{*D} \tilde{c}_l^D}{Z(k)\cdot C_h(k)} - 1 \right) + \sum\limits_{l\in\mathcal{C}_i} \mathcal{I}_{ij} \cdot (\rho_i^{*D} + \rho_j^{*U} - 1) \right)}.$$

Starting within a feasible point ρ and using increasing values for γ, these rules can be iteratively applied and will eventually converge to the optimal point ρ^*.

Proof. We now prove that the above rule indeed minimizes Optimization Problem 4. The proof follows very similar steps with the case addressed in the previous Section where backhaul was

assumed to be provisioned. Thus, we immediately proceed to the inner product $\langle \nabla\Phi(\rho^*), \Delta\rho^* \rangle$

$$\sum_{L=\{D,U\}} \sum_{i\in\mathcal{B}} \left(\frac{1}{\zeta^L \left(1 - \frac{\rho_i^{*L}}{\zeta_i^L}\right)^{\alpha^L}} + 2\gamma \sum_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{*D} + \rho_j^{*U} - 1) + 2\gamma \sum_{j\in\mathcal{B}_h(i)} \mathcal{J}^L(j) \left(\frac{\sum_{k\in\mathcal{B}(j)} \rho_k^{*L} \tilde{c}_k^L}{(Z^L(j)\cdot C_h(j))^2} \tilde{c}_i^L - \frac{\tilde{c}_i^L}{Z^L(j)\cdot C_h(j)} \right) \right) \cdot$$
$$\cdot (\rho_i^L - \rho_i^{*L}) =$$

$$= \int_L \frac{\lambda^L(x)}{\mu^L(x)} \sum_{L=\{D,U\}} \sum_{i\in\mathcal{B}} \left(\frac{1 + 2\gamma \sum_{k\in\mathcal{B}_h(i)} \tilde{c}_i^D \frac{\mathcal{J}^D(k)}{Z(k)\cdot C_h(k)} \left(\frac{\sum_{l\in\mathcal{B}(k)} \rho_l^{*D} \tilde{c}_l^D}{Z(k)\cdot C_h(k)} - 1 \right) + 2\gamma \sum_{l\in\mathcal{C}_i} \mathcal{I}_{ij} \cdot (\rho_i^{*D} + \rho_j^{*U} - 1)}{\zeta_i^L c_i^L(x) \left(1 - \frac{\rho_i^{*L}}{\zeta_i^L}\right)^{\alpha}} \right) \cdot$$
$$\cdot \left(p_i^L(x) - p_i^{L*}(x) \right) dx \geq 0,$$

due to the corresponding maximizers of p_i^{*D}, p_i^{*U}. □

Note that, when the capacity constraints for the backhaul links k are not violated (e.g., $\mathcal{J}^D(k) = 0$), the above rules state that the optimal associations are the same as the one in Eq. (4.11). However, when it becomes congested, an additional term is added in the denominator that penalizes that BS making it less preferable to UEs at location x. Note that this penalty considers the whole backhaul path $\mathcal{B}_h(i)$ that traffic from BS i traverses, and adds a penalty for *every* link along that path that is congested (outer sum in the denominator). Overall, these rules provide the optimal association for a user at x, by optimally weighting (i) access network performance, (ii) cross interference avoidance, and (iii) backhaul congestion.

Remark 1. Note that these rules are still "device centric" and maintain all the desired properties in this context (scalable, simple and flexible performance).

Remark 2. We analytically proved that the rules for optimal user association in DL and UL can be applied in a parallel fashion independently from each other (despite the coupling constraint (4.3)), and convergence to the global optimum is guaranteed. This takes some support from the ACS method, when it comes to the distributed gradient. At each iteration step: instead of *cyclically optimizing* each block of variables, we now suggest to *cyclical update* them based on the local distributed gradient until convergence to the single optimum.

Remark 3. The hierarchical decomposition can be done by a number of different decomposition orders and all would converge to the global optimum under the mentioned certain circumstances (e.g., as discussed in Algorithm 1 or 2). Specifically, upon n optimization problems there are $n!$ (factorial of n) possible decomposition orders; for us this would be $3! = 2 * 3 = 6$ different decomposition orders. However, we believe that the proposed decomposition order lends itself to a natural implementation between different network elements. User association is proposed to run in the fastest timescale to adapt to the high traffic fluctuations across different locations and users. The load of a single BS depends on the sum of its attached users and is subject to fewer fluctuations. It only has to react to (slower) traffic shifts of the aggregate loads, by updating its ζ parameter accordingly. Finally, a backhaul link further aggregates the traffic of multiple BS, and can update its optimal allocation at an even slower timescale.

4.6 Simulations

In this section, we evaluate our proposed algorithms on example scenarios, and discuss related insights.

Since there are multiple non-trivial tradeoffs, we will now consider two network topologies. Firstly, we will consider a simple scenario with only one macro BS and three SCs, in order to better elucidate the qualitative behavior of our algorithm compared to standard practices, as well as better trace its performance benefits and where these come from. We then consider a larger network scenario and demonstrate that similar benefits can be observed there as well. Nevertheless, we are going to re-evaluate our previous schemes (analyzed in Chapters 2 and 3) with the fixed resource allocation policies, for comparison reasons.

Scenario 1: We consider a 2×2 km^2 area. Fig. 4.3 shows a color-coded map of the heterogeneous traffic demand $\lambda(x)$ (*flows/hour* per unit area) with 3 hotspots (blue implying low traffic and red high). We assume that this area is covered by three SCs (referred with BS numbers 1-3), and one macro cell (BS number 4). Without loss of generality, we assume that each SC offloads its traffic through a dedicated backhaul link (corresponding BH link numbers 1-3) to the macro BS, and that the macro BS cross interferes with all SCs (i.e., $C_4 = \{1, 2, 3\}$, $C_1 = C_2 = C_3 = \{4\}$, see B.9). We set $\alpha^D = \alpha^U = 1$ to optimize user throughput.

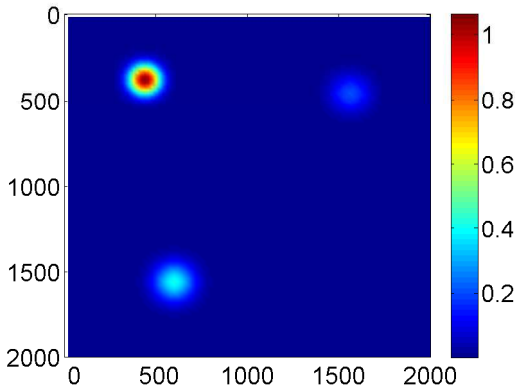


Figure 4.3: Traffic Arrival Rate (blue colour implying low traffic and red colour implying high traffic demand).

Coverage Snapshots: We first look at the coverage maps that different schemes create. Figure 4.4(a), 4.4(b) depict the optimal user associations for fixed LTE-TDD configuration 1 that assumes static UL/DL timeslot ratio 4 : 4 i.e., fixed $\zeta_i = 0.5, \forall i \in \mathcal{B}$. Similarly for the BH links $Z(j) = 0.5, \forall j \in \mathcal{B}_h$. As a first note, we see that in DL most users are associated with the macro BS, and a few to SCs (macro BS attracts more DL users due to the higher transmit power). In the UL, users tend to form Voronoi cells (to minimize path loss and improve UL SINR). Secondly, we note that the DL coverage areas of the various SCs are decreased according to the corresponding traffic arrival intensity: e.g. SC 1 that serves the most intense hotspot see Fig. 4.3 has the smallest coverage area, while SC 3 which sees lower traffic intensity has the largest). The main reason is that the SCs have limited DL backhaul capacities that force some users to the far away macro BS. This alleviates the backhaul link congestion but hurts overall

performance. At the same time, a high amount of the pre-configured UL backhaul resources might remain wasted (due, to asymmetry in DL/UL traffic intensity for example).

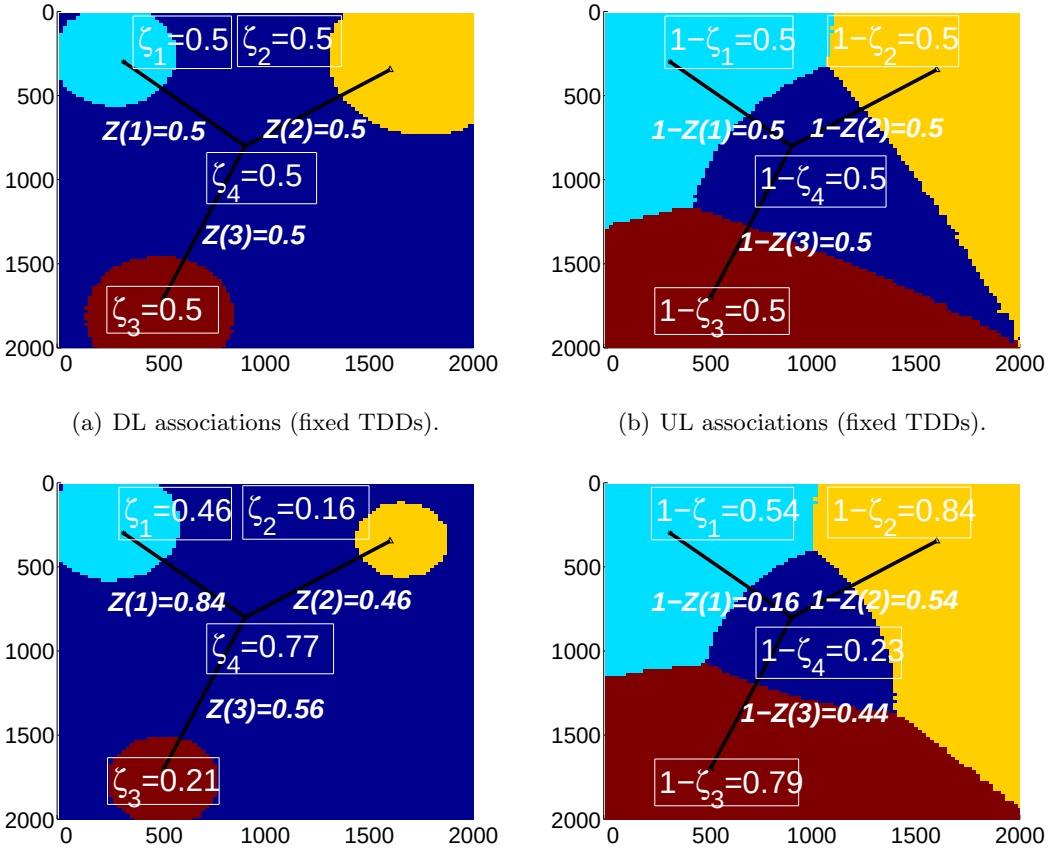
Summarizing, the observed coverage maps for this scenario demonstrate two possible shortcomings of pre-configured TDD: (a) asymmetry in the DL/UL coverage areas and corresponding transmit powers suggest that a TDD allocation other than 50-50% could improve performance; (b) some (usually DL) user associations could be suboptimal, dictated by backhaul capacity limitations arising from the preconfigured fixed allocation on the BH, even if the total BH resources would suffice for the sum of both UL and DL traffic.

To explore these possibilities, we now relax the allocation variables ζ and Z (see B.2 and C.2) and apply our proposed algorithm. Clearly, in this simple example, a single-step improvement in either direction described above ((a) or (b)) could improve performance. We remind the reader that our proposed algorithm goes beyond this single step, alternating between optimizing coverage maps and TDD resource allocation, until it finds the best possible combination. The resulting coverage maps (i.e. optimal ρ values) and radio/BH allocations (optimal ζ and Z values) are shown in Fig. 4.4(c), 4.4(d). We first note that macro BS increases its $\zeta_4 = 0.77$ to serve more DL users, and SC increase their UL resources $1 - \zeta_1 = 0.54, 1 - \zeta_2 = 0.84, 1 - \zeta_3 = 0.79$ to serve more UL, bewaring to avoid *cross interference*. Interestingly, such an allocation simultaneously improves both UL and DL performances (we will explicitly show this later). Also, the DL BH allocated resources ($Z(j)$) are increased to accommodate more DL traffic, while ensuring not to exceed a maximum value that would congest the UL.

User performance: We now go beyond the above qualitative behavior and evaluate the quantitative benefits. We first focus on user-centric performance and consider various τ values (we remind the reader that τ is a parameter that balances the importance of DL vs UL performance). We compare the performance of the following main schemes. (*ProposedAlg*): our proposed algorithm; (*TDD Fixed*): the optimal allocation algorithm of [95] with equal, preconfiigured UL/DL resources on both radio access and BH. To better understand the importance of considering the cross-interference and BH capacity constraints, we also include results for the following schemes. (*AlgNoCross*): jointly optimal allocation, but not taking cross-interference into account. If there is an eventual asymmetry in the optimal UL/DL schedules, potential cross-interference is included in the SINR to capture its impact. (*AlgNoBH*): jointly optimal allocation without considering the backhaul constraints. Here, we assume that all BSs associated with a BH link that is congested decrease their performance proportionally to the amount of congestion.

In Figures 4.5 and 4.6 we depict the DL and UL user throughput as a function of τ in different scenarios. It is easy to see that our *ProposedAlg significantly outperforms* the *TDD fixed* policy by up to $2.5 - 3\times$. What is more, for most intermediate τ values, it is able to simultaneously improve both DL and UL performance. As τ increases further, the emphasis of *ProposedAlg* moves exclusively to the DL (and vice versa) which is consistent with our expectations, unlike the fixed TDD scheme where DL and UL performances are optimized independently of τ (decoupled objective).

Regarding the impact of the cross interference constraint, *AlgNoCross* can still offer some improvement on the DL for $\tau > 0.5$, compared to the baseline (*TDD Fixed*). However, it does so with a significant penalty on UL performance (up to $3\times$ worse), which is the most sensitive to cross-interference (this DL-to-UL interference is a key problem for future Flexible TDD [112]). This underlines the importance of directly considering cross interference constraints



(a) DL associations (fixed TDDs).

(b) UL associations (fixed TDDs).

(c) DL associations (flexible TDDs).

(d) UL associations (flexible TDDs).

Figure 4.4: Optimal user associations in the following scenarios: (a) downlink associations when fixed TDD resource allocation (50%-50%) (b) uplink associations when fixed TDD resource allocation (50%-50%), (c) downlink associations with flexible TDD resource allocation, (d) uplink associations with flexible TDD resource allocation. ($\tau = 0.5$).

in our optimization framework through Eq.(4.3). Finally, the performance of *AlgNoBH* shows similar behavior, where it can sometimes provide better performance for the DL or the UL (compared to *TDD fixed*) but not both.

Summarizing, the following important conclusions can be drawn from the above analysis: (a) jointly optimal allocation of user association and DL/UL radio resources can actually lead to considerable performance degradation, unless cross-interference is taken explicitly into account; (b) a jointly optimal allocation, even with cross-interference taken into account, might still be quite suboptimal, if the DL/UL resources on the BH are not also optimized to conform to the new load requirements imposed by the BSs; (c) joint optimization of all these dimensions is feasible, and can offer significant performance improvement for both DL/UL.

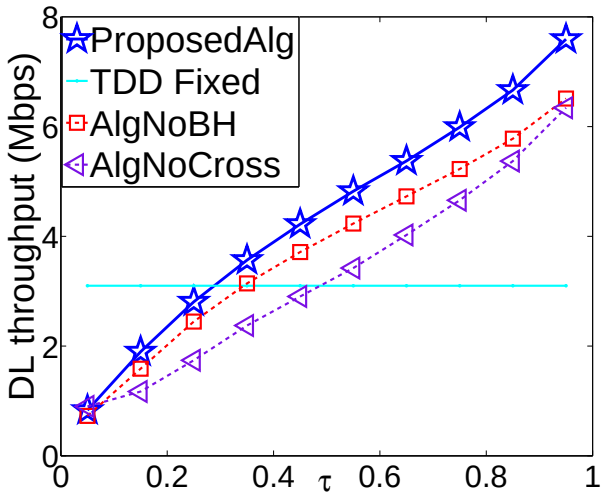


Figure 4.5: User-centric Performance: Downlink user throughput for various fixed and flexible TDD resource allocation schemes (in Mbps).

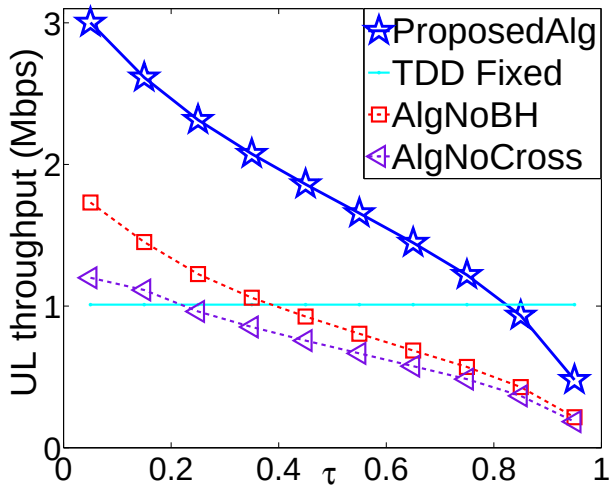


Figure 4.6: User-centric Performance: Uplink user throughput for various fixed and flexible TDD resource allocation schemes (in Mbps).

Network performance. Table 4.1 considers the performance improvements in the same comparison scenario (*ProposedAlg* and *TDD Fixed* [95]), but now from the network perspective when $\tau = 0.5$. We consider two metrics: Spectral Efficiency (SE) in terms of bits/s/Hz, and Load Balancing (LB) in terms of mean square error between different BS loads, similar to what is assumed in [95]. DL/UL spectral efficiency improve up to 44% since *flexible TDD better allocates the resources* with respect to the heterogeneous transmit powers that help physical data rates improve (see B.2-B.3). It also considers related traffic statistics and asymmetries across users (see A.1-A.2) by diminishing the BS load fluctuations (e.g., BS under/over utilizations) and thus LB is improved. It is interesting to note that simultaneous improvement of these metrics implies improvement in user performance, as showed previously and explained in B.7.

Table 4.1: Network performance in terms of Spectral Efficiency (SE) and Load Balancing (LB) (when $\tau = 0.5$)

	Downlink		Uplink	
Performance.	SE	LB	SE	LB
Percentage % of improvement.	42	16	44	54

Scenario 2: Having highlighted the different tradeoffs and sources of performance improvement in the basic scenario above, we now turn our attention to a larger network topology consisting of 4 macro BSs and 13 SCs. Without loss of generality, we now consider uniform traffic demand. Considerable performance improvements can be observed in this scenario as well as seen in Table 4.2 (e.g. 86% better UL user performance). Relative lower improvement values compared to the smaller Scenario 1 are mainly due to: (a) not all BSs experience bad performance now so even if *ProposedAlg* considerably improves the performance of the problematic BSs, average performance is not as affected; (b) the *additional cross interference constraints* posed from the neighboring clusters.

Table 4.2: User and network performance improvements in terms of Throughput, Spectral Efficiency and Load Balancing (when $\tau = 0.5$)

	Downlink			Uplink		
Scenario.	UE	SE	LB	UE	SE	LB
Percentage % of improvement.	29	39	4	86	42	51

4.7 Discussion and Conclusions

In this Chapter, we formulated a novel algorithm that carefully studies the coupled problems of (i) user association, TDD (ii) access, and (iii) backhaul resource allocation under the emerging *backhaul* and *cross interference* constraints. Using optimization theory we proved that under certain circumstances it converges to the global optimum. Simulation results corroborate the correctness of our framework and reveal promising qualitative and quantitative results, e.g., simultaneous and significant performance improvements on both UL and DL dimensions.

Note that the proposed algorithm of this Chapter (Algorithms 1 and 2) supplement the frameworks analyzed in Chapter 2 and 3, by including additional degrees of freedom. We highlight the importance of these degrees of freedom, since as showed can reveal various performance improvements up to 3 times.

Chapter 5

Energy Optimizations subject to user QoS constraints.

5.1 Introduction

As discussed, the growing demand for Internet-enabled wireless devices, and bandwidth-hungry multimedia services from the increasing number of "heavy" users and smartphones create significant capacity problems. The latter motivated the operators to build very dense deployments that suffer with high load fluctuations. Thus, in Chapters 2, 3 and 4 we claimed that user association and dynamic TDD allocation problems need to be revisited in such scenarios, and we offered some initial insights about the involved tradeoffs.

Note that, the intense spatio-temporal fluctuations usually imply that a significant number of BSs will carry no traffic or only a low traffic-load, by questioning their actual contribution. Currently, 15-20% of all sites carry about 50% of the total traffic [113]. Hence, a considerable number of sites waste energy (for staying ON, as well as for cooling), despite serving little or no traffic [114]. Chapter 5 is devoted for this problem.

As pointed out in Chapter 1, a large research effort has been initiated recently in the area of "green" networks. Nevertheless, most past studies are performed in the context of large macrocells under homogeneous traffic profiles, and with large time-scales (e.g. turning off BSs during the night [87]). Furthermore, usually simple QoS requirements are considered when applying such techniques, e.g. signal quality as in [86], or traditional blocking probabilities as in [17]. In modern and future cellular networks, dealing with energy consumption issues becomes more challenging. Significantly more opportunities arise for switching off BSs in smaller time scales (e.g. in the order of some minutes), due to (a) coverage overlaps stemming from heterogeneous and/or independent deployment of cells, (b) larger spatio-temporal load variations due to the smaller number of users associated to each cell, and (c) power-proportional and load-dependent BSs. Yet, exploiting such opportunities must be done without violating agreed QoS performance for users. The evaluation of the latter is a rather daunting task, due to the diversity of user traffic (streaming, voice, web, file download, etc.) and service and performance requirements offered to users. As a result, a number of interesting questions arise: Which QoS metric(s) should be used in such future HetNets? Which types of users and BSs should one consider when making a power management decision? Should the duration of switching-off period, affect our decision, and if so, how?

Towards answering these questions, in this Chapter we identify three QoS constraints, related to different ways that the performance of a UE could deteriorate [115]. We then derive analytically the probability of violating each of them, as a function of user and network parameters and planned switch-off duration. Specifically, we consider the following **BCD** constraints:

- ***B**locking probability* for dedicated flows, i.e., the probability that a flow that requires a certain amount of (dedicated) bandwidth, is blocked due to the lack of the available resources (Section 5.2.2).
- ***C**overage Failure Probability*, i.e., the probability that a random UE experiences poor signal quality when it needs to use the network (e.g. making a call, or sending a web request). (Section 5.2.1).
- ***D**elay* for best-effort flows, i.e., the ongoing delay for the flows that are multiplexed and have to compete for resources. (Section 5.2.3).

Our general methodology is to, first, identify the key parameters for each QoS constraint, and then use analytical tools, mostly coming from queueing theory, to evaluate the probability of violating each one of them, if a BS is switched-off. Our goal in this direction is to strike a tradeoff between realistically capturing some features of new, data-centric cellular systems, while maintaining a certain analytical tractability to provide insights into the QoS vs. Energy savings. The novelty of our methodology is that we can select even a small time-interval, for the sleeping period X, and evaluate the energy-QoS tradeoff by switching to transient analysis (rather than stationary analysis) of the stochastic model in hand (Section 5.2). Based on these QoS constraints and the time duration X, we perform a preliminary study and show that significant energy savings can be achieved even for switching-off periods of the order of some minutes (Section 5.3).

5.2 System Model and Problem Formulation

Since we are interested in short sleeping mode durations, i.e. in the short-term (rather than long-term) dynamics of the system, the standard stationary analysis presented in Chapters 2, 3, 4 as well as the ones proposed in the literature can become incorrect. Thus, we present a new system model and corresponding assumptions that conform to the needs of our new requirement for short-time scale analysis.

We will follow a similar presentation as did in the previous Chapters. Specifically, in (A.1 - A.2) we present our system model related to the user and traffic differentiation, as well as the radio access network. In ((B1)CD- (B3)CD), B(C1)D- B(C3)D) and (BC(D1)- BC(D2))) we present our model for each one of the three QoS constraints.

Note that our focus is on the radio access network, since BSs are the main energy killers in a cellular network [116], and our major concern is on the DL (similar analysis can be done for UL).

(A.1 - User Classification) Our first observation is that in the short-term different users will affect the considered constraints differently. For example, ensuring good signal quality for a UE that has some *ongoing* traffic (e.g. doing a VoIP call, or streaming a video), is more important (and more challenging) than for a UE currently "on" the network but idle. If the former

Table 5.1: Notation

Variable	Meaning
X	Duration of the switch-off period.
z^d, z^b	Probability that a random flow is dedicated or best-effort.
ζ_i^d, ζ_i^b	Resource allocation parameter for dedicated or best-effort flows.
p_f	Threshold for Failure Probability while switching-off BSs (Proposition 1).
p_{block}	Threshold for Blocking Probability while switching-off BSs (Proposition 2)
D_{max}	Threshold for service delay while switching-off BSs (Proposition 3)
$R_{\mathrm{b}}, R_{\mathrm{total}}$	Available peak bit rates for "dedicated'" flows, for "best effort" flows.
$\lambda_{\mathrm{AU}}, \lambda_{\mathrm{CU}}, \lambda_{\mathrm{DU}}$	Data rates for AU,CU,DU.
B, $1/\mu^b$	Dedicated flows demand (in bps), and best-effort flow length (in bits).

experiences poor signal quality, the communication session might be dropped immediately. We consider three different types of users:

- **Active users** (AU): users that are connected to a BS *and* have one or more on-going traffic flows currently. These users reside in EMM (EPS-Mobility Management) REGISTERED and ECM (EPS Connection Management) CONNECTED states.
- **Connected users** (CU): users associated with a BS but without any ongoing traffic sessions[1]. These users are in EMM REGISTERED and ECM IDLE states.
- **Disconnected users** (DU): users *in the vicinity* of the BS, but currently not *ON* (or in airplane mode); while their exact number and location cannot be known their impact should be estimated, especially when the switch-off duration increases, as one of them might decide to switch on the UE and use the network. These users are in the EMM DEREGISTERED and ECM IDLE states.

(A.2 - Traffic Differentiation and Bandwidth Allocation) In addition to the above classification of users, we also need to classify the flows between dedicated and best-effort, since different flows shall affect differently the transient cell load and thus our decision for switch-off. As discussed in Section 2.2, the probability that the next flow generated by a user is a dedicated or best effort flow depends on the aggregate traffic mix (e.g., percentage of VoIP calls vs. video streaming vs. simple browsing, etc.). We maintain the same notation (z^d vs. z^b, respectively). For simplicity, we assume that each BS has a peak data rate R_{total} to allocate among all flows from all serving users, and $0 < \zeta_i < 1$ is the spiting parameter such that: $R_{\mathrm{d}} = \zeta_i \cdot R_{total}$ and $R_{\mathrm{b}} = (1 - \zeta_i) \cdot R_{total}$ are allocated to dedicated and best effort flows, respectively[2].

Our aim is to decrease energy consumption of this cellular network, by dynamically switching off one or more of small cells, during a defined time-duration of X minutes, referred to as

[1]For simplicity, we will ignore background traffic as it usually less delay-sensitive, and often "lightweight" (e.g. email client polling, social network notifications, etc.).

[2]The actual values are operator-specific, which is why in our analysis it is considered as an input parameter. Note also that, depending on the deployment, the available rate R_{total} might not be bounded by the radio access capacity, but rather by the backhaul capacity, as discussed in Chapter 3.

switch-off period, hereafter[3]. To do so, we shall avoid violating each one of the following three constraints.

In the following we start with the more generic constraint i.e. the Coverage Failure probability constraint since is easier to tackle. Then, we formulate the blocking probability and service delay constraints, by modifying accordingly the k-loss and PS system.

5.2.1 Coverage Constraint

When the decision to switch off a BS with users is made, those users will have to be handed-over to an available neighboring BS. This will often result into a weaker than average signal level. Hence, before a decision to switch off a target BS is made, we must ensure that it will not lead to a disconnection or unacceptable quality for one or more handed-over users. To this end, as our first QoS constraint we will consider the probability that a user, originally associated with a switched-off BS, will experience low-signal quality (e.g. a deep fade) *if* it needs to use the network during the switch-off period.

It turns out this probability changes for different types of users, namely AU, CU, and DU. Specifically, an AU with a current ongoing session will be immediately affected by a signal quality drop. In contrast, a CU or DU will be affected only if *both* the following events occur: (i) it becomes active (e.g. initiates a new call or data session) during the switch-off period X, and (ii) the signal quality is low. Consequently, we need to calculate the following quantities:

- the *outage probability*, which is the probability that the signal strength of user is not sufficient to maintain an ongoing service,
- the *activation probability*, which is the probability that a user covered by the BS in question (e.g. a CU or a DU) becomes active during the next X minutes, and
- the *coverage failure probability*, which depends on both the outage (AU, CU, DU) and activation probabilities (CU, DU), and is the quantity we are interested in.

(B(C1)D - Outage probability) For simplicity, we use the SNR to calculate the outage probability[4]. Following Chapter 2, we assume that the SNR for the l^{th} UE associated with the j^{th} BS, is given by:

$$\text{SNR}_{lj} = \frac{G_{lj} R_{lj} p_j}{N_0}. \tag{5.1}$$

The noise power is denoted as N_0, and the transmission power of the j^{th} BS is p_j. G_{lj} represents the nonnegative path loss between the j^{th} BS and the l^{th} UE (it may also encompass antenna and coding gains) that is often modeled as proportional to r_{lj}^{-n} (n is the power fall-off factor and r_{lj} denotes distance). Note that now, we include the Rayleigh fading, as usually done in outage probability consideration (e.g., see [86]). R_{lj} corresponds to a Rayleigh fading component, and is exponentially distributed with unit mean. The distribution of the received power from the j^{th} BS at the l^{th} UE is then exponentially distributed with mean value $E[G_{lj} R_{lj} p_j] = G_{lj} p_j$.

[3]"Idle" power consumption (related to both electronics, but also cooling) is a major component one could thus save. The additional "load-dependent" power consumption would essentially be shifted over to a neighboring base station, not leading to significant further gains. We defer exploring more complex power management techniques (e.g. cell zooming [117]) to future work.

[4]The use of SINR could also be introduced in this constraint, but would make our analysis more complex.

Thus, the outage probability for the l^{th} AU or CU associated with the j^{th} BS is:

$$\mathrm{P}_{\mathrm{out}}(r_{lj}) = P(\mathrm{SNR}_{lj} < \gamma) = 1 - e^{-\frac{\gamma N_0}{G_{lj} p_j}} = 1 - e^{-\frac{\gamma N_0}{r_{lj}^{-n} p_j}}. \tag{5.2}$$

The above formula is applicable for AUs and CUs, as their actual distance r_{lj} is known. In the case of DUs, their location and the total number is unknown. Assuming that there are ρ_{DU} DUs per m^2, and the transmission range of a base station is r_{max}, the expected number of DUs in the considered cell is:

$$N_{DU} = \rho_{DU} \pi r_{max}^2. \tag{5.3}$$

If we now consider a specific DU that becomes active, and whose "local" BS is switched off, it will try to connect to one of the neighboring cells. Let r_d denote the distance of the chosen BS from the local BS (its mean value is a function of deployment density). Thus, we can replace r_{lj} in Eq.(5.2) with r_d to get an estimate for the DU outage probability:

$$P_{out}^{DU} = \left(1 - e^{-\frac{\gamma N_0}{(r_d)^{-n} p_j}}\right). \tag{5.4}$$

(B(C2)D - Activation Probability) We now consider the probability that a CU or DU becomes active during the next X minutes. We denote these probabilities as $P_{act}^{CU}(X)$ and $P_{act}^{DU}(X)$, respectively. For simplicity, we assume that the time until a CU or a DU generates a new session (call, data session, etc.) is exponentially distributed with rate λ_{CU} and λ_{DU}, respectively (we assume $\lambda_{DU} \leq \lambda_{CU}$). Hence, we can calculate the activation probabilities as follows:

$$P_{act}^{CU}(X) = 1 - e^{-\lambda_{CU} X}, \tag{5.5}$$
$$P_{act}^{DU}(X) = 1 - e^{-\lambda_{DU} X}. \tag{5.6}$$

The above equations can be easily extended to general user session interarrival distributions. However, Poisson arrivals are often assumed for user-initiated sessions [14].

(B(C3)D Coverage Failure Probability) Assume that the candidate BS serves N_{AU} active and N_{CU} connected users. We denote the set of active and connected users as $\mathcal{N}_{AU}$ and $\mathcal{N}_{CU}$, respectively, and we assume that some DUs are also in the covered region, whose number is given by Eq.(5.3). If the BS is switched off, then let $J(i)$ denote the BS that user i is handed-over to[5]. Finally, assume that the desired QoS is described by a maximum failure probability p_f, chosen by the operator or indicated in a Service Level Agreement (SLA). Then, the following Proposition captures the first system constraint:

Proposition 1. (Constraint I) *A BS cannot be switched off if the average user associated with it will experience a coverage failure probability, during the switch-off period X, that exceeds a threshold p_f. This probability is given by*[6]*:*

[5]Note that in the real system, this is done using RSSI and RSRP measurements coupled with the received system information assuming that the terminal is eligible. In our analysis, we will assume that either maximum SNR or, simply, distance is used as the criterion.

[6]Instead of this weighted average approach, one could also consider a very conservative, worst-outage probability minimization approach, that has been considered in [118], using the Perron-Frobenius theorem.

$$\frac{\sum_{i\in\mathcal{N}_{AU}} P_{out}(r_{iJ(i)}) + \sum_{i\in\mathcal{N}_{CU}} P_{act}^{CU}(X)P_{out}(r_{iJ(i)}) + N_{DU}P_{act}^{DU}(X)P_{out}^{DU}}{N_{AU}+N_{CU}+N_{DU}}. \tag{5.7}$$

Impact of switch-off duration X: The above analysis gives qualitative insight about the impact of the switch-off duration. If X is short, compared to the average inactivity time for CUs (DUs), one can more aggressively switch off BSs as a smaller percentage of node is affected. However, for large X Eq.(5.5) and (5.6) converge to 1. In that case, all users in the vicinity of a BS must be considered, and the decision only depends on the average outage probability.

5.2.2 Dedicated flows: Blocking Probability Constraint

In this subsection, we focus on dedicated flows. Specifically, we are interested in the impact of switching off a BS on the admission control mechanism of the neighboring BSs, where users with dedicated flows will have to be handed over.

As discussed in Chapter 2, for such flows we consider the k-loss system. There, a BS is allocated a finite set of k resources (or, servers). If a user initiates a new session (e.g., call) when the BS is already using all its k resources, this session will be *blocked*. We are interested in calculating this blocking probability, and trying to keep it lower than a pre-defined threshold.

While in stationary systems (e.g., as the ones defined in Chapters 2,3 and 4) this probability can be given by the well-known Erlang-B formula [14], for systems that have not yet converged this formula can become incorrect.

To that end, we now focus on (i) the arrival/service rate under the system model discussed earlier, and (ii) explain how k can be approximated in such a system where our focus is on the short-term statistics. Then, (iii) by using Transient Analysis theory we calculate the desired blocking probability and construct our QoS constraint. In the remaining discussion, when we consider a given BS, we use the term "handed over" or "remote" (sub-/superscript "HO") to refer to users that have been "transferred" to this BS from a neighboring BS that is switched off, and "local" (sub-/superscript "l") for existing users of this BS. Handed-over users are generally further away from the BS than local ones.

((B1)CD - Arrival and service rate of dedicated flows) Consider a given BS being switched off, whose users are handed over to (different) neighboring BSs. Consider now one of this neighboring BSs, and let us denote as N_i^l, and N_i^{HO} the number of local and remote users, respectively, of type i ($i \in \{AU, CU, DU\}$), associated with this BS. Let further λ_i denote the flow arrival rate *per user* of type i. The total load for this BS is the sum of all flow rates across these users, and we'll assume that the actual arrival process is Poisson with the sum rate. This assumption is motivated by the Palm-Khintchine theorem, which states that the sum of many independent arrival processes becomes Poisson in the limit [14]. Finally, assume that each arriving flow requires dedicated resources with a probability z^d. Then, due to Poisson splitting, the arrival process remains Poisson with total rate λ_d, given by:

$$\lambda_d = z^d \left(\lambda_{AU}\cdot(N_{AU}^l + N_{AU}^{HO}) + \lambda_{AU}\cdot \sum_{i\in\{CU,DU\}} (N_i^l + N_i^{HO})\cdot \lambda_{AU} P_{act}^i(X) \right), \tag{5.8}$$

where $P_{act}^i(X)$ are the activation probabilities defined in (5.5) and (5.6). We will also assume that the flow sizes are exponentially-distributed with parameter μ_d, that is, approximately, the

average one between dedicated flows. Thus, we can replace λ and μ in the Markov chain of Fig. 5.1, with λ_d and μ_d, for the case of dedicated flow admission control.

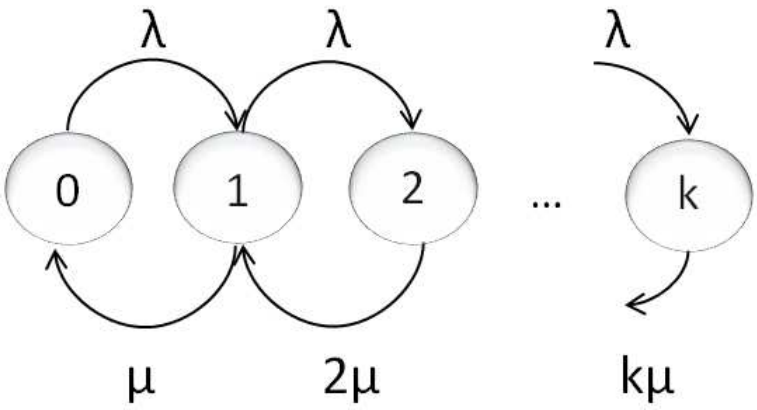


Figure 5.1: Continuous Time Markov chain (CTMC) for the considered k-Loss queueing system of the dedicated flows being served at a given base station.

((B2)CD - Resource constraint k) As explained earlier, k is the resource constraint in the k-loss system related to countable resources (e.g., servers). In the context considered, the available peak rate for dedicated flows R_{d}, is a flexible resource, whose allocation is a function of the number of flows, respective dedicated rate demand, and user channel quality. Thus, we apply a "softer", estimated value of k in our loss system.

Assume that B_d is the average bit rate demand per dedicated flow. If a peak rate R_d is available at the BS, the resource constraint k could be approximated in the k-loss system as $\frac{R_d}{B_d}$, since an "average" flow consumes a percentage $\frac{B_d}{R_d}$ of the available rate. However, this nominal peak rate is only available when the SNR is ideal, or more simply, within a certain distance from the BS (assuming e.g., a simple log-distance path loss model). Hence, to better estimate the maximum number of "average" dedicated flows that can be served, we need to also consider the (potential) distances of different users generating these flows. For this purpose, we adopt a simple and low-complexity model associating peak rate to distance, proposed in [76,119], stating that the peak rate available drops with distance r_{ij} from a BS j as:

$$c(r_{lj}) = \begin{cases} 1, & r_{lj} \leq r_0 \\ (\frac{r_0}{r_{lj}})^n, & \text{otherwise} \end{cases} \tag{5.9}$$

where r_0 is some threshold range within which the maximal rate is obtained, and n, is the attenuation factor.

Hence, if all dedicated flows where requested from a distance r_{lj}, then the total rate available to them would be only $c(r_{lj})R_d$ ($\leq R_d$), or stated differently, the effective rate requirement per average flow would be higher at a large distance r and given by

$$B(r_{lj}) = \frac{B_d}{c(r_{lj})}. \tag{5.10}$$

We can now approximate the peak rate drop factor $c(r)$ based on the combination of UEs and distances, e.g. using again a weighted average. Specifically, our estimated resource constraint k for dedicated flows is given by

$$k = \frac{R_d}{\tilde{B}_d}, \tag{5.11}$$

where

$$\tilde{B}_d = \frac{\sum_{l=1}^{N_{AU}} B(r_{lj}) + \sum_{m=1}^{N_{CU}} B(r_{mj})}{N_{AU} + N_{CU}}, \tag{5.12}$$

and $N_i = N_i^l + N_i^{HO}$, $i \in \{AU, CU\}$, denotes the total number of users, local and remote, of type i. We can thus replace k with the approximated value of k in the loss system of Fig. 5.1[7]. Finally, note that we have assumed that DUs will not affect the peak data rate of the considered BSs (but only affect this constraint through Eq.(5.8), where we assume that DUs might also switch on and generate some flows during X).

((B3)CD - Transient analysis of k-loss system) So far, we have shown how to calculate the necessary parameters for the Markov chain of Fig. 5.1. However, to calculate the probability that a newly arrived flow that needs dedicated resources will be blocked, it does not suffice to replace these parameters in the Erlang B formula. The latter gives the *stationary* blocking probability, that requires the respective chain to be converged, and thus corresponds to large values of X. Instead, we need to apply *transient analysis* to this system, and estimate the blocking probability via the *occupation time* in state k during the intended switch-off duration X.

The initial state for the Markov chain, at time 0 (the beginning of the switch-off period), corresponds to the current number of active dedicated flows. Denoted as s, it is:

$$s = z^d \cdot (N_{AU}^l + N_{AU}^{HO}) \cdot \xi, \tag{5.13}$$

where we use ξ to denote the expected number of ongoing flows per AU (this is an input parameter). Starting from s, the occupation time in state i, denoted as $O_i(X)$ $(0 \leq i \leq k)$, is the time that the MC spends in state during the next X minutes (or time units). We are interested in deriving the quantity $\frac{E[O_k(X)]}{X}$. This corresponds to the percentage of time that the system is in state k (all resources are used), during the switch-off period X and starting from state s. Hence, due to the PASTA (Poisson Arrivals See Time Averages) property, this also corresponds to the probability that a newly arrived dedicated flow will be blocked due to non-available capacity.

[7]We should stress that, as mentioned earlier, this is only an estimate. In practice, there will be a few times when the system is serving more than k dedicated flows (e.g. when all users are close-by or flows require lower rates than average), and times when a new flow might be blocked even if less than k flows are served. However, since we are interested in the short time-scale statistics where the system dynamics do not change significantly, we believe that this chance is quite small.

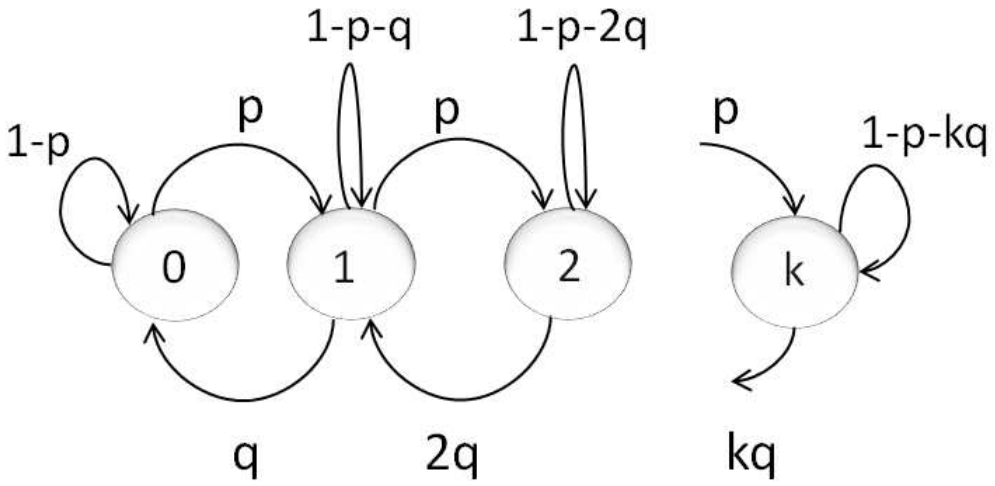


Figure 5.2: Discrete Time Markov chain (DTMC) for the considered k-Loss queueing system of the dedicated flows being served at a given base station.

We can estimate this percentage of time either by *uniformization* in CTMC (Continuous Time Markov Chain) [120], or by converting it to DTMC (Discrete Time Markov Chain), as an approximation. To simplify our discussion, we follow the second approach. Let Δt be a small time inteval. The DTMC depicted in Fig. 5.2, is the discrete-time approximation of our system continuous-time k-loss system, where a state transition occurs every Δt time units. If P_{ij} denotes the probability that the chain goes from state i to state j $(0 \leq i, j \leq k)$, then it follows from standard properties of the Poisson distribution [120] that (see Fig. 2):

$$p = \lambda_d \cdot \Delta t, \quad P_{i,i+1} = p, \;\; 0 \leq i < k$$

$$q = \mu_d \cdot \Delta t, \quad P_{i,i-1} = i \cdot q, \;\; 0 < i \leq k$$

$$P_{i,i} = 1 - P_{i,i+1} - P_{i,i-1}, \;\; 0 \leq i \leq k$$

Hence, if $\mathbf{P} = \{P_{i,j}\}$ denotes the probability transition matrix, and $\mathbf{P}^n = \{P^n_{i,j}\}$ the n-step transition matrix ($\mathbf{P}^n = (\mathbf{P})^n$), then the expected occupation time is given by:

$$E[O_k(X)] = \sum_{n=0}^{\frac{X}{\Delta t}} P^n_{\mathrm{s,k}}, \tag{5.14}$$

where s denotes the initial state (initial number of active dedicated flows) and $\frac{X}{\Delta t}$ the total switch-off duration (counted in discrete time steps of duration Δt).

Proposition 2. (Constraint II) *Assume a desired maximum blocking probability is given for dedicated flows, defined as p_{block}. A given BS can be switched off only if the following inequality holds for all neighboring BSs to which users of the switched-off BS are handed-over:*

$$\frac{\sum_{n=0}^{\frac{X}{\Delta t}} P^n_{s,k}}{X/\Delta t} \leq p_{block}. \tag{5.15}$$

Impact of switch-off duration X. The computational complexity of Proposition 2 can be traded off with accuracy by increasing the time step Δt. In addition, as X becomes large (specifically, larger than the mixing time for the MC of Fig. 5.2), the condition of Eq.(5.15) converges to the Erlang-B formula.

Remark 1. When $X \to \infty$, the condition of Eq. (5.15), converges to

$$\frac{(\frac{\lambda_{\mathrm{d}}}{\mu_d})^k/k!}{\sum_{j=0}^{k}(\frac{\lambda_{\mathrm{d}}}{\mu_d})^j\frac{1}{j!}} \leq p_{block}. \tag{5.16}$$

Proof.

$$\begin{aligned}\lim_{\Delta t\to 0}\lim_{X\to\infty}\frac{E[O_k](X)}{X/\Delta t} &= \lim_{\Delta t\to 0}\lim_{X\to\infty}\frac{\sum_{n=0}^{\frac{X}{\Delta t}}P_{s,k}^n}{X/\Delta t}\\ = \lim_{\Delta t\to 0}\pi_k(\Delta t) &= \pi_k,\end{aligned}$$

where $\pi_k(\Delta t)$ is the stationary probability for state k in the DTMC approximation with unit step (Δt). As $\Delta t \to 0$ this quantity converges to π_k, the stationary probability of state k for the CTMC corresponding to the standard k-loss loss system of Fig.5.1, which is the Erlang B formula [14]. □

5.2.3 Best-effort flows: Service Delay Constraint

As our last constraint, we consider the delay for a best-effort flow, i.e. a flow that requires elastic resources. Consider again a given BS being switched-off, whose users are handed over to different neighboring BSs, and let's pick one of them and focus on it. As before, this BS will have some local users and some remote users, that were handed over from the switched-off BS, all of which might generate (new) best effort flows. While there are no guarantees for such flows, we might still want to keep their expected delay below a certain threshold. Our goal is to model and analytically bound this delay.

As discussed in Chapter 2 such best-effort flows are usually scheduled under a PS scheduling discipline. To analyze the delay of such a system we need to know λ_b, the arrival rate of best-effort flows, and μ_b the service rate for best-effort flows.

(BC(D1) - Arrival and service rate of best-effort flows) Let z^b denote the probability that a new flow arrival is best effort. It follows that incoming "best-effort" flows are Poisson distributed with total rate:

$$\lambda_b = \lambda - \lambda_d. \tag{5.17}$$

To find the service rate for best-effort flows, let R_{b} denote again the peak bit rate for best-effort flows. As explained before, if a *single* best effort flow exists in the system for a user at distance r_{lj}, then the actual bit rate received is only $c(r_{lj})R_{\mathrm{b}}$, where $c(r_{lj})$ is given by Eq.(5.9). The actual average peak rate is:

$$\tilde{R}_b = R_b \cdot \frac{\sum_{l=1}^{N_{AU}} c(r_{lj}) + \sum_{m=1}^{N_{CU}} c(r_{mj})}{N_{AU}+N_{CU}}, \tag{5.18}$$

where $N_i = N_i^l + N_i^{HO}$, $i \in \{AU, CU\}$. The above estimated rate corresponds to a single flow. If there are n total best-effort flows currently in the system, then PS would split this rate equally, and each flow would be served with a bit rate $\frac{\tilde{R}_b}{n}$.

To find the actual service rate μ_b of the PS queue, the number of *flows served per time unit* (note that this is not equal to $\tilde{R}_b$, which is just the effective bit rate), we also need to know the average length of best effort flows. If we assume that the sizes of the best-effort flows are exponentially distributed with mean Y_b, then $\mu_b = \frac{\tilde{R}_b}{Y_b}$. When the system is stationary, i.e. when X is quite large, the expected delay for a newly arriving flow corresponds to the delay of a PS queueing system:

$$E[D_b] = \frac{1}{\mu_b - \lambda_b} \quad (X \to \infty).$$

(BC(D2) - Transient analysis of PS system) However, for general values of X, the Markov chain corresponding to the PS system is not stationary. Thus, we must again apply transient analysis, assuming an initial state. Let s again denote the initial number of best-effort flows in the BS, at the beginning of the switch-off period X, where $s = z^b \cdot (N_{AU}^l + N_{AU}^{HO})\xi$, similar to Eq. (5.13). Consider now a new flow of size $1/\mu^b$ arriving at some time $t \in [0, X]$. The number of active best effort flows in the system that has to share the PS capacity with, is a random variable, denoted as n. Our approach will be to find the expected delay conditional on this value of n, and then take the average.

If our flow of size $1/\mu^b$ finishes transmitting while in state n (i.e. no new flows arrive and no existing flows finish), the service rate remains fixed at $R_n = \tilde{R}_b/n$ and the expected delay for this flow is $\frac{1/\mu^b \cdot n}{\tilde{R}_b}$. However, if a state transition occurs before all $1/\mu^b$ bits are transmitted, then the remaining bits will be transmitted at a lower ($\tilde{R}_b/(n+1)$) or higher rate ($\tilde{R}_b/(n-1)$), if a new flow arrived, or an existing finished, respectively. Let us denote as T_n the time spent in this state until the next transition. This time is exponentially distributed with rate $\lambda_b + \mu_b$, so $E[T_n] = \frac{1}{\lambda_b + \mu_b}$. Hence, putting everything together, we can define the following recursion to derive the (conditional) delay of a flow of $1/\mu^b$ bits finding another n ongoing flows when it arrives. R_n denotes the transmission rate at state n.

$$D^n(1/\mu^b) = \begin{cases} \frac{1/\mu^b}{R_n}, & \text{if } \frac{1/\mu^b}{R_n} \le E[T_n] \\ E[T_n] + D^{n+1}(1/\mu^b - R_n \cdot E[T_n]), & \text{if } \frac{1/\mu^b}{R_n} > E[T_n] \\ & \text{and } n \to n+1 \\ E[T_n] + D^{n-1}(1/\mu^b - R_n \cdot E[T_n]), & \text{if } \frac{1/\mu^b}{R_n} > E[T_n] \\ & \text{and } n \to n-1 \end{cases} \tag{5.19}$$

It is: $P(n \to n+1) = \frac{\lambda_b}{\lambda_b + \mu_b}$, and $P(n \to n-1) = \frac{\mu_b}{\lambda_b + \mu_b}$.

However, the actual number of initial active flows n at time t is also a random variable, which depends on the evolution of the system, starting at initial state s until time t. To find these probabilities, we will again use a DTMC approximation and n-step transitions as before. Since the procedure is symmetric as for the dedicated flows, we omit here the details and give the final result which is

$$E[D_b] = \sum_{t=0}^{X/\Delta t} \sum_{n=1}^{\infty} \frac{P_{s,n}^{t/\Delta t} D^n(1/\mu^b)}{X/\Delta t}, \tag{5.20}$$

In practice, we can add up only a finite number of terms in the inner sum, to reduce the calculations.

Proposition 3. (Constraint III) *Assume a desired maximum delay for best effort flows, D_{max}. A given BS can be switched off only if the following inequality holds for all neighboring BSs to which users of the switched-off BS are handed-over:*

$$\sum_{t=0}^{X/\Delta t} \sum_{n=1}^{\infty} \frac{P_{s,n}^{t/\Delta t} D^n(1/\mu^b)}{X/\Delta t} \leq D_{max}.$$

Impact of switch-off duration X. The computational complexity of Prop. 3 can be traded off with accuracy by increasing the step Δt. Also, as X becomes large, the individual probabilities of (5.20) converge to their stationary distribution.

5.3 Simulation Results

In this section we briefly present some numerical results and discuss some initial insights they offer.

To evaluate our QoS constraints, we consider a network composed of 120 small cells , and 2 macro cells that are uniformly distributed in an area of $45km^2$. The simulation parameters remain similar as the ones adopted in Section 2.4. Specific differences will be elaborated when necessary.

We assume that there are 500 AUs and CUs, plus 120 DUs. We also assume total peak rate R_{total}= 70 Mbps; average length and bit-rate for best-effort and dedicated flows $1/\mu^b = 20$ Kbytes and $B = 200$ kbps, respectively; coverage threshold $\gamma = 50dB$[8]; $\lambda_{\text{AU}}, \lambda_{\text{CU}}, \lambda_{\text{DU}}$ 10, 2, 1 and 0.1 flows/hour, respectively. Finally, the maximum number of concurrent users that each SC can handle is set to[9] 11.

We are interested in investigating how the different values of the predefined thresholds p_f (failure probability), p_{block} (blocking probability) and D_{max} (service delay) affect the portion of energy savings[10]. In Fig. 3(a), 3(b) and 4(a), we assume switching-off duration $X = 10$ minutes. Each figure contains two curves; the "top" curve corresponds to the portion of energy saved when we consider only a certain constraint active, while the "bottom" curve considers all constraints to be active, at fixed thresholds (when not explicitly mentioned, we assume them to be $p_f = 0.3$, $p_{block} = 10^{-3}$ and $D_{max} = 50$msec).

In the "top" curve of Fig. 3(a), on the x-axis we increase the p_f and plot the savings. It can be seen that, increasing the threshold (making the constraint less strict) increases savings, as it allows for more BSs to be switched off. For instance, we can save up to 68% for $p_f = 0.4$. As for the "bottom" curve, savings increase too, but less sharply, as the other two constraints can overrule the switch-off decision, especially for large p_f. For example, with $p_f = 0.4$ and the other two thresholds fixed, the energy savings can be up to 30%.

[8]The threshold γ is an input parameter and is chosen to ensure the coverage constraint with a relatively good signal quality.

[9]This number can vary, depending on the type of the small cell [121], and does not affect the blocking probability.

[10]This portion is equal to the energy we can save, divided by the energy needed for all BSs switched-on during X i.e. $\frac{E_{ALL}-E_{part}}{E_{ALL}}$, where E_{ALL} is the energy needed if all BSs are switched-on, and E_{part} is the (decreased) energy needed if we safely switch-off some BSs based on our policies.

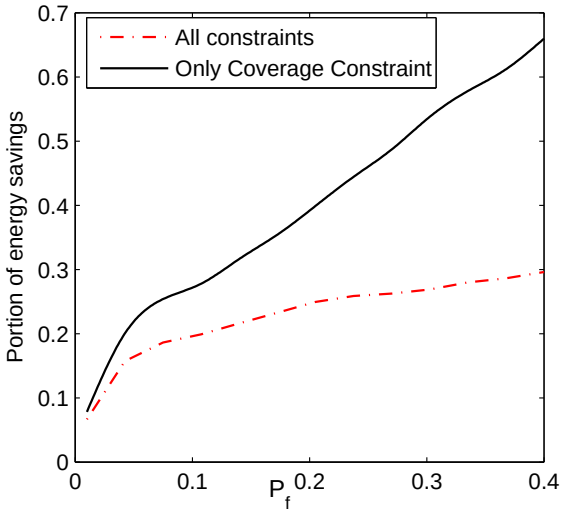


Figure 5.3: Portion of Energy Saving versus the Failure probability in the considered HetNet scenario when switching-off duration is $X = 10$ minutes.

Similarly, Fig. 3(b) and 4(a) depict the portion of the energy saved, by taking into account the blocking probability and service delay constraints. For example the top (bottom) curve of Fig. 3(b), shows that the portion of energy savings can be up to 50% (28%), by considering only the blocking probability constraint (plus the other two with fixed). Finally, Fig. 4(a) shows that the portion of energy savings for the delay constraint can be 70% by maintaining only the $D_{\max}$ in 100msec, and 30% by holding the other two fixed.

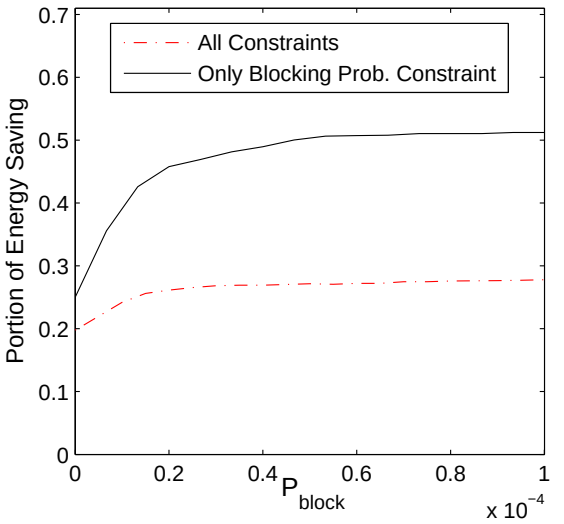


Figure 5.4: Portion of Energy Saving versus the Blocking probability in the considered HetNet scenario when switching-off duration is $X = 10$ minutes.

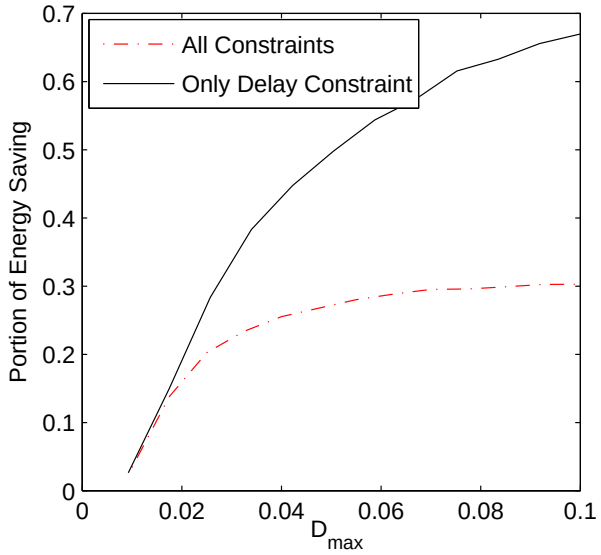


Figure 5.5: Portion of Energy Saving versus the Service delay (sec) in the considered HetNet scenario when switching-off duration is $X = 10$ minutes.

Another interesting parameter is X, the switch-off duration. Fig. 4(b) depicts the portion of energy saved for different values of X with fixed constraint thresholds ($p_f = 0.4$, $p_{block} = 10^{-3}$, $D_{max} = 200$msec). To be more precise, energy savings are maximum when X is relatively small, but start decreasing and eventually flatten out, as X increases. The reason is that, for small X, one needs to only consider the impact of AUs when evaluating the constraint and the impact of hand overs to neighboring BSs. However, as X increases, there is a higher chance that CUs and DUs will add traffic to the total transferred load (see Eqs .(5.8) and (5.17)), which might prevent us from switching off a BS. Finally, the plot corresponding to each constraint is not always linear, as some additional phenomena, such as convergence to stationarity for the stochastic systems we use in constraints 2 and 3, also affect systems' behavior.

Thus, smaller switch-off duration X promises larger energy savings, but also implies that the system will (a) have to re-evaluate the state of the system and repeat its decision quite frequently (computation complexity) and (b) it might lead to some additional energy wastage (and performance degradation) due to the fixed power (and delay) needed to switch off and back on a BS. This suggests an interesting trade-off that we plan to explore in future.

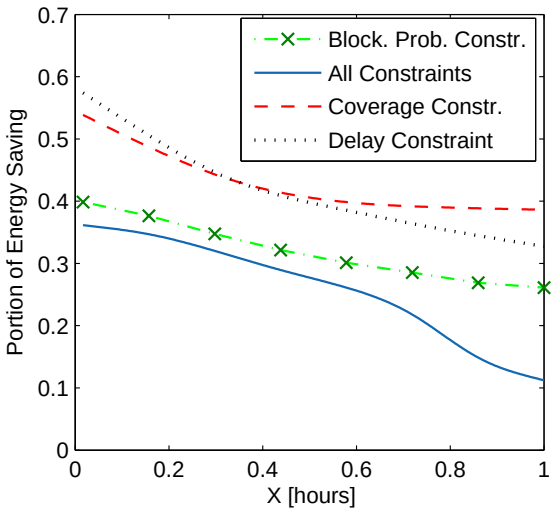


Figure 5.6: Portion of Energy Saving versus switching-off period in the considered HetNet scenario when the various constraint thresholds are $p_f = 0.4$, $p_{block} = 10^{-3}$, $D_{max} = 0.2$sec.

5.4 Conclusion

In this chapter, we considered the problem of energy saving in future HetNets by switching off underloaded BSs while focusing on short sleeping durations.

Specifically, we have shown how the potential degradation of user QoS could be analytically captured and bounded along different dimensions, namely coverage probability, blocking probability (for dedicated flows), and delay (for best effort flows). We used transient analysis tools to encompass the feature of the short-time system dynamics. Based on the proposed framework, we then showed how a significant amount of energy could be saved while maintaining some desired QoS levels. We also investigated through extensive simulations the tradeoff between power savings and duration of sleeping mode. While there is inefficient research to draw the importance of this tradeoff, in this work we highlight its importance since it stresses an arisen opportunity in future and dense HetNets.

Chapter 6

Conclusions and Future Research.

Nowadays, operators believe that the aggressive densification of network deployments, overlaying the conventional macro-cells, with a high number of small cells is the only way of dealing with the arising traffic crunch. Such a densification increases the spatio-temporal fluctuations of the traffic load within the network, by strongly affecting the dynamics of the system and thus system performance. This suggests that a plethora of radio access functionalities, that were rather simplistic in the conventional networks, shall be revisited for the next-generation networks (e.g., in 5G systems).

Such revisions shall not only (i) address the arising trends, challenges and bottlenecks that such dense networks born, but also (ii) better understand the corresponding performance improvements of such deployments and the conditions under which they hold. As a first step in all of the chapters of this dissertation, we considered various models, we investigated network optimization problems, as well as we performed the required analysis and provided the corresponding solutions using tools mostly coming from probability, queueing and optimization theory. The derived expressions can be used from the operators to design efficient network protocols, scheduling algorithms, and various distributed implementations.

Our contributions are summarized as following:

- We begin with Chapter 2 by revisiting the famous *user association* problem while being focused on the radio access network. Having the popular α-fair objective function as our starting point [10] we significantly extended it to realistically capture some features of the next-generation, data-centric cellular systems. These features include: (i) traffic differentiation between best-effort (traffic that demand for elastic resources) and dedicated (traffic that demand for non-elastic resources) flows, (ii) joint uplink and downlink traffic performance, as well as (iii) both Split and Joint UL/DL association schemes. We thus sketched a complete optimization framework capturing all the above characteristics, and eventually we derived various "device centric" user association rules. Interestingly, when considering multiple objectives these rules resemble a weighted version of the harmonic or arithmetic mean of the individual rules. This suggests that one can flexibly add different flow types or dimensions in our framework, and flexibly derive the optimal user association rules, without any analytical calculations.

- Later, in Chapter 3, motivated by the emerging bottlenecks arising in the backhaul network, we extend our objective function to capture the backhaul network limitations. Our aim is to better understand the impact of backhaul (i) link capacities, and (ii) network

topology, in under-provisioned backhaul scenarios. In particular, we extend our optimization problem by adopting appropriate penalty functions to capture the various additional constraints and avoid congested backhaul links. Eventually we analytically derive novel backhaul-aware and "device centric" user association rules, by handling the penalty constraints in either a "soft" or "hard" manner. Simulation results corroborate the correctness of our framework, and provide not only qualitative insights on the emerging backhaul bottleneck, but also offer quantitative results to better illustrate its impact on the wide user and network performance.

- While in the previous chapters we assumed static and fixed bandwidth allocation (e.g., between UL and DL for a BS, or a backhaul link), in this chapter we explore the opportunity for a more flexible and dynamic allocation, in the context of TDD. Thus, we include the related resource allocation parameters in the objective function and include the (i) flexible TDD allocation between UL/DL for BSs, (ii) TDD allocation for backhaul links, along with the (iii) user association problem. We derive a novel algorithm that decomposes this problem into three different levels, possibly running into different elements (e.g., at the UE, BS, backhaul link), and using optimization theory we prove that this algorithm converges to the global optimum. Simulation results show performance improvements up to 3 times, when the operator is allowed to flexibly allocate its resources between DL and UL depending on the traffic statistics.

- Eventually, in Chapter 5, we present a novel framework that studies the load fluctuations in dense HetNets in order to improve energy efficiency. Specifically, one of our key parameters is the duration of the switching-off period, and we claim that short sleeping periods in small cells can promise high energy savings. To that end, we propose a framework that switches off BSs subject to three sophisticated QoS constraints: coverage failure probability, blocking probability, service delay. These constraints are derived by considering the short time-scale system dynamics, and are related to different ways that system performance could deteriorate. Simulation results show that significant energy savings can be achieved, as a function of (i) the three different QoS constraints, (ii) the switching-off period. The latter reveals a novel and promising way for energy savings in future networks.

6.1 Future Work

In the context of future heterogeneous networks, various optimizations have been proposed and studied in the radio access network part. Nevertheless, there is a rather insufficient research into the backhaul network. Thus, the maximum performance improvements as well as the arising dependencies between the access-backhaul networks for next generation systems (e.g., 5G) are not clear yet. Having proposed a tractable and analytical framework for joint access and backhaul network optimization, the future research directions we are planning to focus on in this context follow.

- *Joint radio and L3 backhaul routing.* Mesh backhaul topologies with multiple available routing paths are expected to be the rule, rather than the exception in future networks. Our assumption of fixed, L2 backhaul routing is restrictive, and as we saw in Chapter 4, it also penalizes performance. It would be interesting to jointly optimize (a) the BS that each

user should be associated with, as well as (b) the routing path up to an aggregation point (L3 routing). Our goal is twofold: to consider (a) *per-BS offloading*, where each BS should offload all flows by using the same routing path upto an aggregation point, (b) *per-location offloading*, where flows at different locations of a certain BS can follow different routing paths to improve system performance. It remains to be investigated whether these two options retain the convexity and other desirable properties of the original problem.

- *Dual Connectivity* operation of LTE-A networks allows control plane signaling to be maintained on the macro layer while aggregating SCs to provide them extra (user plane) backhaul capacity. Within our framework, this is applicable since regardless of the backhaul topology and path, all traffic is routed through an (high coverage area) eNB. So, one can consider the amount of traffic related to the control plane, and offload it directly to the macro eNB, by eventually lighten both the SC and the corresponding backhaul link capacities and allow them to "carry" more user-plane traffic. This is a promising way to improve performance.

Other future work steps, include the consideration of C-RAN, or distributed SDN-based control to achieve higher performance improvements.

- *Fronthaul Network and C-RAN.* In the proposed framework we only considered the backhaul network and the constraints related to it. However, modern networks tend to increasingly focus on Centralized-Radio Access Network (C-RAN) architectures, fact that has lead fronthaul networks to be rather under-provisioned and their architecture to be revisited. Thus, the introduction of the fronthaul in our framework, along with the potentially influenced by, backhaul network, and their interaction is another promising extension. Furthermore, the investigation of C-RAN functionalities that can be *flexibly* centralized, depending not only on simple current fronthaul-network status, but also on other factors, as the backhaul constraint, or the *expected* system performance, the overall *expected* network performance, as initially investigated in this paper, along with radio access, backhaul and fronthaul resources utilization.

Chapter 7

Résumé.

Les réseaux cellulaires sans fil sont généralement constitués d'un ensemble d'équipements utilisateurs (User Equipments – UE) et d'une collection de stations de base (Base Stations – BS) qui se connectent au coeur du réseau (Core Network – CN) par l'intermédiaire d'un ensemble de liens backhaul (BH). Dans les réseaux traditionnels, l'intensité du trafic et la demande à travers différents UE restent généralement similaires. En outre, les BS ont des niveaux de puissance d'émission semblables, des modèles d'antenne ainsi que la connectivité backhaul au core.

De nos jours, la demande de trafic croît de façon exponentielle pour les services UE provenant de différentes applications [1]. En particulier, les applications de «réalité augmentée», de réseautage social, ainsi que les diverses applications de Machine Type Communication (MTC) (par exemple, le suivi des systèmes de soins de santé ou des systèmes énergétiques) présentent des exigences élevées de capacité et de latence. Les opérateurs qui luttent pour faire face à cette augmentation du trafic tendent à construire des déploiements de réseaux plus denses pour améliorer la réutilisation spatiale. Plus précisément, ils construisent des petites cellules (small cells – SC) additionnelles avec les macro BS (MBS) déjà existantes. Les niveaux de puissance du MBS sont généralement entre 5 et 40 W, ceux pour les SC sont seulement 0.25 et 2 W. Les réseaux composés d'un mélange de différentes BS avec des niveaux de puissance différents (et donc des tailles de cellules différentes) sont appelés reseaux hétérogènes – (Heterogeneous Networks – HetNets).

Alors qu'une topologie HetNet dense représente une opportunité prometteuse pour répondre à la demande croissante de trafic, elle nécessite une planification et une maintenance des backhaul méticuleuses pour desservir le grand nombre de SC. À cette fin, les recherches actuelles semblent étudier les nouvelles exigences de backhaul en termes de dépenses en immobilisations (Capital Expenditure – CAPEX) et de dépenses d'investissement (Operational Expenditure – OPEX), de couverture, de capacité, de sécurité, de latence, de synchronisation, de conception physique et de gestion en comparaison aux exigences traditionnelles posées par les macrocellules.

7.1 Motivation et contributions de la thèse

Cependant, beaucoup de choses restent plutôt floues pour de tels HetNets. Dans cette thèse, nous nous concentrons sur divers problèmes importants de ces réseaux:

- l' association d'utilisateurs,
- l' accès et le backhaul de l'allocation TDD,

- la gestion de l'énergie.

Notre principale motivation est de mieux refléter certaines lacunes, les bouleversements et les hypothèses traditionnelles que nous rencontrons habituellement dans le travail existant. Plus précisément, les chapitres de cette thèse sont organisés selon:

Chapitre 2 - Association des usagers qui considèrent le trafic (Ici nous essayons de capturer la différenciation du trafic)

Chapitre 3 - Association d'utilisateurs qui considèrent le backhaul (Ici nous essayons de capturer les limitations de backhaul comme un problème fondamental pour les réseaux 5G)

Chapitre 4 - Association des utilisateurs et allocation TDD (Ici nous essayons de capturer la TDD dynamique)

Chapitre 5 - Minimisation d'énergie (Ici nous essayons de capturer la minimisation d'énergie pour les nouvelles générations HetNets).

Dans la suite, nous donnons un bref résumé de chaque chapitre et soulignons nos principaux résultats.

Chapter 2 - Association des usagers qui considèrent le trafic.

Dans ce chapitre, nous voulons dériver les règles d'association utilisateur pour les réseaux de la génération prochaine et se concentrer sur la différenciation du trafic.

(A. Modèle). Nous pouvons résumer ici nos hypothèses/ modèle:

Les probabilités des différents types de flux sont

- ζ_b est la probabilité d'un meilleur flux d'effort (par exemple, une page facebook ou youtube),
- $\zeta_d = 1 - \zeta_b$ est la probabilité d'un flux dédié (ou dévoué) (par exemple un appel VoIP).

Aussi, les probabilités des flux descendantes (downlink – DL) ou montantes (uplink – UL) sont

- ζ^D est la probabilité pour un flux descendante,
- $\zeta^U = 1 - \zeta_D$ est la probabilité pour un flux montante,

Ainsi, il existe 4 processus indépendants d'arrivée de flux de Poisson avec des taux (figure 7.1):

- $\lambda^{D,b}(x) = z^D \cdot z^b \cdot \lambda(x)$,
- $\lambda^{D,d}(x) = z^D \cdot z^d \cdot \lambda(x)$,
- $\lambda^{U,b}(x) = z^U \cdot z^b \cdot \lambda(x)$,
- $\lambda^{U,d}(x) = z^U \cdot z^d \cdot \lambda(x)$.

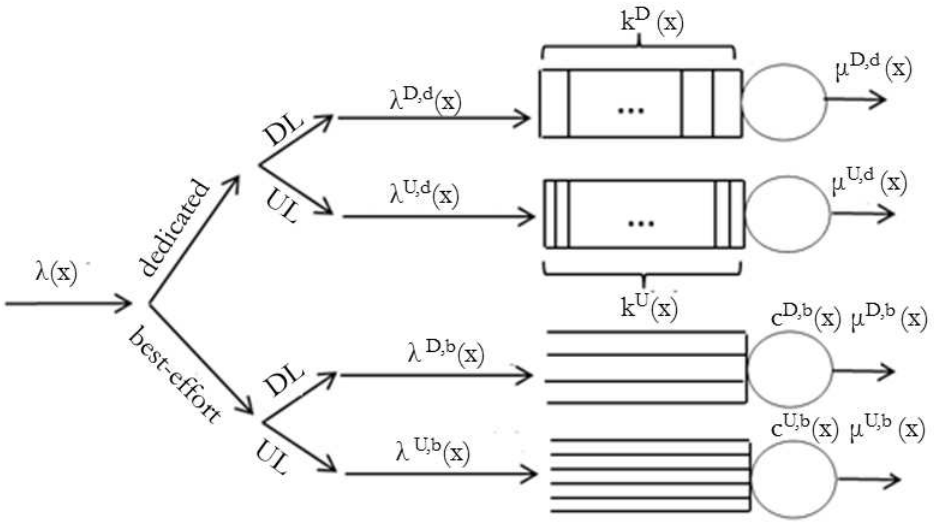


Figure 7.1: Les 4 processus.

Nos symboles peuvent être résumés ci-dessous.

Table 7.1: Notation

Variable	Meilleur d'effort (Flux) descendante	Meilleur d'effort (Flux) montante	Dévoué (Flux) descendante	Dévoué (Flux) montante
Exposant	D,b	U,b	D,d	U,d
Probabilité	$z^D \cdot z^b$	$z^U \cdot z^b$	$z^D \cdot z^d$	$z^U \cdot z^d$
Bandwidth BS i	$w_i \cdot \zeta_i \cdot \xi_i^D$	$w_i(1-\zeta_i) \cdot \xi_i^U$	$w_i \cdot \zeta_i(1-\xi_i^D)$	$w_i(1-\zeta_i)(1-\xi_i^U)$
Rate (taux) x	$\lambda^{D,b}(x)$	$\lambda^{U,b}(x)$	$\lambda^{D,d}(x)$	$\lambda^{U,d}(x)$
Max. rate | les serveurs – BS i	$c_i^{D,b}(x)$	$c_i^{U,b}(x)$	$k_i^D(x)$	$k_i^U(x)$
Charge – BS i	$\rho^{D,b}(x)$	$\rho^{U,b}(x)$	$\rho^{D,d}(x)$	$\rho^{U,d}(x)$
LB parameter $\in [0,\infty)$	$\alpha^{D,b}$	$\alpha^{U,b}$	$\alpha^{D,d}$	$\alpha^{U,d}$
Total charge – i BS	$\rho^{D,b}$	$\rho^{U,b}$	$\rho^{D,d}$	$\rho^{U,d}$
Probabilité d'association	$p_i^{D,b}(x)$	$p_i^{U,b}(x)$	$p_i^{D,d}(x)$	$p_i^{U,d}(x)$
Taille de flux (bits) | durée (sec)	$1/\mu^{D,b}$	$1/\mu^{U,b}$	$1/\mu^{D,d}$	$1/\mu^{U,d}$
Demande de flux (bps)	-	-	B^D	B^U
BH Capacité j	$C_h^D(j)$	$C_h^U(j)$	-	-
BH j – congestion	$\mathcal{I}^D(j)$	$\mathcal{I}^U(j)$	-	-

(B. Optimisation) Nous présentons maintenant quelques définitions de notre problème d'optimisation.

Definition 13. *(Faisabilité) $l \in \{U,D\}, t \in \{b,d\}$, et ϵ est une petite constante positive. La set $f^{l,t}$ des charges BS faisable $\rho^{\mathbf{l,t}} = (\rho_1^{l,t}, \rho_2^{l,t}, \ldots, \rho_{\|\mathcal{B}\|}^{l,t})$*

$$\begin{aligned} f^{l,t} = \Big\{ \rho^{\mathbf{l,t}} \mid \rho_i^{l,t} = \int_{\mathcal{L}} p_i^{l,t}(x)\rho_i^{l,t}(x)dx, \\ 0 \le \rho_i^{l,t} \le 1-\epsilon, \\ \sum_{i\in\mathcal{B}} p_i^{l,t}(x) = 1, \\ 0 \le p_i^{l,t}(x) \le 1, \forall i \in \mathcal{B}, \forall x \in \mathcal{L} \Big\}. \end{aligned} \tag{7.1}$$

Lemma 9. *Les sets $f^{D,b}, f^{D,d}, f^{U,b}, f^{U,d}$,$[f^{D,b}; f^{D,d}]$, $[f^{U,b}; f^{U,d}]$, $[f^{D,b}; f^{U,b}]$, $\mathcal{F} = [f^{D,b}; f^{D,d}; f^{U,b}; f^{U,d}]$, sont convexes.*

Definition 14. *(DL Objectif): Notre (descendant) objectif est*

$$\phi_{\alpha^{\mathbf{D}}}(\rho^{\mathbf{D}}) = \sum_{i\in\mathcal{B}} \theta \cdot \frac{(1-\rho_i^{D,b})^{1-\alpha^{D,b}}}{\alpha^{D,b}-1} + (1-\theta)\cdot \frac{(1-\rho_i^{D,d})^{1-\alpha^{D,d}}}{\alpha^{D,d}-1}, \text{ if } \alpha^{D,d}, \alpha^U \neq 1. \tag{7.2}$$

Si $\alpha^{D,b} = 1$, la fraction doit être remplacée par $\log(1-\rho_i^{D,b})^{-1}$.

Definition 15. *(DL/UL Objectif) Notre (descendant et montant) objectif est*

$$\phi_\alpha(\rho) = \tau \cdot \phi_{\alpha^{\mathbf{D}}}(\rho^{\mathbf{D}}) + (1-\tau)\cdot \phi_{\alpha^{\mathbf{U}}}(\rho^{\mathbf{U}}). \tag{7.3}$$

Lemma 10. *$\phi_\alpha(\rho)$ est convexe.*

Definition 16. *(Problème d'optimisation 1) Notre problème d'optimisation (en fait c'est problème de minimisation) est*

$$\underset{\rho}{\text{minimize}}\{\phi_\alpha(\rho)|\rho \in \mathcal{F}\}. \tag{7.4}$$

Lemma 11. *Problème 1 est convexe.*

Notez que les différentes valeurs d' α optimisent les différentes métriques.

- $\alpha = 0$: Optimisation de: *efficacité spectrale*,
- $\alpha = 1$: Optimisation de: *débit*,
- $\alpha = 2$: Optimisation de: *latence*,
- $\alpha \to \infty$: Optimisation de: *Efficacité de l'équilibrage de la charge.*

Notez également que θ pèse le meilleur-effort ($\theta \to 0$) vs performance dédiée ($\theta \to 1$).

Maintenant, nous dérivons les règles optimales pour divers scénarios. Nous commençons par le scénario plus simple: Split UL/DL. Dans ce scénario, un utilisateur *peut s'associer à deux BS*: une pour DL et une pour UL.

Theorem 1.1. *(Split UL/DL – Règles optimales) Si $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ désigne le vecteur optimal, les règles optimales sont:*

$$i(x) = \arg\max_{i \in \mathcal{B}} \left(\underbrace{c_i(x)}_{\text{Connaissances des utilisateurs}} \cdot \overbrace{P_i}^{\text{BS message de diffusion}} \right) \tag{7.5}$$

et chaque BS diffuse:

$$P_i = \frac{\left(1-\rho_i^{*b}\right)^{\alpha^b} \cdot \left(1-\rho_i^{*d}\right)^{\alpha^d}}{e^b \cdot \left(1-\rho_i^{*d}\right)^{\alpha^d} + e^d \cdot \left(1-\rho_i^{*b}\right)^{\alpha^b}}.$$

Notez que, $e^b = \frac{\theta z^D z^b}{\mu^b \zeta_i \xi_i^D}$ aussi bien que $e^d = \frac{(1-\theta) z^D z^d B^D}{\mu^d \zeta_i (1-\xi_i^D)}$.

Proof. Nous prouvons que la règle ci-dessus en effet minimise l'objectif. Comme nous l'avons vu, le problème est convexe. Par conséquent, il convient de vérifier la condition

$$\langle \nabla\phi(\rho^*), \Delta\rho^* \rangle \geq 0 \tag{7.6}$$

pour tous $\rho \in f$, $\Delta\rho^* = \rho - \rho^*$. $p(x)$ et $p^*(x)$ sont les vecteurs de probabilité de routage associés pour ρ et ρ^*, respectivement. En utilisant équation 7.5, la règle optimale:

$$p_i^*(x) = \mathbf{1}\left\{ i = \arg\max_{i \in \mathcal{B}} c_i^D(x) \frac{\left(1-\rho_i^{*b}\right)^{\alpha^b} \cdot \left(1-\rho_i^{*d}\right)^{\alpha^d}}{e^b \cdot \left(1-\rho_i^{*d}\right)^{\alpha^d} + e^d \cdot \left(1-\rho_i^{*b}\right)^{\alpha^b}} \right\}. \tag{7.7}$$

Le produit scalaire est:

$$
\begin{aligned}
\langle \nabla\phi\left(\rho^*\right), \Delta\rho^* \rangle &= \sum_{z=\{b,d\}} \frac{\partial\phi}{\partial\rho_z}\left(\rho^*\right)\left(\rho_z - \rho_z^*\right) \\
&= \frac{\partial\phi}{\partial\rho^b}(\rho^*)(\rho^b - \rho^{*b}) + \frac{\partial\phi}{\partial\rho^d}(\rho^*)(\rho^d - \rho^{*d}) \\
&= \theta \sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^b)^{\alpha^b}}(\rho_i^b - \rho_i^{*b}) + (1-\theta)\sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^d)^{\alpha^d}}(\rho_i^d - \rho_i^{d*}) \\
&= \sum_{i\in\mathcal{B}} \frac{\theta\int_L \rho_i^b(x)(p_i(x) - p_i^*(x))dx}{(1-\rho_i^b)^{\alpha^b}} + \frac{(1-\theta)\int_L \rho_i^d(x)(p_i(x) - p_i^*(x))dx}{(1-\rho_i^d)^{\alpha^d}} \\
&= \int_L \lambda(x) \sum_{i\in\mathcal{B}} (p_i(x) - p_i^*(x)) \left(\frac{e^b(1-\rho_i^{*d})^{\alpha^d} + e^d(1-\rho_i^{*b})^{\alpha^b}}{c_i(x)\cdot(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}} \right) dx.
\end{aligned} \tag{7.8}
$$

Notez que,

$$
\sum_{i\in\mathcal{B}} p_i(x) \frac{e^b(1-\rho_i^{*d})^{\alpha^d} + e^d(1-\rho_i^{*b})^{\alpha^b}}{c_i(x)\cdot(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}} \geq \sum_{i\in\mathcal{B}} p_i^*(x) \frac{e^b(1-\rho_i^{*d})^{\alpha^d} + e^d(1-\rho_i^{*b})^{\alpha^b}}{c_i(x)\cdot(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}} \tag{7.9}
$$

. □

Nous explorons maintenant le scénario Joint UL / DL. Dans ce scénario, un utilisateur doit s'associer *à une BS* pour le déchargement de trafic DL et UL.

Theorem 1.2. *(Joint UL/DL – Règles optimales) Si* $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ *désigne le vecteur optimal, les règles optimales sont:*

$$
i(x) = \arg\max_{i\in\mathcal{B}} \frac{1}{\frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)}} \tag{7.10}
$$

et maintenant,

$$
P_i^D = \frac{\sum\limits_{t\in\{b,d\}} e^{D,t} \prod\limits_{c\in\Omega\neq(D,t)} ((1-\rho^{*c})^{\alpha^c})}{\prod\limits_{c\in\Omega} ((1-\rho^{*c})^{\alpha^c})}
$$

et

$$
\Omega \in \{(D,d),(D,b),(U,d),(U,b)\}
$$

$$
e^{D,b} = \tau \frac{\theta^D z^D z^b}{\mu^{D,b}\zeta\xi^D}
$$

$$
e^{D,d} = \tau \frac{(1-\theta^D) z^D z^d B^D}{\mu^{D,d}\zeta(1-\xi^D)}
$$

$$
e^{U,b} = (1-\tau) \frac{\theta^U z^U z^b B^U}{\mu^{U,b}(1-\zeta)\xi^U}
$$

et

$$
e^{U,d} = (1-\tau) \frac{(1-\theta^U) z^U z^d B^U}{\mu^{U,d}(1-\zeta)(1-\xi^U)}
$$

.

Proof. Les étapes de la preuve sont similaires à celles du scénario de Split UL / DL, où l'on doit aussi exiger $p_i^D(x) = p_i^U(x)$. Ensuite, le produit scalaire:

$$\begin{aligned}
&\langle \nabla\phi\left(\rho^*\right), \Delta\rho^* \rangle = \sum_{z=\{b,d\}} \frac{\partial\phi}{\partial\rho_z}\left(\rho^*\right)\left(\rho_z - \rho_z^*\right) \\
&= \theta^D \sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^{D,b})^{\alpha^{D,b}}}(\rho_i^{D,b} - \rho_i^{*D,b}) + (1-\theta^D)\sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^{D,d})^{\alpha^{D,d}}}(\rho_i^{D,d} - \rho_i^{D,d*}) \\
&+ \theta^U \sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^{U,b})^{\alpha^{U,b}}}(\rho_i^{U,b} - \rho_i^{*U,b}) + (1-\theta^U)\sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^{U,d})^{\alpha^{U,d}}}(\rho_i^{U,b} - \rho_i^{U,d*}) \\
&= \int_L \lambda(x) \sum_{i\in\mathcal{B}} (p_i(x) - p_i^*(x)) \left(\frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)}\right) dx \geq 0,
\end{aligned} \tag{7.11}$$

□

(C. Évaluation) Maintenant, nous faisons quelques simulations. Nos paramètres sont sur Table 7.2.

Table 7.2: Paramètres de simulation

Variable	**Valeur**
$P_{eNB}/P_{SC}/P_{UE}$	43/24/12 dBm
w/W	10/10 MHz
N_0	-174 dBm/Hz
ζ_i^D, ζ_i^U	0.5/0.5
$\frac{1}{\mu^{D,b}}/\frac{1}{\mu^{U,b}}$	100/20 Kbytes
$B^D(x)/B^U(x)$	512, 128 kbps
z^b, z^D	0.3,0.6

Nous considérons une zone $2 \times 2\ km^2$. La figure 7.2 montre une carte de couleur de la demande de trafic hétérogène $\lambda(x)$ ($flows/hour$ par unité de surface) (bleu impliquant trafic faible et rouge élevé) avec 2 hotspots. Nous supposons que cette zone est couverte par 2 BS macro et 8 SC.

Couverture Nous décrivons la zone de couverture dans différents scénarios (α scénarios). Par exemple:

- $\alpha^{D,b} = \alpha^{D,d} = 0$: Figure 7.3(a),
- $\alpha^{D,b} = \alpha^{D,d} = 10$: Figure 7.3(b).

Dans la Table 7.3, vous pouvez voir que différents scénarios optimisent (en effet) des métriques différentes (i.e. le nombre de serveurs (dédiés) $E[k^D]$ et équilibrage de charge ou efficacité

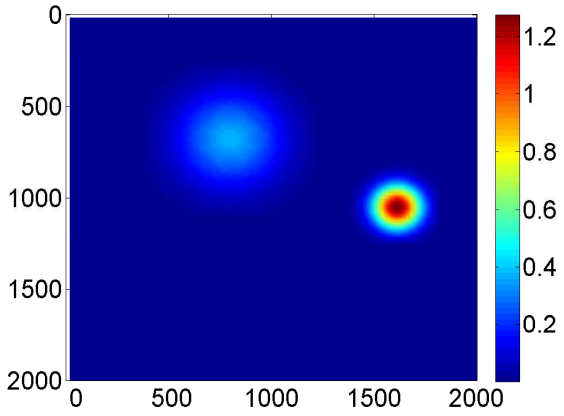
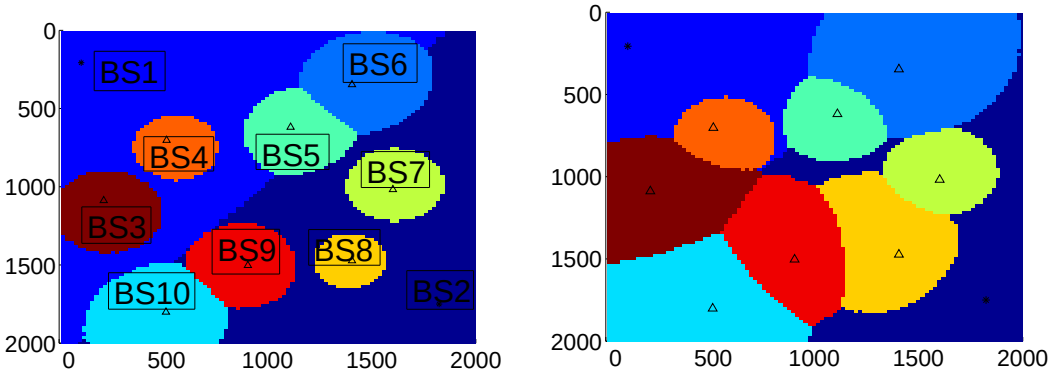


Figure 7.2: Taux d'arrivée.



(a) $\alpha^{D,b} = \alpha^{D,d} = 0$. (b) $\alpha^{D,b} = \alpha^{D,d} = 10$.

Figure 7.3: Associations optimales.

Table 7.3: Résultats de performance (Figure 2.3).

	Rates et serveurs		**équilibrage de charge**	
	$E[c^{D,b}]$ (Mbps)	$E[k^D]$	1-$MSE^{D,b}$	1-$MSE^{D,d}$
Fig. 7.3(a)	16.3	32	0.77	0.78
Fig. 7.3(b)	14.3	27	0.96	0.995

d'utilisation 1-$MSE^{D,b}$). En outre, dans la figure 7.4, vous pouvez voir comment θ affecte les résultats.

Le travail de ce chapitre corresponde a la publication suivante

- *N. Sapountzis, T. Spyropoulos, N. Nikaein, U. Salim, An analytical framework for optimal*

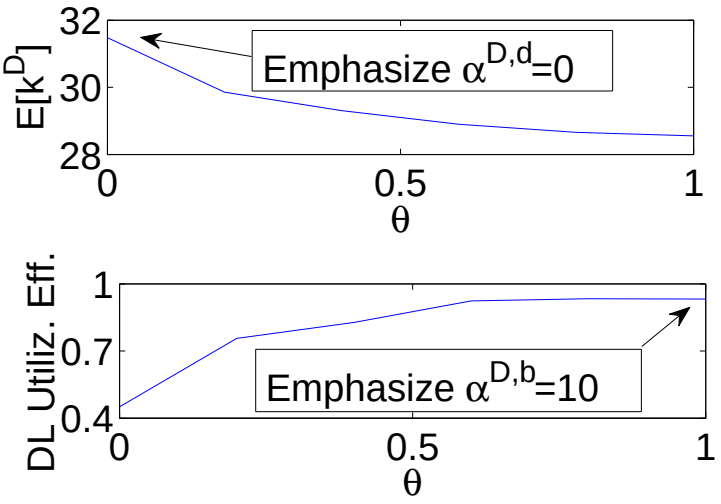


Figure 7.4: Impact de θ.

downlink-uplink user association in HetNets with traffic differentiation, in Proc. IEEE Global Communications (GLOBECOM) Conference, San Diego, CA, USA, 2015.

Chapter 3 - Association des usagers qui considèrent le backhaul.

Lorsque les capacités de backhaul ne sont pas suffisantes, les règles de la section précédente sont (effectivement) incorrectes.

Ici, nous prenons en compte le backhaul. Et nous retrouvons les règles optimales. Pour plus de simplicité, dans ce chapitre, nous ne considérons que les flux de meilleurs efforts.

(A. Modèle) Nous étendons notre modèle pour inclure le réseau de backhaul (Figure 7.5).

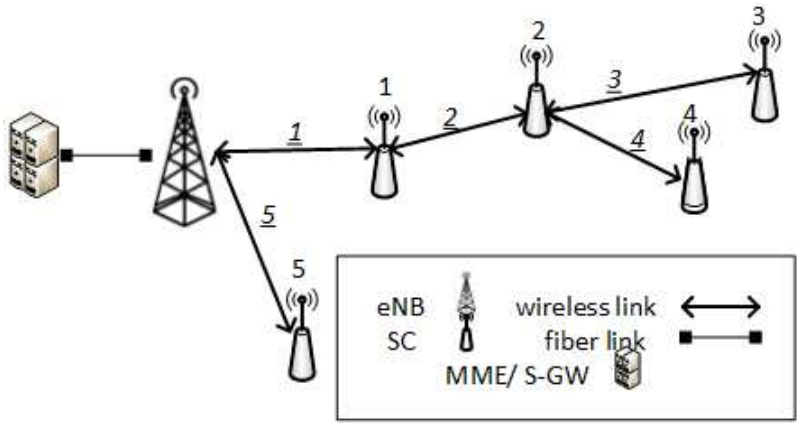


Figure 7.5: Topologie du backhaul.

La charge d'un lien backhaul j est

$$\sum_{i\in\mathcal{B}(j)} \frac{\frac{\rho_i^D}{\zeta_i}\cdot(\zeta_i\cdot\tilde{c}_i^D)}{Z(j)\cdot C_h(j)} = \sum_{i\in\mathcal{B}(j)} \frac{\rho_i^D\cdot\tilde{c}_i^D}{Z(j)\cdot C_h(j)}. \tag{7.12}$$

où $\tilde{c}_i^D$ est le taux maximal de DL de BS i.

Contrainte: nous devons exiger pour chaque lien de backhaul

$$\sum_{i\in\mathcal{B}(j)} \frac{\rho_i^D\tilde{c}_i^D}{Z(j)\cdot C_h^D(j)} < 1,\ \forall j\in\mathcal{B}_h. \tag{7.13}$$

(B. Optimisation) Maintenant, notre problème comprend les contraintes backhaul.

Definition 17. ***(Objectif avec backhaul contraintes)*** *Notre objectif avec backhaul contraintes est*

$$\begin{aligned} &\underset{\rho}{\text{minimize}} \left\{\phi_\alpha(\rho)|\rho\in\mathcal{F}\right\}, \\ &\textit{subject to} \sum_{i\in\mathcal{B}(j)} \frac{\rho_i^D\tilde{c}_i^D}{Z(j)\cdot C_h^D(j)} \prec 1,\ \forall j\in\mathcal{B}_h, \end{aligned} \tag{7.14}$$

Nous considérons les contraintes comme des *fonctions de pénalité*. Nous illustrons immédiatement les règles optimales.

Il existe deux topologies backaul: star (one-hop) et tree (multi-hop). Nous commençons par la plus simple (star).

Theorem 1.3. *(Star topologie – Règles optimales (backhaul-conscient) – Split UL/DL) Si* $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ *désigne le vecteur optimal, les règles optimales sont:*

$$i(x) = \arg\max_{i \in \mathcal{B}} \left(\underbrace{c_i(x)}_{\text{Connaissances des utilisateurs}} \cdot \overbrace{P_i}^{\text{BS message de diffusion}} \right), \tag{7.15}$$

et

$$P_i = \frac{(1-\rho_i^*)^\alpha}{1 + 2\gamma \cdot (1-\rho_i^*)^\alpha \cdot \tilde{c}_i \cdot \frac{\mathcal{I}(i)}{Z(i)C_h(i)} \cdot \left(\frac{\rho_i^* \tilde{c}_i}{Z(i)C_h(i)} - 1 \right)}.$$

Proof. Les étapes de la preuve sont similaires à celles du scénario de Chapitre 2. Ensuite, le produit scalaire:

$$\begin{aligned}
\langle \nabla \phi(\rho^*), \Delta \rho^* \rangle &= \sum_{i \in \mathcal{B}} \left(\frac{1}{(1-\rho_i^*)^\alpha} + \gamma \mathcal{I}(i) \frac{2\rho_i^* \tilde{c}_i^2 - 2\tilde{c}_i Z(i) C_h(i)}{Z(i) C_h(i)^2} \right) (\rho_i - \rho_i^*) \\
&= \sum_{i \in \mathcal{B}} \frac{1 + 2\gamma \mathcal{I}(i)(1-\rho_i^*)^\alpha \frac{(\rho_i^* \tilde{c}_i^2 - \tilde{c}_i Z(i) C_h(i))}{Z(i)C_h(i)^2}}{(1-\rho_i^*)^\alpha} \int_{\mathcal{L}} \rho_i(x) \left(p_i(x) - p_i^*(x) \right) dx \\
&= \int_{\mathcal{L}} \frac{\lambda(x)}{\mu(x)} \sum_{i \in \mathcal{B}} \left(\frac{1 + 2\gamma (1-\rho_i^*)^\alpha \tilde{c}_i \frac{\mathcal{I}(i)}{Z(i)C_h(i)} \left(\frac{\rho_i^* \tilde{c}_i}{Z(i)C_h(i)} - 1 \right)}{c_i(x)(1-\rho_i^*)^\alpha} \right) (p_i(x) - p_i^*(x))\, dx.
\end{aligned}$$

Notez que,

$$\begin{aligned}
&\sum_{i \in \mathcal{B}} p_i(x) \left(\frac{1 + 2\gamma (1-\rho_i^*)^\alpha \tilde{c}_i \frac{\mathcal{I}(i)}{Z(i)C_h(i)} \left(\frac{\rho_i^* \tilde{c}_i}{Z(i)C_h(i)} - 1 \right)}{c_i(x)(1-\rho_i^*)^\alpha} \right) \geq \\
&\sum_{i \in \mathcal{B}} p_i^*(x) \left(\frac{1 + 2\gamma (1-\rho_i^*)^\alpha \tilde{c}_i \frac{\mathcal{I}(i)}{Z(i)C_h(i)} \left(\frac{\rho_i^* \tilde{c}_i}{Z(i)C_h(i)} - 1 \right)}{c_i(x)(1-\rho_i^*)^\alpha} \right)
\end{aligned}$$

. □

Theorem 1.4. *(Tree topologie – Règles optimales (backhaul-conscient)– Split UL/DL) Si* $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ *désigne le vecteur optimal, les règles optimales sont:*

$$i(x) = \arg\max_{i \in \mathcal{B}} \left(\underbrace{c_i(x)}_{\text{Connaissances des utilisateurs}} \cdot \overbrace{P_i}^{\text{BS message de diffusion}} \right), \tag{7.16}$$

et

$$P_i = \frac{(1-\rho_i^*)^\alpha}{1 + 2\gamma \cdot (1-\rho_i^*)^\alpha \cdot \tilde{c}_i \sum\limits_{j \in \mathcal{B}_h(i)} \frac{\mathcal{I}(j)}{Z(j)C_h(j)} \cdot \left(\frac{\sum\limits_{k \in \mathcal{B}(j)} \rho_k^* \tilde{c}_k}{Z(j)C_h(j)} - 1 \right)}.$$

Proof. Les étapes de la preuve sont similaires à celles du scénario de Chapitre 2. Ensuite, le produit scalaire:

$$\begin{aligned}
&\langle \nabla \Phi_\alpha (\rho^*), \Delta \rho^* \rangle = \\
&= \sum_{i \in \mathcal{B}} \left(\frac{1}{(1-\rho_i^*)^\alpha} + 2\gamma \sum_{j \in \mathcal{B}_h(i)} \mathcal{I}(j) \left(\frac{\sum_{k \in \mathcal{B}(j)} \rho_k^* \tilde{c}_k}{Z(j) C_h(j)^2} \tilde{c}_i - \frac{\tilde{c}_i}{Z(j) C_h(j)} \right) \right) (\rho_i - \rho_i^*) \\
&\quad \cdot \int_{\mathcal{L}} \rho_i(x) \left(p_i(x) - p_i^*(x) \right) dx = \\
&= \int_L \frac{\lambda(x)}{\mu(x)} \sum_{i \in \mathcal{B}} \left(\frac{1 + 2\gamma (1-\rho_i^*)^\alpha \tilde{c}_i \sum\limits_{j \in \mathcal{B}_h(i)} \frac{\mathcal{I}(j)}{Z(j) C_h(j)} \cdot \left(\frac{\sum\limits_{k \in \mathcal{B}(j)} \rho_k^* \tilde{c}_k}{Z(j) C_h(j)} - 1 \right)}{c_i(x) (1-\rho_i^*)^\alpha} \right) \cdot \\
&\quad \cdot \left(p_i(x) - p_i^*(x) \right) dx \geq 0,
\end{aligned} \tag{7.17}$$

□

Enfin, nous voyons le scénario Joint UL/DL, pour la topologie generale (tree).

Theorem 1.5. *(Tree topologie – Règles optimales (backhaul-conscient)– Joint UL/DL) Si* $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ *désigne le vecteur optimal, les règles optimales sont:*

$$i(x) = \arg\max_{i \in \mathcal{B}} \frac{1}{\frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)}}, \tag{7.18}$$

et $g^D = \tau, g^U = 1 - \tau$,

$$P_i^D = \frac{e^D \cdot (1 - \rho^{U*})^{\alpha^U}}{(1-\rho^{*D})^{\alpha^D} \cdot (1-\rho^{U*})^{\alpha^U}},$$

$$P_i^U = \frac{e^U \cdot (1 - \rho^{*D})^{\alpha^D}}{(1-\rho^{*D})^{\alpha^D} \cdot (1-\rho^{U*})^{\alpha^U}},$$

$$e^l = \frac{z^l \left(g^l + 2\gamma \left(1 - \rho_i^{*l}\right)^{\alpha^l} \sum\limits_{j \in \mathcal{B}_h(i)} \frac{\mathcal{I}^l(j)}{Z(j) C_h^l(j)} \left(\frac{\sum\limits_{k \in \mathcal{B}(j)} \rho_k^{*l} \tilde{c}_k^l}{Z(j) C_h^l(j)} - 1 \right) \right)}{\mu^l(x)}, l \in \{D, U\}.$$

Proof. Les étapes de la preuve sont similaires à celles du scénario de Chapitre 2. Ensuite, le produit scalaire:

$$
\begin{aligned}
&\langle \nabla \Phi_\alpha(\rho^*), \Delta\rho^* \rangle = \\
&= \sum_{i\in\mathcal{B}} \left(\tau \cdot \frac{1}{(1-\rho_i^{*D})^{\alpha^D}} + 2\gamma \sum_{l\in\mathcal{B}_h} \mathcal{I}^D(l) \left(\frac{\sum_{j\in\mathcal{B}(l)} \rho_j^D \tilde{c}_j^D}{Z(l)C_h(l)^2} \tilde{c}_i^D - \frac{\tilde{c}_i^D}{Z(l)C_h(l)} \right) \right) \rho_i^D - (\rho_i^{*D})) + \\
&+ \sum_{i\in\mathcal{B}} \left((1-\tau) \cdot \frac{1}{(1-\rho_i^{*U})^{\alpha^U}} + 2\gamma \sum_{l\in\mathcal{B}_h} \mathcal{I}^U(l) \left(\frac{\sum_{j\in\mathcal{B}(l)} \rho_j^U \tilde{c}_j^U}{Z(l)C_h(l)^2} \tilde{c}_i^U - \frac{\tilde{c}_i^U}{Z(l)C_h(l)} \right) \right) (\rho_i^U - (\rho_i^{U*})) = \\
&= \sum_{i\in\mathcal{B}} \left(\tau \cdot \frac{1}{(1-\rho_i^{*D})^{\alpha^D}} + 2\gamma \sum_{l\in\mathcal{B}_h} \mathcal{I}^D(l) \left(\frac{\tilde{c}_i^D (\sum_{j\in\mathcal{B}(l)} \rho_j^D \tilde{c}_j^D - Z(l)C_h(l))}{Z(l)C_h(l)^2} \right) \right) \cdot \\
&\quad \cdot \int_{\mathcal{L}} \rho_i^D(x) \left(p_i(x) - p_i^*(x) \right) dx + \\
&+ \sum_{i\in\mathcal{B}} \left((1-\tau) \cdot \frac{1}{(1-\rho_i^{*D})^{\alpha^U}} + 2\gamma \sum_{l\in\mathcal{B}_h} \mathcal{I}^U(l) \left(\frac{\tilde{c}_i^U (\sum_{j\in\mathcal{B}(l)} \rho_j^U \tilde{c}_j^U - Z(l)C_h(l))}{Z(l)C_h(l)^2} \right) \right) \cdot \\
&\quad \cdot \int_{\mathcal{L}} \rho_i^U(x) \left(p_i(x) - p_i^*(x) \right) dx = \\
&= \int_L \lambda(x) \sum_{i\in\mathcal{B}} \begin{pmatrix} \frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)} \\ 1 \end{pmatrix} \cdot \left(p_i(x) - p_i^*(x) \right) dx.
\end{aligned}
\tag{7.19}
$$

Notez que,

$$
\sum_{i\in\mathcal{B}} p_i(x) \left\{ \begin{matrix} \frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)} \\ 1 \end{matrix} \right\} \geq \sum_{i\in\mathcal{B}} p_i^*(x) \left\{ \begin{matrix} \frac{P_i^D}{c_i^D(x)} + \frac{P_i^U}{c_i^U(x)} \\ 1 \end{matrix} \right\}. \tag{7.20}
$$

□

(C. Évaluation) Maintenant, nous faisons quelques simulations.
Nous examinerons les deux cas de topologies de backhaul

- filaire – la capacité maximale est toujours garantie $C_h(j)$,
- sans fil – la capacité maximale pour un lien de longueur r_i tombe $d(r_i) \cdot C_h(j)$:

$$
d(r_i) = \begin{cases} 1, & r_i \leq r_0 \\ (\frac{r_0}{r_i})^n, & \text{otherwise,} \end{cases} \tag{7.21}
$$

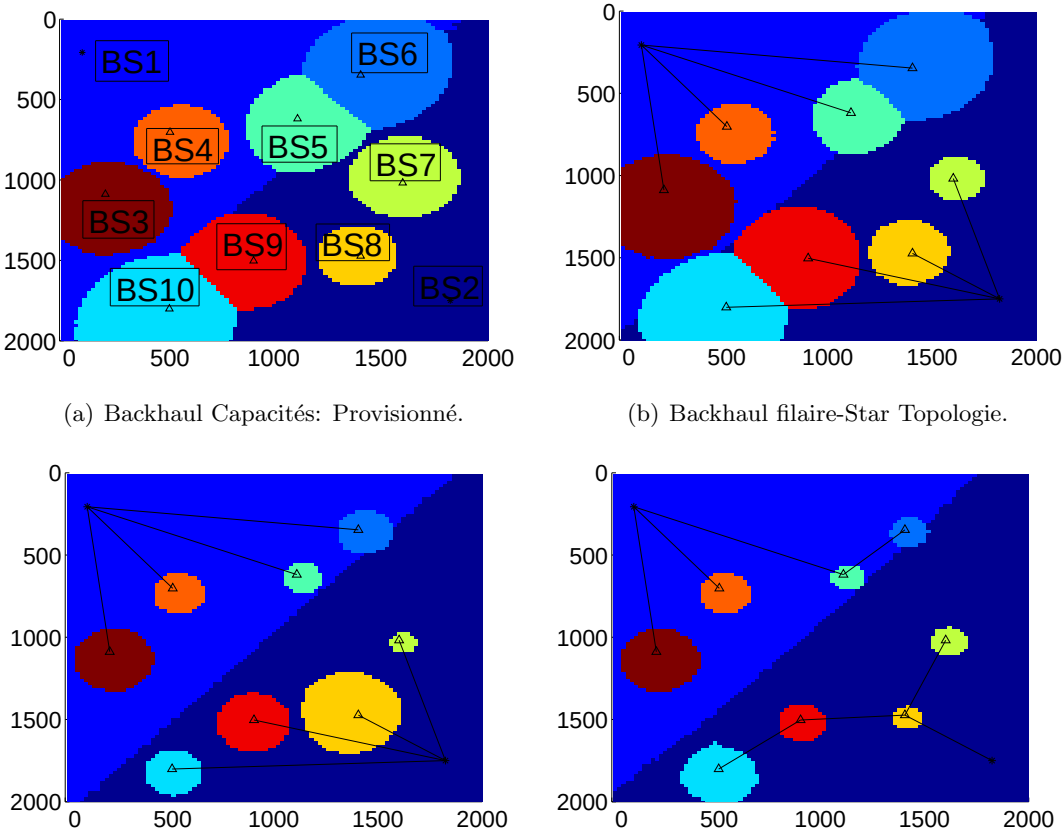
Nous décrivons la zone de couverture dans différents scénarios. Par exemple:

- Descendante (DL) zone quand backhaul $C_h \to \infty$: Figure 7.6(a),

- Descendante (DL) zone quand backhaul $C_h \to 400Mbps$ (filaire – Star topologie) : Figure 7.6(b),

- Descendante (DL) zone quand backhaul $C_h \to 400Mbps$ (sans fil – Star topologie) : Figure 7.6(c),

- Descendante (DL) zone quand backhaul $C_h \to 400Mbps$ (sans fil – Tree topologie) : Figure 7.6(d).



(a) Backhaul Capacités: Provisionné.

(b) Backhaul filaire-Star Topologie.

(c) Backhaul sans fil-Star Topologie.

(d) Backhaul sans fil-Tree Topologie

Figure 7.6: Associations optimales.

Maintenant, nous voulons voir des résultats quantitatifs pour comprendre l'impact des capacités de backhaul non-sifficient sur la performance. Nous supposons le scénario Split UL / DL.

Dans la figure 7.7, 7.8, nous voyons comment C_h affecte DL et UL débit.

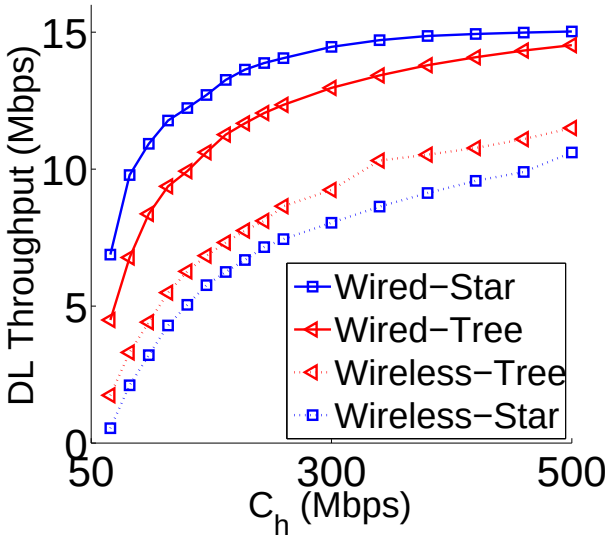


Figure 7.7: Débit descendant pour différentes topologies de backhaul.

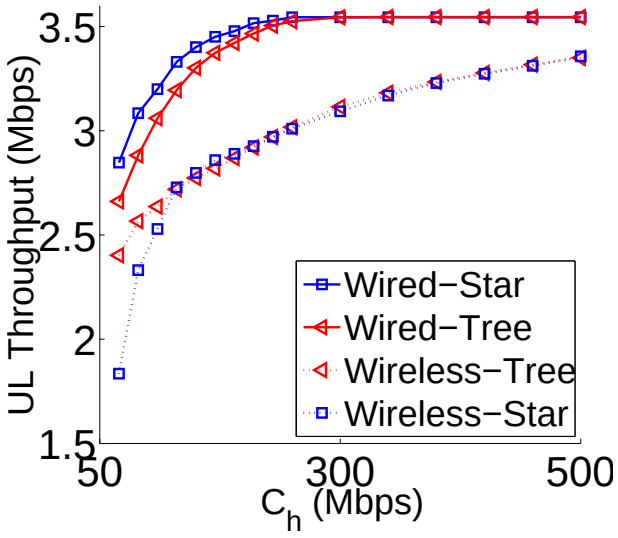


Figure 7.8: Débit montant pour différentes topologies de backhaul.

Dans la figure 7.9, 7.10, nous voyons comment C_h affecte l'efficacité spectrale (spectral efficiency) et l'efficacité d'équilibrage de charge (load balancing or utilization efficiency).

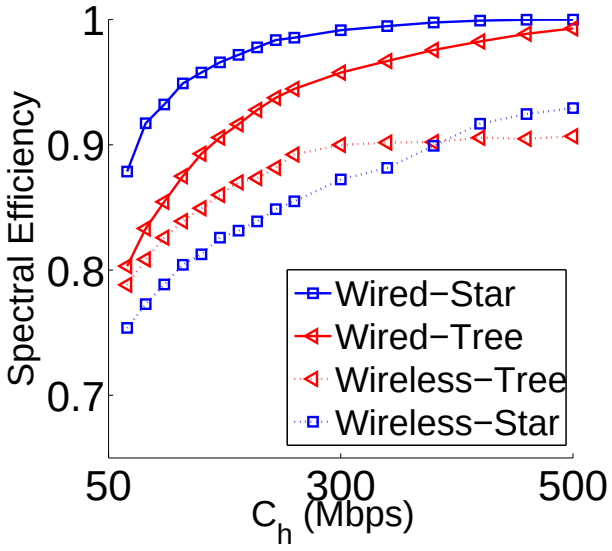


Figure 7.9: Efficacité de la liaison descendante spectrale pour différentes topologies de backhaul (Normalisé).

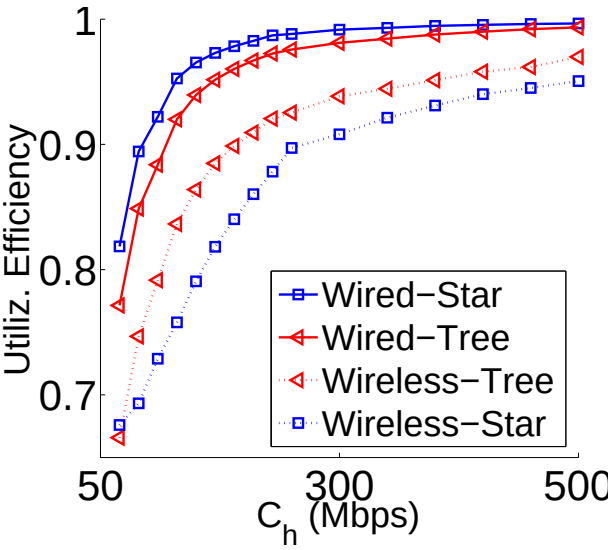


Figure 7.10: Efficacité de la liaison descendante d'équilibrage de charge pour différentes topologies de backhaul (Normalisé).

Résultats similaires pour le scénario conjoint UL / DL (voir Table 7.4).

Table 7.4: Split Vs. Joint UL/DL

Performance	$\tau = 0$	$\tau = 0.5$	$\tau = 1$
Débit DL / UL	6% / 32%	4% / 35%	0% / 37%
DL / UL Spectr. Eff.	4% / 29%	3% / 31%	0% / 33%
DL / UL Uiliz. Eff.	7% / 34%	4% / 38%	0% / 41%

Le travail de ce chapitre corresponde a la publication suivante

- *N. Sapountzis, T. Spyropoulos, N. Nikaein, U. Salim, Optimal Downlink and Uplink User Association in Backhaul-limited HetNets, in Proc. IEEE International Conference on Computer Communications (INFOCOM), San Francisco, CA, USA, 2016.*
  - *Best Presentation Award in Heterogeneous Networks Session.*
- *N. Sapountzis, T. Spyropoulos, N. Nikaein, U. Salim, User Association in HetNets: Impact of Traffic Differentiation and Backhaul Limitations, pending major revision, IEEE/ ACM Transactions on Networking (ToN), May 2016.*

Chapitre 4 - Association des utilisateurs et allocation TDD.
Dans ce chapitre, nous voulons résoudre ensemble les problèmes suivants:

- (Problème 1) Association d'utilisateurs,
- (Problème 2) Accès TDD allocation,
- (Problème 3) Backhaul TDD allocation.

(A. Modèle) Nous étendons notre modèle pour inclure les problèmes 2 et 3. Nous introduisons les paramètres suivants

- (paramètre pour le problème: Accès TDD allocation) ζ_i est le paramètre qui capture la portion de ressources prévue pour DL pour BS i. Donc, $1-\zeta_i$ est le paramètre qui capture la portion de ressources prévue pour UL pour BS i.
- (paramètre pour le problème: Backhaul TDD allocation) Z_i est le paramètre qui capture la portion de ressources prévue pour DL pour backhaul lien j. Donc, $1-Z_j$ est le paramètre qui capture la portion de ressources prévue pour UL pour backhaul lien j.

Nous devons faire attention en obtenant ζ. Nous devons éviter les "interférences croisées" (cross interference), dans le cas où les BS voisins transmettent dans la direction opposée (Figure 7.11).

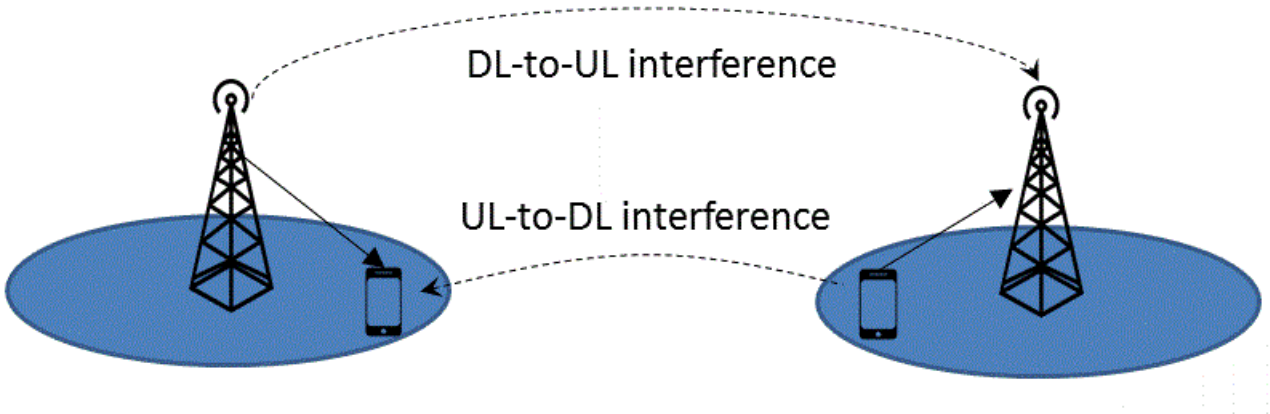


Figure 7.11: "Interférences croisées", dans le cas où les BS voisins transmettent dans la direction opposée.

Nous supposons que les liens de backhaul sont "interférences croisées" libres.

(B. Optimisation) Nous commençons par résoudre les problèmes 1 et 2 (Association d'utilisateurs et Accès TDD allocation).

Definition 18. *(Problèmes 1 et 2) Notre objectif pour association d'utilisateurs et accès TDD allocation est*

$$\begin{aligned} &\underset{\rho,\zeta}{\text{minimize}} \left\{ \phi_\alpha(\rho,\zeta) | (\rho,\zeta) \in \mathcal{F} \right\}, \\ &\text{subject to } \rho_i^D + \rho_j^U \leq 1, \forall i \in \mathcal{B}, j \in \mathcal{C}_i. \end{aligned} \tag{7.22}$$

Lemma 12. *Il s'agit d'un problème d'optimisation biconvexe.*

Nous allons utiliser la théorie de la décomposition pour le résoudre de manière optimale. La schema de l'algorithme est:

Algorithm 3 Decomposition.

1: **Répéter** jusqu'à $\|\zeta^{(k)} - \zeta^{(k-1)}\| < \epsilon$.
2: *(Mettre à jour le problème maître.)*
3: Allocation des ressources: $\zeta \rightarrow$ DL, $1-\zeta \rightarrow$ UL.
4: *(Résoudre les deux sous-problèmes).)*
5: Dériver ρ^{*D} compte tenu des ressources disponibles (ζ).
6: Dériver ρ^{*U} compte tenu des ressources disponibles ($1-\zeta$).

On peut montrer que cet algorithme converge. Et le point de convergence est l'optimum global.

Mettre à jour le problème maître (ζ). Les méthodes de descente suggèrent:

$$\zeta^{(k+1)} = \zeta^{(k)} + t^{(k)} \Delta\zeta^{(k)}, \tag{7.23}$$

tel que $\phi(\rho^*, \zeta^{(k+1)}) < \phi(\rho^*, \zeta^{(k)})$, $\Delta\zeta^{(k)}$ est un *direction de descente*, et $t^{(k)}$ est la *taille de pas*.

Dériver ρ^{*D}, ρ^{*U} compte tenu des ressources disponibles ($\zeta, 1-\zeta$). Voir le théorème prochain.

Theorem 1.6. *Si $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{||\mathcal{B}||}^*)$ désigne le vecteur optimal, les règles optimales sont:*

$$i^D(x) = \arg\max_{i \in \mathcal{B}} \left(\underbrace{c_i^D(x)}_{\text{Connaissances des utilisateurs}} \cdot \overbrace{P_i^D}^{\text{BS message de diffusion}} \right) \tag{7.24}$$

et

$$P_i^D = \frac{\zeta_i \cdot \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D}}{1 + 2\gamma \left(1 - \frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D} \sum\limits_{j \in \mathcal{C}_i} \mathcal{I}_{ij} (\rho_i^{*D} + \rho_j^{*U} - 1)}.$$

Proof. Ici, le produit scalaire:

$$\sum_L \sum_{i\in\mathcal{B}} \left(\frac{1}{\zeta_i^L \left(1-\frac{\rho_i^{*L}}{\zeta_i^L}\right)^{\alpha^L}} + 2\gamma \sum_{j\in\mathcal{C}_i} I(L)(\rho_i^{*L}+\rho_j^{*\tilde{L}}-1) \right)(\rho_i^L-\rho_i^{*L}) =$$

$$\sum_L \sum_{i\in\mathcal{B}} \left(\frac{1+2\gamma\left(1-\frac{\rho_i^{*L}}{\zeta_i^L}\right)^{\alpha^L} \sum\limits_{j\in\mathcal{C}_i} I(L)(\rho_i^{*L}+\rho_j^{*\tilde{L}}-1)}{\zeta_i^L\left(1-\frac{\rho_i^{*L}}{\zeta_i^L}\right)^{\alpha^L}} \right) \cdot \int_{\mathcal{L}} \rho_i^L(x)\left(p_i^L(x)-p_i^{*L}(x)\right)dx =$$

$$= \int_{\mathcal{L}} \sum_L \frac{\lambda^L(x)}{\mu^L(x)} \sum_{i\in\mathcal{B}} \left(\frac{1+2\gamma\left(1-\frac{\rho_i^{*L}}{\zeta_i^L}\right)^{\alpha^L} \sum\limits_{j\in\mathcal{C}_i} I(L)(\rho_i^{*L}+\rho_j^{*\tilde{L}}-1)}{\zeta_i^L c_i^L(x)\left(1-\frac{\rho_i^{*L}}{\zeta_i^L}\right)^{\alpha^L}} \right) \cdot \left(p_i^L(x)-p_i^{L*}(x)\right)dx.$$

Notez que,

$$\sum_{i\in\mathcal{B}} p_i^D(x) \left(\frac{1+2\gamma\left(1-\frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D} \sum\limits_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{*D}+\rho_j^{*U}-1)}{\zeta_i c_i^D(x)\left(1-\frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D}} \right) \geq$$

$$\sum_{i\in\mathcal{B}} p_i^{D*}(x) \left(\frac{1+2\gamma\left(1-\frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D} \sum\limits_{j\in\mathcal{C}_i} \mathcal{I}_{ij}(\rho_i^{*D}+\rho_j^{*U}-1)}{\zeta_i c_i^D(x)\left(1-\frac{\rho_i^{*D}}{\zeta_i}\right)^{\alpha^D}} \right).$$

□

De même pour le problème 3 (Backhaul TDD Allocation) aussi. Nous devons ajouter une troisième échelle de temps qui met à jour Z.

(C. Évaluation)

Maintenant, nous faisons quelques simulations. Nous considérons une zone 2×2 km^2. La figure 7.12 montre une carte de couleur de la demande de trafic hétérogène $\lambda(x)$ ($flows/hour$ par unité de surface) (bleu impliquant trafic faible et rouge élevé) avec 2 hotspots. Nous supposons que cette zone est couverte par 1 BS macro et 3 SC.

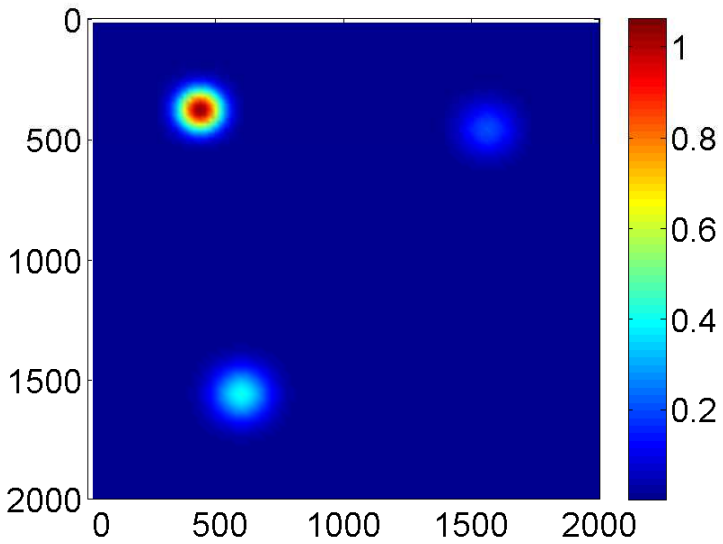


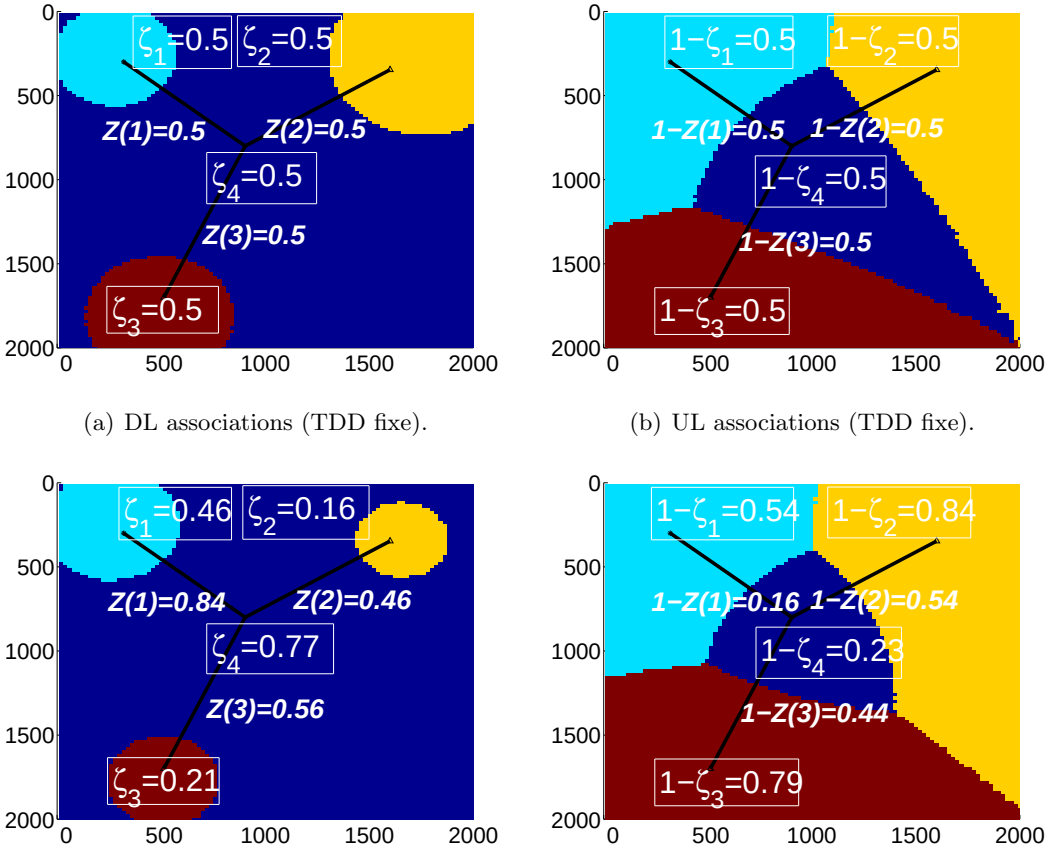
Figure 7.12: Taux d'arrivée.

Couverture Nous décrivons la zone de couverture dans différents scénarios (ζ scénarios). Par exemple:

- Descendante (DL) zone quand nous considérons le (fixé) scénario: $\zeta_i = 0.5\ \forall i \in \mathcal{B}$: Figure 7.13(a),
- Montante (UL) zone quand nous considérons le (fixé) scénario: $1-\zeta_i = 0.5\ \forall i \in \mathcal{B}$: Figure 7.13(b),
- Descendante (DL) zone quand nous considérons le (dynamique) scénario: dynamique ζ_i $\forall i \in \mathcal{B}$: Figure 7.13(c),
- Montante (UL) zone quand nous considérons le (dynamique) scénario: dynamique $1-\zeta_i$ $\forall i \in \mathcal{B}$: Figure 7.13(d),

Maintenant, nous voulons voir des résultats quantitatifs pour comprendre l'impact des capacités de backhaul non-sifficient sur la performance.

Dans la figure 7.14, 7.15, nous voyons comment τ affecte DL et UL débit.



(a) DL associations (TDD fixe).

(b) UL associations (TDD fixe).

(c) DL associations (TDD dynamique).

(d) UL associations (TDD dynamique).

Figure 7.13: Associations optimales et TDD fixe/dynamique ($\tau = 0.5$).

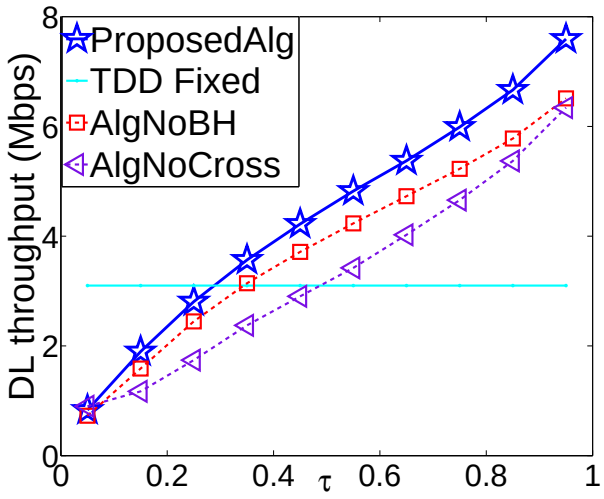


Figure 7.14: Débit descendant pour différentes valeurs de τ.

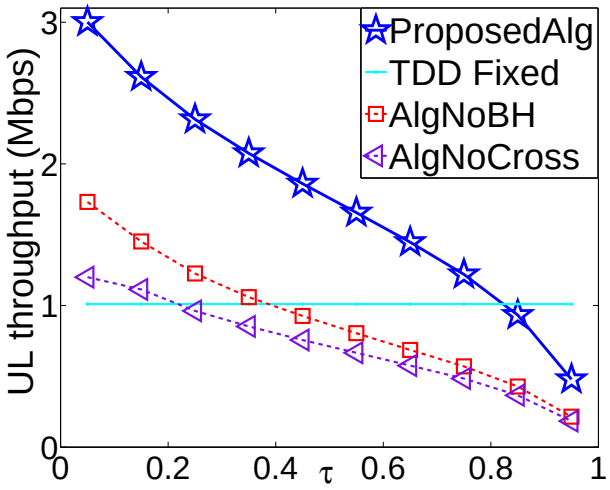


Figure 7.15: Débit montant pour différentes valeurs de τ.

Le travail de ce chapitre corresponde a la publication suivante

- *N. Sapountzis, T. Spyropoulos, N. Nikaein, U. Salim, HoP: Hierarchize then optimize: A distributed framework for user association and flexible TDD allocation for access and backhaul networks, Tech-Report RR-16-328, Eurecom, 2016.*

Chapter 5 - Minimisation d'énergie.

Dans ce chapitre, nous voulons minimiser l'énergie du réseau. Nous essayons donc de désactiver BSs en considérant:

- (1) l' échec de la couverture
- (2) la robabilité de blocage
- (3) la latence.

(A. Modèle) Donc, notre objectif est de construire 3 contraintes de QoS différentes qui tiennent compte des métriques indicateurs ci-dessus. Nous utilisons principalement des chaînes de Markov pour les dériver.

Proposition 4. (Contrainte I) *Une BS ne peut pas être désactivée si l'utilisateur moyen associé à celle-ci subit une probabilité de panne de couverture, pendant la période de déconnexion X, qui dépasse un seuil p_f. Cette probabilité est donnée par:*

$$\frac{\sum\limits_{i\in\mathcal{N}_{AU}} P_{out}(r_{iJ(i)}) + \sum\limits_{i\in\mathcal{N}_{CU}} P_{act}^{CU}(X)P_{out}(r_{iJ(i)}) + N_{DU}P_{act}^{DU}(X)P_{out}^{DU}}{N_{AU}+N_{CU}+N_{DU}}. \tag{7.25}$$

Proposition 5. (Contrainte II) *Supposons que la probabilité de blocage maximale souhaitée est donnée pour les flux dédiés, définis comme p_{block}. Une BS donnée ne peut être désactivée que si l'inégalité suivante est valable pour toutes les BS voisines auxquelles les utilisateurs de la BS déconnectée sont remis:*

$$\frac{\sum_{n=0}^{\frac{X}{\Delta t}} P_{s,k}^{n}}{X/\Delta t} \leq p_{block}. \tag{7.26}$$

Proposition 6. (Contrainte III) *Supposons un délai maximum souhaité pour les flux d'effort optimal, D_{max}. Une BS donnée ne peut être désactivée que si l'inégalité suivante est valable pour toutes les BS voisines auxquelles les utilisateurs de la BS déconnectée sont remis:*

$$\sum_{t=0}^{X/\Delta t}\sum_{n=1}^{\infty}\frac{P_{s,n}^{t/\Delta t}D^{n}(1/\mu^{b})}{X/\Delta t} \leq D_{max}.$$

(B. Optimisation) Nous exécutons un algorithme itératif. À chaque étape, nous essayons de désactiver une BS si les contraintes ci-dessus ne sont pas violées. L'ordre des BS est basé sur la charge BS.

(C. Évaluation)

Maintenant, nous illustrons quelques résultats quantitatifs. Nous commençons par illustrer l'impact des seuils prédéterminés sur les économies d'énergie. Voir la figure 7.16 pour la première contrainte.

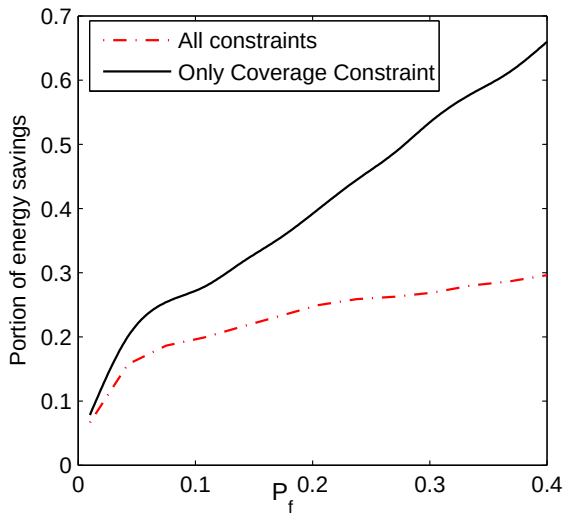


Figure 7.16: Économies d'énergie versus p_f.

Un autre paramètre intéressant est la durée de la période d'arrêt. Nous illustrons son impact sur les économies d'énergie à la figure 7.17.

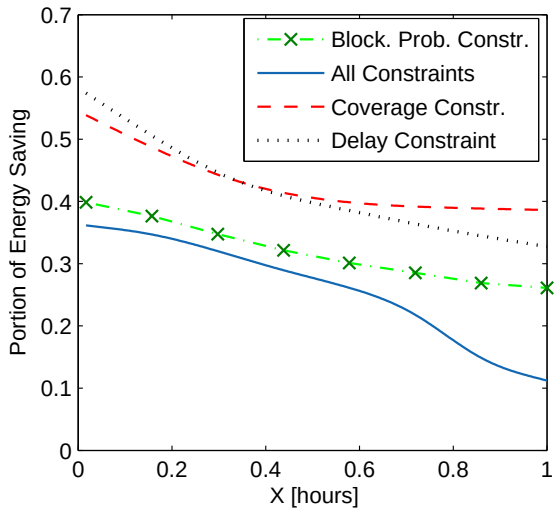


Figure 7.17: Économies d'énergie versus X.

Le travail de ce chapitre corresponde a la publication suivante

- *N. Sapountzis, T. Spyropoulos, N. Nikaein, U. Salim, Reducing the energy consumption of small cell networks subject to QoE constraints, in Proc. IEEE Global Communications*

(GLOBECOM) Conference, Austin, TX, USA, 2014.

- *D. Wang, E. Karathanaras, A. Quddus, N. Sapountzis, L. Cominardi, F. Kuo, P. Rost, C.J. Bernardos, I. Berberana, SDN-based Joint Backhaul and Access design for Efficient Network Layer Operations, in Proc. IEEE European Conference on Networks and Communications (EuCNC), Paris, France, 2015.*

Bibliography

[1] N. Radio, Y. Zhang, M. Tatipamula, and V. K. Madisetti, "Next-generation applications on cellular networks: Trends, challenges, and solutions," *IEEE Proc.*, 2012.

[2] *Small cell backhaul requirements*, NGMN Alliance, 2014.

[3] *Backhaul technologies for small cells*, Small Cell Forum, 2014.

[4] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser mimo systems," *IEEE Transactions on Communications*, 2013.

[5] I. Ku, C. X. Wang, and J. Thompson, "Spectral-energy efficiency tradeoff in relay-aided cellular networks," *IEEE Transactions on Wireless Communications*, 2013.

[6] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE Journal on Selected Areas in Comm.*, 2011.

[7] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Communications Magazine*, 2010.

[8] T. Chen, Y. Yang, H. Zhang, H. Kim, and K. Horneman, "Network energy saving technologies for green wireless access networks," *IEEE Wireless Communications*, 2011.

[9] L. Xiang, F. Pantisano, R. Verdone, X. Ge, and M. Chen, "Adaptive traffic load-balancing for green cellular networks," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, 2011.

[10] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed alpha -optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking*, 2012.

[11] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in hetnets: Old myths and open problems," *IEEE Wireless Communications*, 2014.

[12] Z. Kaleem, B. Hui, and K. Chang, "Qos priority-based dynamic frequency band allocation algorithm for load balancing and interference avoidance in 3gpp lte hetnet," *EURASIP Journal on Wireless Communications and Networking*, 2014.

[13] K. Son, S. Chong, and G. D. Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *IEEE Transactions on Wireless Communications*, 2009.

[14] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems.* Imperial college press, 2010.

[15] I. Siomina and S. Wanstedt, "The impact of QoS support on the end user satisfaction in lte networks with mixed traffic," in *Proc. of IEEE PIMRC*, 2008.

[16] O. Brickley, S. Rea, and D. Pesch, "Load balancing for qos optimisation in wireless lans utilising advanced cell breathing techniques," in *Proc. IEEE Vehicular Technology Conference*, May 2005.

[17] L. Chiaraviglio, D. Ciullo, M. Meo, and M. Marsan, "Energy-efficient management of UMTS access networks," in *Teletraffic Congress*, 2009.

[18] A. Mesodiakaki, F. Adelantado, L. Alonso, and C. Verikoukis, "Joint uplink and downlink cell selection in cognitive small cell heterogeneous networks," in *Proc. IEEE Globecom*, 2014.

[19] T. Bonald and A. Proutiere, "Wireless downlink data channels: User performance and cell dimensioning," in *Proc. Mobile Computing and Networking (MobiCom)*, 2003.

[20] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, 2001.

[21] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. Vehicular Technology Conference*, 2000.

[22] H. S. Dhillon, M. Kountouris, and J. G. Andrews, "Downlink mimo hetnets: Modeling, ordering results and performance analysis," *IEEE Transactions on Wireless Communications*, 2013.

[23] A. A. Abdulkafi, T. S. Kiong, D. Chieng, A. Ting, and J. Koh, "Energy efficiency improvements in heterogeneous network through traffic load balancing and sleep mode mechanisms," *Wireless Personal Communications*, 2014.

[24] P. Hande, S. Patil, and H. Myung, "Distributed load-balancing in a multi-carrier wireless system," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, 2009.

[25] T. Han and N. Ansari, "A traffic load balancing framework for software-defined radio access networks powered by hybrid energy sources," *IEEE/ACM Transactions on Networking*, 2016.

[26] I. Siomina and D. Yuan, "Load balancing in heterogeneous lte: Range optimization via cell offset and load-coupling characterization," in *Proc. IEEE ICC*, 2012.

[27] P. Fotiadis, M. Polignano, D. Laselva, B. Vejlgaard, P. Mogensen, R. Irmer, and N. Scully, "Multi-layer mobility load balancing in a heterogeneous lte network," in *Proc. IEEE Vehicular Technology Conference (VTC Fall)*, 2012.

[28] R. Amin, J. Martin, J. Deaton, L. A. DaSilva, A. Hussien, and A. Eltawil, "Balancing spectral efficiency, energy consumption, and fairness in future heterogeneous wireless systems with reconfigurable devices," *IEEE Journal on Selected Areas in Communications*, 2013.

[29] A. Khandekar, N. Bhushan, J. Tingfang, and V. Vanghi, "LTE-Advanced: Heterogeneous networks," in *Proc. European Wireless Conference*, 2010.

[30] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications*, 2011.

[31] M. Al-Rawi, "A dynamic approach for cell range expansion in interference coordinated lte-advanced heterogeneous networks," in *Proc. IEEE Communication Systems (ICCS)*, 2012.

[32] P. Ökvist and A. Simonsson, "Lte hetnet trial - range expansion including micro/pico indoor coverage survey," in *Proc. of IEEE Vehicular Technology Conference (VTC Fall)*, 2012.

[33] P. T. V. Bhuvaneswari, S. Indu, N. L. Shifana, D. Arjun, and A. S. Priyadharshini, "An analysis on cell range expansion in 4g lte networks," in *Proc. of Signal Processing, Communication and Networking (ICSCN)*, 2015.

[34] Z. Ning, Q. Song, L. Guo, M. Dai, and M. Yue, "Dynamic cell range expansion-based interference coordination scheme in next generation wireless networks," *China Communications*, 2014.

[35] H. Dahrouj and W. Yu, "Coordinated beamforming for the multicell multi-antenna wireless system," *IEEE Transactions on Wireless Communications*, 2010.

[36] S. He, Y. Huang, S. Jin, and L. Yang, "Coordinated beamforming for energy efficient transmission in multicell multiuser systems," *IEEE Transactions on Communications*, 2013.

[37] J. Lee, Y. Kim, H. Lee, B. L. Ng, D. Mazzarese, J. Liu, W. Xiao, and Y. Zhou, "Coordinated multipoint transmission and reception in LTE-advanced systems," *IEEE Communications Magazine*, 2012.

[38] H. A. Mustafa, M. Z. Shakir, Y. A. Sambo, K. A. Qaraqe, M. A. Imran, and E. Serpedin, "Spectral efficiency improvements in HetNets by exploiting device-to-device communications," in *Proc. IEEE Globecom Workshops*, 2014.

[39] H. Boostanimehr and V. Bhargava, "Unified and distributed qos-driven cell association algorithms in heterogeneous networks," *IEEE Transactions on Wireless Communications*, 2015.

[40] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Transactions on Wireless Communications*, 2013.

[41] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking*, 2000.

[42] T. Han and N. Ansari, "Smart grid enabled mobile networks: Jointly optimizing bs operation and power distribution," in *Proc. IEEE ICC*, 2014.

[43] J. Bartelt, A. Fehske, H. Klessig, G. Fettweis, and J. Voigt, "Joint bandwidth allocation and small cell switching in heterogeneous networks," in *Proc. IEEE Vehicular Technology Conference*, 2013.

[44] D. Bharadia, E. McMilin, and S. Katti, "Full duplex radios," in *Proceedings of the ACM SIGCOMM*, 2013.

[45] V. Pauli and E. Seide, *Dynamic TDD for LTE-A and 5G*, 2015.

[46] 3GPP, "TS 36.300, Release 13 (version 13.2.0)," 2016.

[47] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, "Downlink and uplink decoupling: A disruptive architectural design for 5G networks," in *Proc. IEEE Globecom*, 2014.

[48] M. Ding, D. L. Perez, A. V. Vasilakos, and W. Chen, "Dynamic TDD transmissions in homogeneous small cell networks," in *Proc. IEEE ICC Communications Workshops*, 2014.

[49] H. Ji, Y. Kim, S. Choi, J. Cho, and J. Lee, "Dynamic resource adaptation in beyond LTE-A TDD heterogeneous networks," in *Proc. IEEE ICC Communications Workshops*, 2013.

[50] M. S. ElBamby, M. Bennis, W. Saad, and M. Latva-aho, "Dynamic uplink-downlink optimization in TDD-based small cell networks," in *International Symposium on Wireless Communications Systems*, 2014.

[51] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental trade-offs on green wireless networks," *IEEE Communications Magazine*, 2011.

[52] T. Edler and S. Lundberg, "Energy efficiency enhancements in radio access networks," *Ericsson Review*, 2004.

[53] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Communications Magazine*, 2011.

[54] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Communications Surveys Tutorials*, 2011.

[55] S. Deng and H. Balakrishnan, "Traffic-aware techniques to reduce 3g/lte wireless energy consumption," in *in Proc. of IEEE Emerging Networking Experiments and Technologies Conference*, ser. CoNEXT, 2012.

[56] M. Gupta, S. C. Jha, A. T. Koc, and R. Vannithamby, "Energy impact of emerging mobile internet applications on lte networks: issues and solutions," *IEEE Communications Magazine*, 2013.

[57] A. Bousia, E. Kartsakli, L. Alonso, and C. Verikoukis, "Energy efficient base station maximization switch off scheme for lte-advanced," in *in Proc. IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2012.

[58] S. Bhaumik, G. Narlikar, S. Chattopadhyay, and S. Kanugovi, "Breathe to stay cool: Adjusting cell sizes to reduce energy consumption," in *in Proc. ACM SIGCOMM Workshop on Green Networking*, ser. Green Networking, 2010.

[59] P. Frenger, P. Moberg, J. Malmodin, Y. Jading, and I. Godor, "Reducing energy consumption in lte with cell dtx," in *in Proc. IEEE Vehicular Technology Conference (VTC Spring)*, 2011.

[60] T. Han and N. Ansari, "Ice: Intelligent cell breathing to optimize the utilization of green energy," *IEEE Communications Letters*, 2012.

[61] W. Tomaselli, D. Sabella, V. Palestini, and V. Bernasconi, "Energy efficiency performances of selective switch off algorithm in lte mobile networks," in *Proc. IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2013.

[62] G. Micallef, P. Mogensen, and H. O. Scheck, "Cell size breathing and possibilities to introduce cell sleep mode," in *in Proc. Wireless Conference (EW)*, 2010.

[63] A. Bousia, E. Kartsakli, L. Alonso, and C. Verikoukis, "Dynamic energy efficient distance-aware base station switch on/off scheme for lte-advanced," in *Proc. IEEE Globecom*, 2012.

[64] G. Cili, H. Yanikomeroglu, and F. R. Yu, "Cell switch off technique combined with coordinated multi-point (comp) transmission for energy efficiency in beyond-lte cellular networks," in *Proc. IEEE ICC*, 2012.

[65] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Transactions on Wireless Communications*, 2013.

[66] A. Fehske, H. Klessig, J. Voigt, and G. Fettweis, "Concurrent load-aware adjustment of user association and antenna tilts in self-organizing radio networks," *IEEE Transactions on Vehicular Technology*, 2013.

[67] H. Kim, H. Y. Kim, Y. Cho, and S.-H. Lee, "Spectrum breathing and cell load balancing for self organizing wireless networks," in *Proc. IEEE Communications Workshops*, 2013.

[68] A. Elwalid and D. Mitro, "Design of generalized processor sharing schedulers which statistically multiplex heterogeneous QoS classes," in *Proc. IEEE Infocom*, 1999.

[69] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Communication Surveys and Tutorials*, 2013.

[70] J. Ghimire and C. Rosenberg, "Revisiting scheduling in heterogeneous networks when the backhaul is limited," *IEEE Journal on Selected Areas in Communications (JSAC)*, 2015.

[71] D. Chen, T. Quek, and M. Kountouris, "Backhauling in heterogeneous cellular networks: Modeling and tradeoffs," *IEEE Transactions on Wireless Communications*, 2015.

[72] M. Shariat, E. Pateromichelakis, A. Quddus, and R. Tafazolli, "Joint TDD backhaul and access optimization in dense small cell networks," *IEEE Transactions on Vehicular Technology*, 2013.

[73] A. K. Gupta, H. S. Dhillon, S. Vishwanath, and J. G. Andrews, "Downlink coverage probability in mimo hetnets with flexible cell selection," in *Proc. IEEE Globecom*, 2014.

[74] O. Simeone, O. Somekh, H. Poor, and S. Shamai (Shitz), "Downlink multicell processing with limited-backhaul capacity," *in EURASIP Journal on Advances in Signal Processing*, 2009.

[75] S. Corroy, L. Falconetti, and R. Mathar, "Dynamic cell association for downlink sum rate maximization in multi-cell heterogeneous networks," in *Proc. of IEEE ICC*, 2012.

[76] S. Aalto and L. Pasi, *Impact of Size-Based Scheduling on Flow Level Performance in Wireless Downlink Data Channels*, 2007.

[77] 3GPP, "TR 36.842, Release 12 (version 12.0.0)," 2014.

[78] P. Rost, A. Maeder, and X. Perez-Costa, "Asymmetric uplink-downlink assignment for energy-efficient mobile communication systems," in *Vehicular Technology Conference (VTC Spring)*, 2012.

[79] S. E. Nai, T. Quek, and M. Debbah, "Shadowing time-scale admission and power control for small cell networks," in *Proc. IEEE Wireless Personal Multimedia Communications (WPMC)*, Sept 2012.

[80] S. Liu and J. Virtamo, "Inter-cell coordination with inhomogeneous traffic distribution," in *Proc. IEEE Next G. Int. Design and Eng.*, 2006.

[81] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE Journal on Selected Areas in Communications (JSAC)*, 2010.

[82] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets," *IEEE/ACM Transactions on Networking*, 2014.

[83] M. Usman, A. Västberg, and T. Edler, "Energy efficient high capacity hetnet by offloading high qos users through femto," in *17th IEEE International Conference on Networks*, 2011.

[84] P. Donegan, "Small cell backhaul: What, why and how?" 2012.

[85] O. Tipmongkolsilp, S. Zaghloul, and A. Jukan, "The evolution of cellular backhaul technologies: Current issues and future trends," *in IEEE Communications Surveys and Tutorials*, 2011.

[86] S. E. Nai, T. Quek, and M. Debbah, "Shadowing time-scale admission and power control for small cell networks," in *in Wireless Personal Multimedia Communications (WPMC)*, 2012.

[87] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "On the effectiveness of single and multiple base station sleep modes in cellular networks," *in Comp. Networks*, 2013.

[88] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. IEEE Infocom*, 2006.

[89] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *Proc. IEEE Infocom*, 2003.

[90] A. Mesodiakaki, F. Adelantado, L. Alonso, and C. Verikoukis, "Energy-efficient user association in cognitive heterogeneous networks," *IEEE Communications Magazine*, 2014.

[91] T. Bonald, S. Borst, N. Hegde, M. Jonckheere, and A. Proutiere, "Flow-level performance and capacity of wireless networks with user mobility," 2009.

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

[93] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "An analytical framework for optimal downlink-uplink user association in hetnets with traffic differentiation," in *Proc. IEEE Globecom*, 2015.

[94] S. Sesia, I. Toufik, and B. M., *LTE - The UMTS Long Term Evolution: From Theory to Practice, 2nd Edition*. Wiley, 2011.

[95] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "Optimal downlink and uplink user association in Backhaul-limited HetNets," in *Proc. IEEE Infocom*, 2016.

[96] 3GPP, "TR 36.931 Release 13 (version 13.2.0)," 2016.

[97] E. Metsala and J. Salmelin, *Mobile Backhaul*. Wiley, 2012.

[98] Z. G. Raphael T. Haftka, *Elements of Structural Optimization*. Springer Netherlands, 1992.

[99] D. G. Luenberger, Y. Ye *et al.*, *Linear and nonlinear programming*. Springer, 1984, vol. 2.

[100] Alcatel-Lucet, "Mobile bakhaul architecture for hetnet, https://www.alcatel-lucent.com/solutions/mobile-backhaul," 2015.

[101] G. 36.133, "Evolved universal terrestrial radio access (E-UTRA) and radio access network (E-UTRAN); overall descriptiom," 2012.

[102] R. Sivaraj, I. Broustis, N. K. Shankaranarayanan, V. Aggarwal, R. Jana, and P. Mohapatra, "A QoS-enabled holistic optimization framework for LTE-advanced heterogeneous networks," in *Proc. IEEE Infocom*, 2015.

[103] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: a survey and extensions," *Mathematical Methods of Operations Research*, 2007.

[104] D. Bertsekas, *Nonlinear Programming.* Athena Scientific, 1995.

[105] R. E. Wendell and A. P. Hurter, "Minimization of a non-separable objective function subject to disjoint constraints," *Journal on Operation Research*, 1976.

[106] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear gauss-seidel method under convex constraints," *Operations Research Letters*, 2000.

[107] C. A. Floudas, *Deterministic Global Optimization: Theory, Methods and Applications.* Springer-Verlag New York, Inc., 2000.

[108] C. Hildreth, "A quadratic programming procedure," *Naval Research Logistics Quarterly*, 1957.

[109] Cambridge Broadband Networks, "Solutions mobile backhaul," 2015.

[110] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, 2006.

[111] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points- online stochastic gradient for tensor decomposition," in *Proc. of the 28th Conference on Learning Theory*, 2015.

[112] Z. Shen, A. Khoryaev, E. Eriksson, and X. Pan, "Dynamic uplink-downlink configuration and interference management in TD-LTE," *IEEE Communications Magazine*, 2012.

[113] T. K. H. Guan and P. Merz, "Discovery of cloud-RAN," *in Cloud-RAN Workshop*, April 2010.

[114] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *in IEEE Wireless Communications*, 2011.

[115] N. Sapountzis, S. Sarantidis, T. Spyropoulos, N. Nikaein, and U. Salim, "Reducing the energy consumption of small cell networks subject to QoE constraints," in *Proc. IEEE Globecom*, 2014.

[116] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental trade-offs on green wireless networks," *IEEE Communications Magazine*, 2011.

[117] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *in IEEE Communications Magazine*, 2010.

[118] C. wei Tan, "Optimal power control in rayleigh-fading heterogeneous networks," in *IEEE Infocom*, April 2011.

[119] T. Bonald and A. Proutiere, "Wireless downlink data channels: User performance and cell dimensioning," in *Proc. of the International Conference on Mobile Computing and Networking*, ser. MobiCom, 2003.

[120] S. M. Ross, *Introduction to Probability Models*, 10th edition, 2009.

[121] N. S. Networks, "Improving 4G coverage and capacity indoors at hotspots with lte femtocells," *White Paper*, 2011.