ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2017 ENST 0040

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

George ARVANITAKIS

le 27 september 2017

Analytical Modeling of Flow-level Performance of Randomly Placed Wireless Networks

Directeur de thèse : **Florian KALTENBERGER**

Le jury composé de :

Dr. Sara ALOUF	Président du jury
Prof. Alexander PROUTIERE	Rapporteur
Prof. Luiz DASILVA	Rapporteur
Prof. Raymond KNOPP	Examinateur
Dr. Marios KOUNTOURIS	Examinateur
Prof. Andreas POLYDOROS	Examinateur
Prof. Thrasyvoulos SPYROPOULOS	Examinateur
Prof. Florian KALTENBURGER	Directeur de thése

T H È S E

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

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIE
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

George ARVANITAKIS

Analytical Modeling for Flow-level Performance of Randomly Placed Wireless Networks

Defense scheduled on September 27^{th} 2017

before a committee composed of:

Dr. Sara Alouf	President of the Jury
Prof. Alexander Proutiere	Reporter
Prof. Luiz DaSilva	Reporter
Dr. Marios Kountouris	Examiner
Prof. Raymond Knopp	Examiner
Prof. Andreas Polydoros	Examiner
Prof. Thrasyvoulos Spyropoulos	Examiner
Prof. Florian Kaltenberger	Thesis Supervisor

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIE
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 Èlectronique (Paris)

Présentée par

George ARVANITAKIS

Analytical Modeling of Flow-level Performance of Randomly Placed Wireless Networks

Soutenance de thése prévue le 27 september 2017

devant le jury composé de:

Dr. Sara Alouf	Président du jury
Prof. Alexander Proutiere	Rapporteur
Prof. Luiz DaSilva	Rapporteur
Prof. Raymond Knopp	Examinateur
Dr. Marios Kountouris	Examinateur
Prof. Andreas Polydoros	Examinateur
Prof. Thrasyvoulos Spyropoulos	Examinateur
Prof. Florian Kaltenberger	Directeur de thése

Abstract

The recent evolution of mobile communications and the widespread use of "smart" mobile devices have radically changed the behavior and the needs of the mobile user. In the developed world, people that use their mobile devices just to place a call or text an sms could be characterized as endangered species. Ubiquitous access to Internet, video/song streaming and upload/download data flows on the fly are some of the modern demands of the mobile user. It would not be exaggeration to say that modern users have almost the same demands regardless if they are connected through wire to the network from their personal computers or if they have established a wireless connection through their cellular devices.

On the one hand, the overall network load increases. On the other hand base stations have limited resources due to the fact that they are able to operate only to a limited part of the electromagnetic spectrum. Some (temporal) solutions for the aforementioned problem are the expansion of frequency operational bands (mm wave) or the use of more antennas (massive MIMO) or the denser deployment of small cells. In this thesis we are interested in the promising case of denser small cells which will be tighter integrated with the macro cell. Unfortunately, to deploy optimally a small cell network is not trivial. A small cell network is usually deployed in an ad-hoc style and not all at once, so a part of the network exists already and cannot be planned. Additionally, there are natural obstacles and physical limitations that do not allow to deploy base stations wherever we want. Thus, the small cell network's topology is quite different from the traditional (macro cell) well-structured one.

From our point of view, the future of mobile communications will approach the following: A mass of users will cause a nondescript data traffic that will be served from an irregular planed and heterogeneous network. This chaotic picture however makes the problem of network modeling and performance analysis extremely challenging.

To this end, the primary focus of this thesis is to build up an analytical framework in order to analyze the performance of a randomly placed network, which serves randomly placed users. We are interested in the flow-level performance of this network. The users are not assumed to be saturated, instead, each one generates flows according to a Poisson point process (PPP). Additionally, our framework does not assume that neighboring base stations are saturated or that they constantly contribute to the interference in order to obtain closed form results, but deals with the problem of coupling between network load and user quality of service. The second goal is to propose, based on this analysis, some general design guidelines and/or insights about specific communication scenarios.

Specifically, Chapter 1 provides a short introduction of flow-level performance of a random placed network, the motivation of our work as well as and the main contributions of our work. Subsequently, Chapter 2 presents a brief introduction to the two main mathematical tools that we used in this thesis (stochastic geometry and queueing theory) and an overview of the related

work.

In Chapter 3 we model and study the PHY and the MAC layer of LTE and WIFI radio access techniques (RAT). The results of this chapter such as the rate thresholds and the scheduler policy for each RAT will be used widely in the chapters that follow.

We derive the flow-level performance of a random placed network which serves randomly placed users in Chapter 4. Our model does not assume saturated BS or Users and derives network performance metrics such as average users delay, network's load and BS's congested probability. Finally we applied our framework to the popular LTE and WIFI RATs and we provide useful insights about how the PHY characteristics affect the users' flow-level performance and the tensions between the users' experience (e.g. flow delay) and the network's parameters (e.g. BS density).

In Chapter 5 we study the energy consumption of randomly placed networks. We mainly emphasize into how the energy consumption of the network is scaling with respect to the density of the BS. Additionally we provide insights about the tradeoff between energy consumption and users delay. Additionally, in this chapter we derive a simple rule that indicates when the decrease of the BS leads to decrease of the energy consumption as well. Furthermore, we consider the case where the density of the users is drastically changing in a given area, and we answer the question, how the BS's density should adapt to the new situation.

Chapter 6 analyzes the performance of orthogonal heterogeneous networks (HetNet) providing flow-level metrics for such systems (delay, load, congested probability). We model mathematically popular user association criteria, such as Off-load, Max-SINR association, and Min-Delay association and we apply them to the case of a popular 2-tier HetNet scenario, based on LTE and WiFi, in order to further understand their performance differences.

Finally, we conclude our findings and discuss future research direction in Chapter 7.

Acknowledgements

The present thesis marks the end of a journey. Looking back, I have this weird feeling of relative time that cannot be captured by any mathematical equation. How it is possible to feel that the same event was a long process and at the same time that it lasted only for one breath?

I tried a lot to decompose this feeling and understand what is beyond that. The duration of the journey was long because I learned a lot of things, I tackled circumstances that were totally new for me and I evolved a lot as a scientist but most importantly as a person. On the other hand, the feeling that the journey was short stems from the fact that it was a pleasant trip. The dilated time caused by the (hard) process of evolution and knowledge expansion, as Hegel said, is a lonely and personal case. However, the contraction of time was caused by other people, who, at this point, I would like to thank.

To begin with, I would like to thank my supervisor Prof. Florian Kaltenberger – it is certain that nothing would have ever happened without him. Florian has supported and protected me a lot (scientifically – and not). I would like to express my regards to him mainly because he offered me the privilege to search, get lost, and, finally, to find my way. This usually painful but very valuable procedure needs time; I have come to admire how a person that works under time pressure has the internal will to not only refrain from transmitting this pressure, but instead to provide you with the time and space in order to learn and evolve.

Furthermore, I want to specially thank Prof. Akis Spyropoulos. All of those years would have been much more difficult without him. Except for the scientific support that he has provided me, Akis was always the Deus ex Machina when things went tough: his couch became my bed when I was "homeless", he set everything up on my sudden, after midnight, hospitalization. He was my first call after my bike crashed. So many other stories and memories without him would not be pleasant. He is a real problem solver...

I would also like to express my gratitude to Prof. Petros Elia. I am sure that he never understood how valuable my weekly short visits at his office were (I heavily doubt that he even remembers them).

My friend from my undergrad years, Praxitelis – I was lucky enough that he did his PhD in theoretical physics at the same period in Paris. Our endless conversations until sunrise are part of my PhD; even if they are not included in this manuscript.

My former colleagues and friends: Ropokis, Pavlos, Sap, Robin, Ayse, Waki, Dimitris, Pantelis, Katsalis, TTT, Miltos, Reza and Dimitra – who created a home-felt environment full of intriguing conversations (not only scientifically oriented). Most of all, I would like to thank my officemate, Panos, with his outstanding capability to convince you that he is the kindest person in the universe and, at the same time, to mock you right in your face.

My friend, Dorin; we both graduated from the Physics department of Athens being totally

ignorant to each other, in order to meet for the first time in Cote d' Azur.

My friends, George, Vagos, Dr and Ira: together we explored the futility of existence in Nice, Berlin, Amsterdam, Munich, Bilbao, Firenze (but why?).

My friend, Jorli, who, even if she never aimed for that, helped me a lot through all those years with her unique way of caring and her indifference about everything at the same time.

I would like to thank my flat mates George, Beren, Christos and Thodoris that proved in action that Sartre was wrong – "paradise" is the others!!! Together we made small and almost abandoned places to smell like home.

There is a large list of my friends from my home city that were always with me. Specifically, Kostaras who really supported me from the first moment - literally - (Him and I we did the first travel from Greece to France with a motorbike and all of my belongings in a backpack), Lef and Kazos, together our ups but, most important our downs, were always a reason for celebration, almost all of our fears, sorrows and anxieties could demystify with blasphemous friends!!! Finally, I would like to thank Dr. Dagres, one of the most reasonable persons that I know, for his unreasonable love and support all of those years.

Last but not least, I would like to thank my parents, Argyro and Spyros, for their unconditional love and support over the years.

If the journey that you have just finished fulfilled you as person in such a level that when you arrive on the land you have the fire, the need and the impatience to start a new trip, we should conclude that the journey was successful.

Thank you all
George

Contents

List of Figures

List of Tables

Acronyms

Here is the list of acronyms used in the text.

3GPP	Fourth generation.
3G	Third Generation of wireless mobile telecommunications technology.
4G	Fourth Generation of wireless mobile telecommunications technology.
5G	Fifth Generation of wireless mobile telecommunications technology.
ACK	Acknowledgment.
AP	Access Point.
BLER	Block Error Rate.
BS	Base Station.
ccdf	Complementary Cumulative Distribution Function.
cdf	Cumulative Distribution Function.
CSI	Channel State Information.
CTS	Clear To Send.
CQI	Channel Quality Indicator.
DPS	Discriminatory Processor Shearing.
FCFS	First Come First Served.
HetNet	Heterogeneous networks.
h-PPP	Homogenous Poisson Point Process.
I-TBS	Index Transfer Block Size.
ICT	Information and Communications Technology.
LTE	Long Term Evolution.
MAC	Medium Access Control.

MCS Modulation and Code Scheme.

MIMO Multiple-Input Multiple-Output.

OFDM Orthogonal Frequency-Division Multiplexing.

pdf Probability Density Function

PHY Physical layer.

pmf Probability Mass Function

PPP Poisson Point Process.

PS Processor Sharing.

QoS Quality of Service.

QoS Quality of Experience.

RAT Radio Access Technology.

RB Resource Block.

RTS Request To Send.

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

SIR Signal-to-Interference Ratio.

SNR Signal-to-Noise Ratio.

TDMAI Time-Division Multiple Success.

TTI Transmission Time Interval.

UDP Ultra Dense Networks.

WiFi Wireless Fidelity.

W.l.o.g Without loss of generality.

Chapter 1

Introduction

Nowadays, a cellular device is so much more than a mean to talk or exchange some personal messages with someone who is away. A lot of the actions that a user was used to take in his personal computer, is now completing them through his personal cellular device (web browsing, reading news papers, hearing music, watch video etc). In addition, almost all of the data storage or transport equipment (floppy disk, CD, DVD, external hard-drive) had been replaced by the internet. Even the local memory of our devices has been partially replaced by the internet cloud. The unfortunate event of one day without internet access, reminds us how few things are stored in our personal devices. Therefore, except for the absolute increase of the number of cellular users, the average cellular user is becoming more and more demanding. Full internet coverage, video streaming, uploading and downloading high-quality pictures, transport of large data files, cloud computing are just some of the modern needs of the cellular user. Of course all of the aforementioned should happen with high rates and low delay.

The modern cellular networks have to deal with the heavy increase of data traffic were the evolution of wireless communications has lead to. More precise, studies show that data traffic increases exponentially and this trend is expected to continue for the foreseeable future [1]. If we take into account that the available resources of wireless telecommunications are limited, we conclude that the satisfaction of those high load demands by the network is not trivial and a lot of research is taking place in order to tackle this problem (even temporally).

One direction of research that aims to relieve the problem, is the exploration of new frequency bands. The goal is to find new suitable frequency bands and to develop radio systems that will be able to operate on those. This radio frequency exploitation and system design should take into account some important components i) The spectrum management should tackle effectively regulation issues in national, regional and global levels ii) The tradeoff between penetration depth and frequency reuse – low frequencies have lower propagation losses, so longer penetration depth and because of that smaller frequency reuse, besides, higher frequencies have high propagation losses, so short penetration depth and high frequency reuse– iii) In high frequencies the antennas size changes fundamentally so we will face compatibility issues iv) Around the resonant frequencies of oxygen O_2 or water H_2O the propagation of electromagnetic waves suffers by high losses that caused by scattering on the aforementioned molecules. Summarizing, even if the radio frequency range is almost 300GHz, is not an easy task to find frequency bands and to design the radio systems for those.

Another direction is to use multiple antenna elements in both transmitters and receivers.

MIMO (multi-input multi-output) and massive-MIMO systems aim to increase the achievable data rate of the network. MIMO systems focus on increasing network's data rate by the use of multiple antenna elements in three main ways i) *Precoding*, is the technique that uses multi-stream beam-forming in order to maximize the total throughput by transmitting simultaneously to multiple-antenna receiver systems ii) *Spatial multiplexing* where each antenna transmits a different data stream in the same frequency channel. The signals arrive at the receiver with sufficiently different spatial signatures and if the receiver has accurate channel state information (CSI), it can separate these streams into parallel channels iii) *Diversity coding*, a single stream is transmitted, but the signal is coded using techniques called space-time coding. Space-time coding relies on transmitting multiple, redundant copies of a data stream to the receiver in order to decode the one with the best channel conditions.

Another main concept in order to tackle the aforementioned data tsunami problem, is to benefit from the properties of the small-cells. The lower transmission power gives the capability to small-cell BS to operate in the same frequency band much closer than the macro-cells without increasing dramatically the interference between them. The development of denser networks in order to split the total amount of traffic in different BS that will reuse the same frequency bands is one of the most promising solutions, especially for urban areas, the futuristic approach of ultra dense networks (UDN) which becomes more and more attractive. Some proposals of UDN consider the development of the BS antennas on the top of traffic lights, or the use of the existing indoor WiFi network. It is worth mentioning that even in macro cells the trend is to decrease their cell-radius, which results in an increase of their density.

A modern cellular network consists of different BS, with different radio access technologies (RAT), protocols, transition powers or even operational frequency bands. This kind of networks are called *Heterogenous Wireless Networks* but for simplicity we will refer to them as Heterogenous Networks (HetNet). Each of the different radio access techniques named tiers of the network. An important goal of the research community is the tighter integration and collaboration between the tiers.

The gain of the tighter collaboration of the tiers could be many-fold. It is expected that modern HetNets will i) Improve *spectrum efficiency* by making use of RATs which may have few users through load balancing across tiers ii) *Coverage* might be improved because different RATs may fill holes in coverage that any one of the single tier networks alone would not be able to fill iii) *Reliability and handover* will be improved because when one particular RAT within the HetNets fails, it might still be possible to maintain a connection by falling back to another RAT.

Summarizing, our assessment is that in the near future networks will:

- operate to a *wider range* of the radio-electromagnetic bands.
- use *multiple antennas* in both transmitters - receivers, in order to increase networks throughput.
- be *denser* in order to increase the spatial-frequency utilization of the network.
- be *less regular planned* or even totally random. Buildings and other natural obstacles are not allowed to place the BSs antennas wherever we plan, especially for the small cells case.
- be *more heterogeneous* with cooperation of different RAT tiers.

In this thesis we are studying the HetNet scenario that combines a macro-cell and a small-cell. We assume that macro and small cells tiers operate in orthogonal frequencies, so we do not have inter-tier interference. The interference is caused only by the BSs of the same tier. Furthermore, we chose that the small cell operates on the unlicensed frequency bands.

Unlicensed Bands

Considering the extremely high cost of the spectrum, the Federal Communication Commission of US in spectrum auction 97 (began on 11/13/2014 and closed on 1/29/2015) sold 65MHz for the price of $44.9 billions [2], operators would be very satisfied if they could release some of their customers' traffic through the free unlicensed band. Considering the aforementioned, it makes perfect sense that HetNets that combine both licensed and unlicensed bands have attracted the attention of numerous leading telecom operators and vendors, such as Qualcomm [3–5], Ericsson [6], Alcatel-Lucent and Samsung [7], and Huawei [8].

There are already various protocols that aim to combine the LTE radio access technology with the unlicensed spectrum. A brief overview is presented Bellow.

- *LWIP* (LTE WLAN integration with IPSec tunnel), is a 3GPP Release 13 feature that enables WiFi to be more optimally integrated into an LTE Access network. The specification of LWIP aims through switching or aggregation parallel data paths in the IP level provides more efficient load balancing between LTE and Wi-Fi or capacity boost. This uses multi-IP to re-sequence and reassemble the parallel data streams at an anchor point in the core network.

- *LWA* (LTE Wi-Fi Aggregation), as LWIP is a technology defined by the 3GPP that the network utilizes both links simultaneously but runs the LTE Layer on top of the Wi-Fi Physical Layer. LWA allows the usage of both links for a single traffic flow and is generally more efficient due to the aggregation of LTE and WiFi at the level of PDCP, while keeping the MAC and the PHY layer of WiFi "as is".

- *LTE-U* (LTE in Unlicensed Spectrum) proposed and originally developed by Qualcomm based on 3GPP Release 12. A relatively unmodified form of LTE operates in the unlicensed spectrum. The protocol proposes dynamic channel selection to avoid Wi-Fi and adaptive duty cycle to fairly coexist. LTE-U does not operate the listen before talk (LBT) mechanism . Due to that, it is only applicable to markets that do not require LBT.

- *LAA* (Licensed-Assisted Access) proposes a modified version of LTE in the unlicensed band in order to boost the downlink speeds based on 3GPP Release 13. The main difference with LTE-U is that LAA in addition to dynamic channel selection (to find the least used sub-band) also employs Listen-Before-Talk mechanism to co-exist with other unlicensed users.

- *eLAA* (Licensed-Assisted Access) is an evolution of LAA proposed in 3GPP Release 14 that also allows the use of unlicensed spectrum for the uplink as well.

- Finally MulteFire uses LAA and eLAA to run an LTE network entirely in the unlicensed spectrum without requiring an "anchor" in licensed spectrum. The goal of MulteFire is the LTE network to operate with Wi-Fi-like deployment simplicity, but with the higher performance taking advantage of the LTE PHY layer capabilities.

In this thesis we examine one of the most promising scenario for such HetNets, the combination of LTE macro (or small) cells with WiFi small cells. We focused on the scenarios that both RATs operates without modifying the actual access technology (LWIP and LWA). Our goal is to mathematically model and analyze a network with the aforementioned characteristics (dense, randomly placed, heterogeneous) in order to provide insights about its performance.

System performance, especially in such complex systems as the telecomm networks is not well defined. There are different performance metrics in every layer of the communication system such as capacity, coverage, frequency utilization, frame error rate, coexistence with other networks or devices, power consumption, latency and security are just some of the performance metrics that are used from electrical engineers in different layers of the telecom systems.

We aim to analyze and provide insights about metrics that describe the flow-level performance of the network, because we believe that dynamic flow-level metrics represent more accurately the actual users' experience. Furthermore, a chapter of this thesis aims to combine the flow-level metrics such as flow delay or the congestion probability with the the energy consumption of the network.

Users Quality of Experience

When a network is designed and developed the goal is to serve and satisfy its users, in other words, to make our users happy (in order to joyfully spend their money on network operators and mobile-phone infrastructures, which as a consequence helps PhD students and senior researchers to find some funding for research!). It turns out that quality of experience QoE (user's satisfaction) can not always be captured by the traditional SINR (channel quality or capacity) metric, e.g., "I have five bars, why is my download so slow!!!" [9]. A better metric might be delay, as one of the key goals of 5G technologies is to minimize users' delay [10, 11].

As we will see later user's delay differs qualitatively from throughput especially while the network traffic is increasing. When a system is under-utilized, indeed the delay is strongly depended on throughput and probably this is the reason why a lot of electrical engineers in wireless systems consider that QoE has same attributes as throughput or SINR, while they under-estimate the effect of load on it.

Civil traffic engineers tackle this kind of problems from the middle of 20th century [12]. If we do the analogy between SINR and road distance, a motivation example about how the traditional metrics sometimes fails to capture the delay is the difference between shortest and fastest route (Fig. 1.1). Of course delay depends on SINR, but not only. The existence of a short road is a necessary but not sufficient condition in order to arrive fast at your destination. So, like in other scientific fields, the electrical engineers should take into consideration more parameters than the SINR when they design a wireless network that aims to minimize the users delay. As we shall see, even the throughput depends on other dynamic characteristics of the network.

We are interested in the flow-level performance of the network from both users and operators perspective. The main flow-level metrics that we are interested in are: a) *Delay*, what will be the mean user's delay until a file is downloaded? b) *Congestion probability*, what is the probability of a BS to be congested and and not being able to serve more users? c) *Load*, what is the load of the network? d) *Utilization*, what is the average utilization of a given network?

Until this point, we mentioned that dense heterogeneous networks are one of the most promising solution in order to tackle the problem of the exponentially increased data demand. Additionally, we referred to the need that the design of those networks should take into account

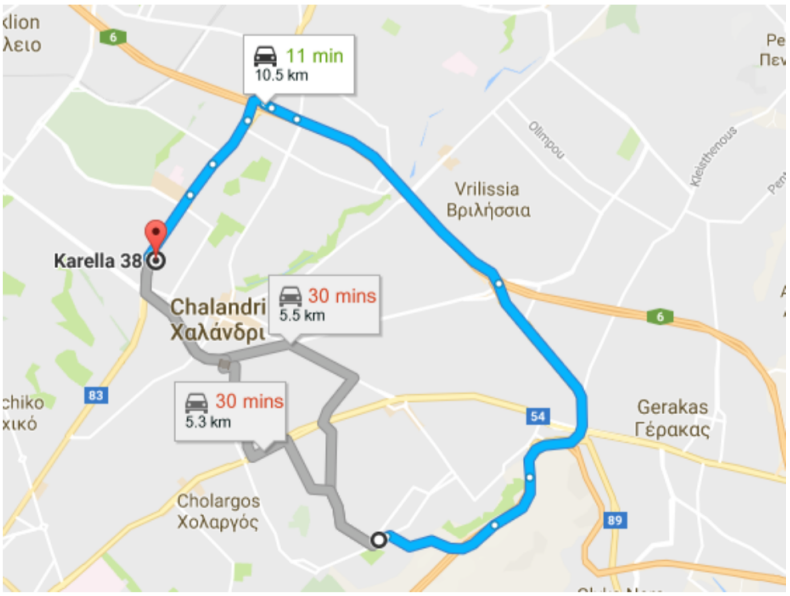


Figure 1.1 – Sortest Vs Fastest route

flow-level metrics in order to satisfy the users QoE. What we totally missed to mention is if those type of networks are feasible regarding the energy consumption. Undoubtedly, the energy consumption is a critical component for the sustainability of a network and would be an omission if not be included in this thesis.

Energy Efficiency

Unfortunately, "there is no such thing as a free lunch". This law of nature holds for wireless networks as well. So, on the one hand the trend of future networks is to become more and more dense and on the other hand it is essential for the ICT sector to contribute to the reduction of its carbon footprint; this purpose includes recycling (in any form), using green-power schemes and reducing power consumption. Indeed, the cost benefits of this procedure could be important, but still, the main target should be environmental protection.

Our interest is to theoretically model the energy consumption of a network in order to provide insights about the dependencies between energy and other performance metrics. A lot of research effort is invested to analyze the trade-off between the users' quality of experience and the energy consumption of a network, and how this relationship scales with respect to network's density. There are two main case studies 1) we would like to know the gain in terms of users' QoE and the loss in terms of energy consumption as we add more and more BS to our network (densification) and 2) if we already have a dense network, what is the energy gain if we turn off a part of the network and what is the loss in terms of QoE of this process.

The performance analysis becomes much more interesting if we take into account the actual energy performance of a BS. In [13] and [14] the authors analyze the energy consumption of BS. Energy consumption could be divided in two main components 1) the transmission cost (digital signal processing, power amplifiers, etc.) and 2) the energy spent in order to have the BS active even if it is not transmitting any signal (power supply losses, cooling, etc.). It turns out that

there is not a dominant component that allows you not to take into consideration the second one.

1.1 Contributions

Our motivation was to create an analytical framework, in order to capture the flow-level performance of randomly placed networks without assuming that both users and BS are saturated. On the one hand we seek to avoid the complex and time-consuming packet-level simulators but on the other hand we desire our framework to be as close as possible to the real world. We tried to keep the balance between mathematical elegance and reality, but when this was not possible we preferred to make our model "uglier" than makeing a total unrealistic assumption.

In order to be consistent with the real world we put effort into understanding and modeling the rate of the systems that we decide to work with, LTE and WIFI. This modeling procedure provides us with *modulation and coding schemes* (MCS's) rates and thresholds in order to avoid the Shannon's formula.

After the proper modeling of the PHY layer, we study the network's flow-level performance. Our analysis is based on the combination of two key theoretical tools that have recently provided us many useful insights in wireless systems: (i) We use *queueing theory* to model the performance of dynamic flow arrival and service via the respective scheduler, at the level of a single BS; (ii) We utilize *stochastic geometry*, in order to understand the impact of topological randomness and interaction/competition between BSs at the network level, in order to derive statistics about the *number of users associated with a base station* at each tier, and the MCSs offered at each BS. Both these quantities serve as key inputs to the BS queueing model: the former to define the total traffic intensity (in terms of flow arrivals) a given BS has to serve, the latter to define the average service rate (in terms of flow departures) that a BS is able to offer.

Finally, based on our framework we want to investigate how the performance of the network scales with respect to it's parameters and to study the trade off between energy consumption and users' QoE. Additionally, based on our framework we want to investigate the performance gain for different tier-association criteria in a HetNet environment and more specifically the 2-tiers case of LTE macro and WiFi small cells.

1.2 Outline

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

Chapter 2 - Background

In this chapter we present a brief introduction to our two basic mathematical tools that are 1) stochastic geometry and 2) queueing theory, focusing on their applications of wireless networks. The goal is to make this thesis self-consistent and to help the readers that are not aware of those tools to read the rest of the chapters uninterrupted.

Afterwards, we present the related work that applies those tools in order to solve problems of wireless (or not) networks. We definitely present the few existing works that are on the intersection between flow-level dynamics and random topologies. Finally we present some works

that do not belong to any of the aforementioned categories but provide us with very useful ideas and insights according to our problems.

Chapter 3 - PHY and MAC layer Modeling of LTE and WiFi RATs

We consider LTE and WiFi networks in order to model both PHY and MAC layers. Our physical layer abstraction consists of a mapping between users' SINR and their corresponding rate. Furthermore, we propose a proper queue model to capture the MAC performance of schedulers of each radio access technique that we are interested in.

Regarding the PHY layer abstraction of LTE, we study the mapping between resources blocks and data rate. Obviously, this procedure depends on channel conditions and the sensitivity level of the receiver. The amount of bits that are transmitted in a resource block is defined through the MCS. The supported MCS are defined at the corresponding protocol description documents [15], but the sensitivity thresholds for each mode are not defined since they heavily depend on the receiver implementation characteristics. In this thesis as reference receivers we used the OpenAirInterface platform in order to obtain a more realistic mapping between SINR thresholds and the corresponding data rate.

According the MAC layer modeling of the LTE BS in the downlink, we assume that the LTE scheduler divides the total amount of resource blocks equally among active users (resource fair scheduler) and serves them simultaneously. Under this assumption we model the LTE scheduler as multi-class M/G/1/PS queue.

In WIFI systems the data rate of each MCS mode is defined at the corresponding protocol description documents [16], but the sensitivity thresholds for each mode depends on each implementation (as LTE case). Additionally, the effective data rate of a WIFI system is less, because we should take into account the overheads of the system like, collisions, RTS and CTS packages etc.

Finally, the WIFI scheduler at each time quantum allocates all frequency - time resources on a single user. The amount of time that each user occupies the medium depends on his data rate. The amount of time that each user occupies the medium multiplied with his data rate will be equal for every user. Thus, we proposed a fair throughput queueing system that captures the performance of the WIFI scheduler.

The work in this chapter corresponds to the following publications:

- *G. Arvanitakis and F. Kaltenberger, "PHY and MAC layer modeling of LTE and WiFi RATs," Research Report RR-16-317, March 24th, 2016.*
- *G. Arvanitakis and F. Kaltenberger, "Stochastic Analysis of Two-Tier HetNets Employing LTE and WIFI," EUCNC 2016, 25th European Conference on Networks and Communications, June 27-30, 2016, Athens, Greece.*

Chapter 4 - Dynamic Performance of Single tier Network

In this chapter, we develop a flexible and accurate analytical model of large networks with random base station (BS) placement and randomly placed users as well. We want to understand the impact of key network parameters like BS density and load on the network performance. The main goal is to understand the flow level dynamics of such a system, assuming non-saturated users and studying the congestion statistics for BSs and the per flow delay. To achieve this,

we base our analysis on two main tools: (a) stochastic geometry, to understand the impact of topological randomness and coverage maps and (b) queueing theory, to model the competition between concurrent flows within the same BS.

In order to derive an analytical framework that captures the flow level performance of a network one key aspect is to compute the interference. Interference has a significant role for the choice of MCS mode and therefore the corresponding rate. The most common way in literature to calculate the interference is to assume that all BS are saturated, so they cause interference with each other all the time. We devoted a lot of effort in order to revoke this assumption and to consider the more realistic scenario where the BSs are contributing to the interference only for the amount of time that serve users. Furthermore, in our model we are considering both interfering cases (always ON and load based).

Finally, we applied our model on popular RATs, such as LTE and WiFi. Our results provide some interesting qualitative and quantitative insights about the performance of those networks. Additionally, we saw that the performance predictions if we assume saturated BS or the load base case differs dramatically.

The work in this chapter corresponds to the following publication:

- *G. Arvanitakis,T. Spyropoulos and F. Kaltenberger, "An Analytical Model for Flow-Level Performance of Large, Randomly Placed Small Cell Networks," GLOBECOM 2016, IEEE Global Communications Conference, 4-8 December 2016, Washington, USA.*

Chapter 5 - Study the Energy Efficiency and User's QoE Tradeoff

Energy consumption is one of the primary concerns of modern dense small cell networks. On the other hand network density is one of the key parameters that affect the users' QoE. How many BS we should add on the network with high traffic in order to achieve a specific QoE. After the deployment of a dense network, one of the key concepts to improve its energy efficiency is to turn off a part of the base stations when they are idle or only lightly loaded, since even then a considerable amount of energy is consumed.

In this chapter we assume a linear energy cost model with respect of transmitting time for each BS, the constant term is the energy that the BS is consuming in order to be ON and the linear term depicts the amount of energy that consuming while the BS is transmitting some information. We analytically investigate the tradeoff between energy efficiency and user experience. The user experience is measured in terms of delay assuming a non-saturated traffic model.

We saw that the behavior of energy efficiency vs delay tradeoff is not straightforward and depends on the ratio between the constant and linear term of the energy cost model. Additionally, as the density of the network is increases, the gap between performance if we pick the base stations that we will turn off randomly or we pick them with more sophisticated criteria is becoming negligible. Furthermore, we are interested in how the bits per joule and the flows per joule metrics are scaling with respect to network's density.

We considered the scenario, where the density of users in a given aria is changing dramatically inside the day in order to define what should be the adaptation of the BS density in order to save energy but without affecting the performance of the remaining users.

We provide an overall network performance analysis with respect to the BS density. We apply our model to the popular LTE radio access technology but it can easily be extended to

others. Our results provide interesting quantitative insights about the capabilities of energy improvement in relation to the user's quality of experience.

The work in this chapter corresponds to the following publication:

- *G. Arvanitakis and F. Kaltenberger, "Energy vs QoE Tradeoff of Dense Mobile Networks," ICNC 2018, International Conference on Computing, Networking and Communications, March 5-8, 2018, Maui, Hawai, USA.*

Chapter 6 - An Analytical Model for Flow-level Performance in Heterogeneous Wireless Networks

Modern cellular networks are increasingly heterogeneous, as a result of operators' efforts to deal with an unprecedented data crunch. This increased complexity however makes performance analysis challenging. In this chapter, we develop a flexible and accurate model in order to analyze the performance of large heterogeneous cellular networks (HetNets), and understand the impact of key network parameters.

This model consists of K tiers of randomly located BSs, with different densities, transmit powers and RATs. Our main goal is to understand the impact of flow level dynamics on such a system, assuming non-saturated users that randomly generate download requests. We do so by deriving analytically the per flow delay achieved by such a network, the average load of the BS across space and time, as well as the congestion probability of BSs in different tiers (i.e., the percentage of BSs that will be overloaded).

We apply our model to the case of a popular 2-tier HetNet, based on LTE and WiFi, in order to further understand the performance differences of popular user association criteria, such as Off-load, Max-SINR association, and Min-Delay association. Our results shown that if we take into account the load of each tier in tier-association we have a significant improvement of the overall network performance especially when the traffic is high.

The work in this chapter corresponds to the following publication:

- *G. Arvanitakis,T. Spyropoulos and F. Kaltenberger, "An Analytical Model for Flow-level Performance in Heterogeneous Wireless Networks," **to appear** on IEEE Transactions on Wireless Communications.*

Chapter 7 - Conclusions and Future Research

Finally we present the conclusions of our study and the next steps of our work.

Appendix - Distribution of the Number of Poisson Points in Poisson Voronoi Tessellation

In this appendix we deal with the following problem: Let two independent sets Φ_1 and Φ_2 follow homogeneous PPP having different densities in two dimensional space. Assuming that Voronoi Tessellations are generated with respect to Φ_1. This appendix calculate analytical the probability distribution of Φ_2 cardinality in an arbitrary tessellation of Φ_1. In other words, consider a single tier of BSs distributed in 2D as a homogeneous PPP with density λ_{BS}, and offering coverage to a set of users distributed as another PPP with density λ_{u}. Assume further that user association within this tier is done using the closest-distance rule, what is the probability of having exactly n users in an arbitrary cell.

The work in this chapter corresponds to the following publication:

- *G. Arvanitakis, "Distribution of the Number of Poisson Points in Poisson Voronoi Tessellation,"Research Report RR-15-304.*

Chapter 2

Background

This chapter focuses on briefly introducing the reader to the main mathematical tools that we use in this thesis and to present the related works in the area of wireless communications. The scope of this chapter is not (and it could not be) to cover and present a summary of all problems and achievements of those really fascinating topics such as stochastic geometry and queueing theory, but to make this thesis as self-consistent as possible, and to help the readers that are not familiar with those topics, to study our work. That being said, we introduce the reader shortly to stochastic geometry and queueing theory, focusing in modeling of wireless communications and sometimes even more narrow to what we are going to use in the rest of the manuscript.

2.1 Stochastic Geometry

Stochastic geometry (former known as geometric probability) studies complicated geometrical patterns in order to provide them with suitable mathematical models and useful statistical tools. Initially, stochastic geometry considered problems of finite number of randomly placed objects. The modern theory initiated by D. G. Kendall, K. Krickeberg and R. E. Miles considers more complex distributions and usually assumes that the pattern is spread in the infinite plane.

Stochastic Geometry aims to describe and to model mathematically a random collection of points, in one, two, three or higher dimensions. Additionally, it intends to study the statistical properties of the aforementioned model and to derive statistical averages over all the possible realizations of such a random collection.

In this thesis, we use the point processes to model the distribution of BSs or users location in the network. This assumption offers us the capability to characterize the performance of a network without assuming a specific realization and to derive more general conclusions.

Definitions

Point process is a random collection of points that lies in a given space. The most common measure space is the d dimensional Euclidian space, $\mathbb{R}^d$.

The set $\Phi = \{x_1, x_2, ...\}$ denotes a point process. Each of the random variables x_i represents an element of this point process, $x_i \in \mathbb{R}^d$.

Voronoi region $V(x)$ of a point x of a general point process Φ, characterizes the area where the distance from x is not larger than the distance to any other point in Φ, Fig. 2.1. Mathematically, we can express this relation by

$$V(x) \triangleq \{y \in \mathbb{R}^d : ||x - y|| \leq ||y - \Phi||\} \ . \tag{2.1}$$

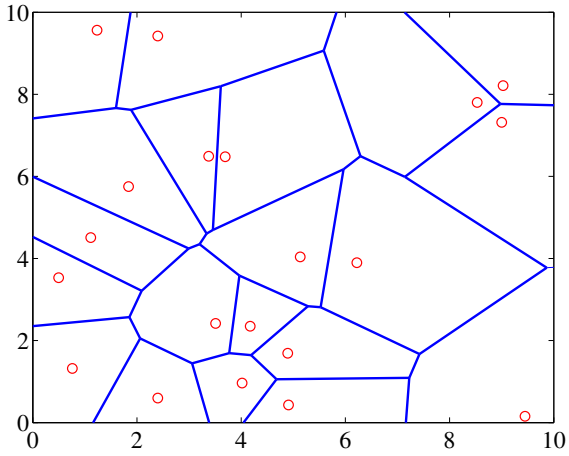


Figure 2.1 – The red cycles represent the points and the blue lines the corresponding Voronoi regions

For a given point process, the *number of points* that falling in the region $B \subset \mathbb{R}^d$ is denoted as $N(B)$. The number of points $N(B)$ is a non negative random variable, $N(\cdot)$ which is also called counting measure.

At this point, we should define two very important metrics that we use in the present thesis:

i) *Coverage probability:* assuming a given point process, we want to calculate the probability on the origin the receiving signal form the closest point divided by the signal of the rest of the points to be larger than an SINR threshold τ.

$$P(SINR > \tau) \ . \tag{2.2}$$

ii) *Congested probability:* assuming a given point process Φ_u, the congestion probability of a given subspace B is the probability that more than M points lie in B

$$P(N(B) > M) \ . \tag{2.3}$$

Is obvious that as subspace B we assume the Voronoi region of a BS and as point process Φ_u we model the distribution of users. So congestion probability is the probability the amount of connected users in a BS to be larger than the maximum amount that the BS can handle.

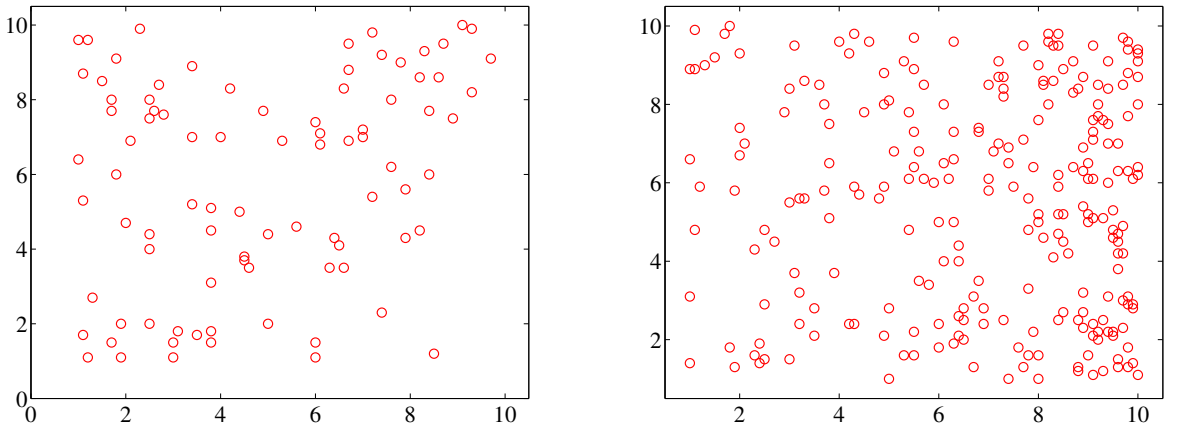
Basic Point Processes

Binomial point process is the simplest example of a point process, which contains only one point uniformly distributed in space. We can expand this point process by adding n independent uniformly distributed points, the more complex pattern named *binomial point process of n points*.

Homogenous (or uniform) Poisson Point Process (h-PPP) with density λ, is a point process in $\mathbb{R}^d$ where the amount of points $N(B)$ of the area B, has Poisson distribution with mean $\lambda|B|$. So, the number of points in the area B is given by

$$P(\Phi(B) = k) = \frac{(\lambda|B|)^k}{k!} \exp(-\lambda|B|) \tag{2.4}$$

Homogenous means that after we defined the total amount of points in the area B according to Eq. (2.4), the points distributed homogeneously in area B, Fig 2.2 *a*. One of the most powerful properties of homogenous Poisson point process is that if $B_1, B_2, ...$ are disjoint bounded sets, then $N(B_1), N(B_2), ...$ are independent random variables. We should mention that we are interested about the two-dimension h-PPP in order to model the location of the BS and users in the $x-y$ plane. We can assume that the point density λ is not a constant but differs regarding the spatial coordinations $\lambda(x, y)$. In this case named *inhomogeneous Poison point process*, Fig. 2.2 *b*.



a: Homogeneous

b: Inhomogeneous, λ is linear with respect to the x-axis

Figure 2.2 – Poisson Point Process

A very useful property of h-PPP is the distance to the nearest-neighbor point. Assuming a h-PPP Φ with density λ on $\mathbb{R}^d$. The radius R of the largest, d-dimensional sphere that fits before hits any of the points of Φ has distribution according to

$$P(R < \upsilon) = 1 - e^{-\lambda\upsilon} \ . \tag{2.5}$$

Generally, we assume that each user is associated with the closest BS, so this result provides us with analytical formula about the distance distribution between the user and the nearest BS in a h-PPP environment.

Hard-core process is a point process which guarantees that the points can not be closer than a minimum distance δ. We can start from a h-PPP with no restrictions and then remove the points that violate the minimum condition. In general, hard-core process is less clustered than the ordinary h-PPP, Fig. 2.3.

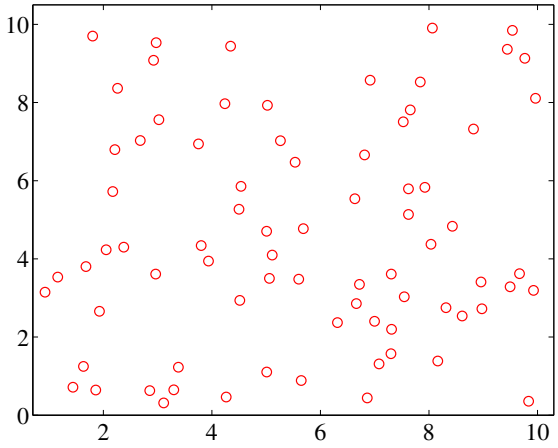


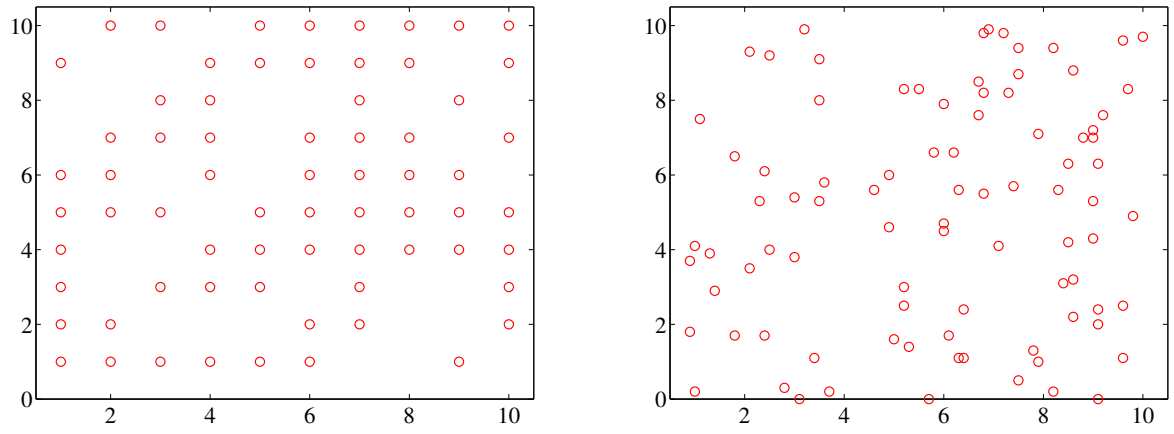
Figure 2.3 – Hardcore point process

Starting with a basic h-PPP Φ_1 with intensity λ_1 in the 2-dimensional space, and by removing all points that have a neighbor within distance δ. The final density resulted by the above process is given by

$$\lambda = \lambda_1 \exp(-\lambda_1 \pi \delta^2) \, . \tag{2.6}$$

Hard-core process is more accurate than h-PPP in order to model the location of the BSs, because in h-PPP there is the probability that two BSs to end up asymptotically close, something that is not inline with the reality where the BSs even if they are totally random, they will not placed one up to the other. Unfortunately, hard-core process is not so analytically tractable for our purposes.

Bernoulli lattice is the point process where starting from a square lattice with a given spacing $\eta > 0$, each point of the lattice is retained independently with probability p. Interestingly, this point process, as we increase the initial density of the square lattice and decrease the retained probability p, the distribution tends to h-PPP with density $\lambda = p\eta^{-d}$, Fig. 2.4.



a: $\eta = 1$, $p = 0.7$ b: $\eta = 0.1$, $p = 0.01$

Figure 2.4 – Bernoulli lattice process

2.2 Queueing Theory

The first paper on what we call now queueing theory was published from Agner Krarup Erlang only in 1909. Nowadays, queueing theory has provided solutions in almost every field of science, physics, computer science, electrical engineering, biology, civil traffic engineers etc. For compactness and completeness we present in this section the basic concept of queueing theory problems as well as the usual notation avoiding the mathematical definitions and derivations. Our goal is to help the non-versed reader to follow the rest of our work without having to indulge into queueing theory, but we strongly recommend him [17]. Additionally, we will "thin" the total amount of information considering that we apply queueing theory to wireless telecommunications.

Introductory Example

Let us build an introductory and motivation example, that we will solve at the end of this section in order to show the importance of queueing theory in the analysis of wireless systems.

We assume an indoor femto cell with four associated users. All users are scrolling to a social media website and looking at photos of other people, each photo have size $s = 4$ MB. Each user takes a quick glance of the photo and immediately demands the next one. The time until the next demand is random but with mean value $1/\lambda_f = 10$ sec. The user a is standing near the femto cell and operates with the best MCS and with rate $r_a = 50$ MB/s, users b and c are in the near room and operate with rate $r_b = r_c = 20$ MB/s and the final user is on the garden and operates with $r_d = 1$ MB/s.

i) What is the load (utilization) of the femto cell?

ii) What is the average delay of the users in this system?

iii) What will happen to the load (utilization) and what to the delay if the file size of each photo increased by $\times 1.5$?

iv) What will happen to the load (utilization) and what to the delay if we "cut" the edge user?

v) How many middle range users with rate $r_M = 20$ MB/s this system can afford.

Those are just some of the very interesting questions that queueing theory can provide analytical solutions. Unfortunately this is not always the case.

Definitions and notation

Literally, queueing theory is the scientific / mathematical subject that studies queues and belongs to the much broader area of mathematics called stochastic modeling and analysis. Queueing theorists aim to model the problems in order to predict system's performance such as queue length, mean delay or delay variability or system load etc.

The main entities that characterize a queue are 1) the time between the arrivals 2) the size of jobs 3) the number of servers and 4) the queueing discipline, how queue is allocating its resources to the queued jobs (scheduling policy).

A widely used notation in order to characterize a single queue system is Kendall's notation with form $A/S/c$ where A describes the time between the arrivals, S indicates the jobs' size and c the number of servers at the node. The Kendall's notation extended in $A/S/c/K/N/D$ where K is the maximum number of customers allowed in the system including those in service (assuming k the buffer space, $K = c + k$), N is the size of the population from which the jobs come, considering that users' request never ends in wireless communications $N = \infty$, finally, letter D indicates the queueing discipline (scheduling policy).

The following Table 2.1 indicates the most common symbols that describe the arrival process *(A)*

Table 2.1 – Symbols regarding the arrival process

Symbol	**Description**
M	Markovian or memoryless, Poisson process (or random) arrival process
M^X	Poisson process with a random variable X for the number of arrivals at one time
D	Deterministic
G	Generic distribution

Table 2.2 presents the most common symbols for distribution of time of the service of a customer (S)

Table 2.2 – Symbols regarding the service time distribution

Symbol	**Description**
M	Markovian or memoryless, exponential service time
M^Y	Exponential service time with a random variable Y for the number of arrivals at one time
D	Deterministic service time
G	Usually refers to independent service time

Regarding the scheduling policies (D), Table 2.3 presents the most common of them

Table 2.3 – Symbols regarding the scheduling policies

Symbol	**Description**
FCFS	First Come First Served
LCFS	Last Come First Served
SIRO	Service In Random Order
SJF	Shortest job first
PNPN	Priority service, including preemptive and non-preemptive
PS	Processor Sharing, the jobs are served in parallel

Regarding the priority queues there are two main categories, non-preemptive where a job in service cannot be interrupted and preemptive where a job in service can be interrupted by a higher-priority job. Additionally, we should note that the term *priority service queue* includes a variety of different systems because the way that the system prioritize each job (or class of jobs) leads to a totally different queue with totally different performance. The same comment holds for the *processor sharing queues* as well. Processor sharing indicates that the jobs inside the queue are served on parallel, but this does not describe how resources are allocated between the jobs. As we will see later on, in this thesis we are interested about two special cases of PS queues, the *resource fair* where the queue equally splits all of its resources to the queued jobs and the *throughput fair* where the queue allocates the resources in order all jobs to achieve equally throughput, e.g. if we have two jobs where the rate of the first one is the half of second's $r_1 = r_2/2$ then the system will give double resources (time or frequency) to the first one.

We could further categorize the queues regarding the Customer's behavior: *Balking* where customers decide not to join the queue if it is too long or *Jockeying* where customers switch between queues if get served faster by doing so or *Reneging* where customers leave the queue if they have been waiting too long for service. In this thesis we are not examining any of the aforementioned special categories, once the user / flow get associated with a specific BS / queue will stay until they get served.

Generally it is possible to design a stochastic model that captures the relationship between the incoming jobs and the system. Unfortunately, these stochastic models do not always provide analytically tractable solutions about performance. We should mention that most theorems and analytical results in queueing theory proved by reducing queues to mathematical systems known as Markov chains. In order a Markov chain to be analytically solvable, our system should be memoryless and the next state of the system should be dependent only on the current state and not on the past. We should be very careful, because the memoryless property could greatly simplify our analysis but on the other hand could lead to very inaccurate performance results and poor system design if this assumption does not really holds.

Base Station of Wireless Communications

Arrival Rate

In respect of the wireless communications, the memoryless property is a reasonable assumption. We can assume that a number N of users lie in the coverage area of a BS and each user generates a demand / flow according to poisson point process with density λ_f, means that the time between two flows is random with mean value $1/\lambda_f$.

Due to poisson merge property, the incoming flows in the queue, is again a poisson point process. Thus, the arrival rate of this system is $\lambda = N \cdot \lambda_f$ with units flows/sec.

Service Rate

In wireless telecommunications each user according to his channel conditions operates with a specific MCS. This means that the flows of the system do not get served with the same rate. Assuming that all flows have equal size s bytes but transmitted with different instantaneous rate r according to a probability distribution $f_R(r)$, the mean service rate of the system is given by

$$\langle \mu \rangle = \left(\sum_r \frac{f_R(r) \cdot \langle s \rangle}{r} \right)^{-1} \quad \text{flows/sec} \, . \tag{2.7}$$

Thus, the service rate is given by the harmonic mean and not by the statistical mean of the flow rates. In order to provide some insight about this result we should consider two flows of equal size 4 MB, the first one gets served with 4 MB/s and the second one with 2 MB/s. The statistical mean predicts that the average rate of the system will be $\frac{2+4}{2} = 3$ MB/s. But this is not quite correct, if we assume that we are measuring the output of the queue, we will measure output of 4 MB/s for one second (the time that the first flow needs until to get served), and 2 MB/s for 2 sec (since the second flow has worse rate, it stays on the system), so the average services rate for the total 3 seconds that the system was operating is $\frac{1}{3}4 + \frac{2}{3}2 = 2.666$. So, we can conclude that the service rate tends to the worst performance.

Mean System's Load

BS's load can be defined as the ratio between the input job rate and the service job rate or formally

$$\rho = \frac{\lambda}{\langle \mu \rangle} \, , \tag{2.8}$$

the system can be characterized as stable when $\rho < 1$. If $\rho > 1$ the system in not stable and the amount of queued job tends to infinity as well as the average flow delay. In order to be inlined with what we present in the stochastic geometry section, when $\rho > 1$ we will refer to the system as *congested*. As we can easily see, system's load is dimensionless entity.

Mean User's Delay

In order to define average users delay, in addition to the arrival rate and the service rate we need to know the scheduling policy as well. Systems with the same characteristics perform totally

different in terms of average user's delay, depending on their scheduling policy. If we assume an ordinary first come first served queue with load ρ, Poisson arrivals of rate λ and S be a random variable that denotes the service time of each arrival, the average user (or flow) delay can be calculated according to

$$E[T]_{FCFS} = E[S] + \frac{\lambda E[S^2]}{2(1-\rho)} \,. \tag{2.9}$$

Where $E[\cdot]$ denotes the mean value and $E[\,\cdot^2]$ the variance. As we can see the average delay depends on the variance of flows' service time. As the variance is increasing, the delay increases as well, even if the load of the system remains the same.

For a system with exactly the same characteristics, but with processor sharing - resource fair - scheduler, the average user delay is given by

$$E[T]_{PS} = \frac{1}{\mu - \lambda} \,. \tag{2.10}$$

As we can see, the delay performance of the resource fair processor sharing is not affected by the variance of flows' service time distribution. Therefore, different scheduling policies give different delay performance and they have different properties. Unfortunately, not all scheduling policies could be expressed by elegant closed form mathematical formulas. Queueing theorists are studying different type of queues (and schedulers) in order to provide closed or semi-closed or asymptotic or lower / higher bounds about their delay performance, this procedure can be considered far from trivial.

Introductory Example (Solution)

We rewrite here our introductory example. We assume an indoor femto cell with four associated users. All users are scrolling to social media website and looking at photos of other people, each photo have size $s = 4$ MB. Each user takes a quick glance of the photo and immediately demands the next one. The time until the next demand is random but with mean value $1/\lambda_f = 10$ sec. The user a is standing near the femto cell and operates with the best MCS and with rate $r_a = 50$ MB/s, users b and c are in the near room and operate with rates $r_b = r_c = 20$ MB/s and the final user is on the garden and operates with $r_d = 1$ MB/s.

i) What is the load (utilization) of the femto cell?

ii) What is the average delay of the users in this system?

iii) What will happen to the load (utilization) and what to the delay if the file size of each photo increased by $\times 1.5$?

iv) What will happen to the load (utilization) and what to the delay if we "cut" the edge user?

v) How many middle range users with rate $r_M = 20$ MB/s this system can afford?

Solution

We should notice that the problem is not complete because the scheduling policy of the femto cell is not specified. In order to show that the scheduling policy affects the flow-level performance

of the system, even if the total resources are the same, we will solve the aforementioned questions for both FCFS and resource fair PS scheduler policies.

i) In order to decide about the load of the system, we should first define the arrival and the service rate of each system.

In both cases we have 4 equal users so the total arrival rate is

$$\lambda_{FCFS} = \lambda_{PS} = 4 \cdot \lambda_f = 0.4 \text{ flows / sec .} \tag{2.11}$$

respectively the service rate of both systems is

$$\begin{aligned} \langle \mu \rangle_{\text{FCFS}} = \langle \mu \rangle_{\text{PS}} &= \left(\sum_r \frac{f_R(r) \cdot s}{r} \right)^{-1} \\ &\approx 0.89 \text{ flows / sec .} \end{aligned}$$

So, in both cases the utilization of the system is

$$\rho_{PS} = \rho_{FCFS} \approx 0.45 . \tag{2.12}$$

ii) For the case of FCFS the average users' delay calculated according to

$$\begin{aligned} E[T]_{FCFS} &= E[S] + \frac{\lambda E[S^2]}{2(1-\rho)} \\ &\approx 2.58 \text{ sec .} \end{aligned}$$

Regarding the case of PS the average users' delay is given by

$$\begin{aligned} E[T]_{PS} &= \frac{1}{\mu - \lambda} \\ &\approx 2.03 \text{ sec .} \end{aligned}$$

Hence, we observe that even if the PHY characteristics are exactly the same, processor sharing scheduler operates roughly 20% better than FCFS in this example.

iii) If the size of the photos increase by $\times 1.5$ the load of both systems will scale linearly

$$\begin{aligned} \rho_{PS}^{s \times 1.5} = \rho_{FCFS}^{s \times 1.5} &\approx 1.5 \times 0.45 \\ &\approx 0.67 . \end{aligned}$$

The same does *not* hold for the case of delay. The delay will increase up to 3 times!

$$\begin{aligned} E[T]_{FCFS}^{s \times 1.5} &\approx 7.2 \text{ sec} \\ E[T]_{PS}^{s \times 1.5} &\approx 5.1 \text{ sec .} \end{aligned}$$

Additionally, it is worth mentioning that the processor sharing now operates 40% better than the FCFS system.

iv) If we "cut" the edge user, we decrease the total incoming load by 1/4 but if we re-calculate the

"new" service rate μ we will notice that the load of the system and the delay improved extremely non linear.

$$\rho_{PS}^{edge} = \rho_{FCFS}^{edge} \approx 0.048$$

$$E[T]_{FCFS}^{edge} \approx E[T]_{PS}^{edge} \approx 0.16 \text{ sec} .$$

That being so, the load decreased almost 10 times and the corresponding delay up to 16 times! Additionally, it is worth noting that in such a low load cases the delay of both schedulers is almost the same.

v) The final question is about the amount of middle rate users that this system can afford. We remind that a system is stable when $\rho < 1$ so assuming that the final amount of users in the system will be $k + 2$ (one high speed user with rate r_H, k middle rate users with rate r_M and one edge user with rate r_L)

$$\begin{aligned}
\mu &> \lambda \\
\left(\sum_r \frac{f_R(r) \cdot s}{r}\right)^{-1} &> \lambda_f \cdot (k+2) \\
\left(\frac{s}{(k+2)r_H} + \frac{s \cdot k}{(k+2)r_M} + \frac{s}{(k+2)r_L}\right)^{-1} &> \lambda_f \cdot (k+2) \\
k &< r_M(1/(s \cdot \lambda_f) - 1/r_H - 1/r_L) \\
k &< 29.6 \text{ !!!}
\end{aligned}$$

So the system can afford 29 middle rage users, so in total 31 users! The number is impressive if we take into account that the system initially was half utilized by serving just 4 users (we don't claim that those users will be happy with system load 0.988 and average delay 26 sec... but is still impressive...).

After this example that aimed to motivate the reader about the value of analyzing a wireless system as a queueing system, we move on to a brief presentation of the literature that assisted and motivated us to complete our work.

2.3 Related Work

2.3.1 Stochastic Geometry

To study the performance of a network assuming that BS are distributed according to homogeneous Poisson point process instead of hexagonal or square grid has been considered from the last decade of the 20th century [18, 19]. the distribution of the interference and the coverage probability studied in [20,21]. The most tractable advantage of this approach is that, it does not only avoid the problem of ideal and simplistic hexagonal or linear topologies but also provides closed form expressions.

The aforementioned framework has been widely used for studies of large and heterogeneous networks. Such examples, [22] expand [20] for the case of K-Tier downlink of heterogeneous cellular networks, [23] analyzes the performance of heterogeneous cellular networks taking into account the capabilities of Carrier Aggregation and [24] tackles the problem of offloading. Additionally, [25] models the downlink coverage probability in MIMO HetNets, [26] studies the problem of fractional frequency reuse for heterogeneous cellular networks, [27] follows the same stochastic approach in order to provide the network performance while turning off BSs and [28] further considers the backhaul network. In [29] authors use stochastic geometry to characterize key performance metrics for neighboring WiFi and LTE networks in unlicensed spectrum in order to provide semi analytical results for some interesting metrics such as e medium access probability, coverage probability and density of successful transmissions. The main drawback of these works is the unrealistic assumptions of saturated users (i.e., not considering flow dynamics) and saturated BSs (i.e., assuming that all BSs interference at full power, all the time).

Regarding the latter shortcoming two notable exceptions are [30] and [31], where the authors consider variable cell loads and load-aware interference models. In [30] the main idea is that the density of the BS is thinned with respect to the load of each tier, in order to obtain more accurately the performance of the network. In [31] authors assume uniform users distribution and by using only PHY layer characteristics authors analyze the coupling problem between cell load factor and user's performance. They prove both sufficient and necessary conditions for the feasibility of the aforementioned load-coupling problem.

The authors in [32], in order to derive some performance metrics for heterogeneous cellular networks with energy harvesting used a simple birth-death Markov model to model the energy levels of a BS. In this work, a BS can be off (not-transmitting) for a period of time, until to harvest some energy. Some results from stochastic geometry and some from queueing theory combined together but without examining the flow-level performance of the network.

In [33] the authors derive the coverage probability and downlink capacity for randomly placed network. Again the BSs are assumed to be saturated, but in comparing with [20] the amount of users in a cell is distributed according to PPP. In Lemma 1 they derive the probability mass function about users cardinality in a randomly placed network. This result was not under our consideration when we derived the same result in Appendix A. Finally, we decide to included this result in the thesis because we follow a totally different approach than [33] and furthermore we present an asymptotic approximation that is very important in terms of computational complexity.

2.3.2 Queueing Theory

In a different research thread, a considerable amount of studies exist that focus on the flow level dynamics in cellular systems. [34] and [35] use queuing models to take into account the random nature of traffic arrivals and departures, in order to obtain the flow-level performance of different schedulers. Those works provide insight and closed form expressions for different queues that can be used to model the schedulers of wireless networks. The queue models are taking into account various properties such as flows' priority, fairness, parallel service, different flow rates.

In [36] the author in order to analyze wireless network systems such as 3G/3G+, models the BS scheduler as a multi-class processor shearing queue and assumes random finite-size service user's demands in order to provide explicit formulas for performance metrics (distribution of the number of active users per class, mean response times, mean throughput, etc.). [37] extends the

previous work, and takes into account spatial component of offered traffic. Authors prove closed form expressions about the flow level performance of a single BS network (no interference), line networks and small hexagonal. We should mention that the interfering BSs are assumed saturated.

However, these works only consider simple cellular topologies (e.g., line networks, or small hexagonal topologies), and assume a known rate distribution and always ON interference. [38] attempts to take into account the performance dependency between nearby BSs, when one considers load-based interference, and propose a methodology to derive some performance bounds. Nevertheless, this work also considers simple topologies.

The modeling of BS as queueing models had applied in [39] where the authors follow a queuing analytic approach to study the impact of turning off a BS on neighboring ones for different traffic models, but they only do a network-wise performance analysis through numerical simulations. In [40] authors solve the optimization problem of user association by taking into account the energy consumption as well as the flow-level performance (delay).

2.3.3 Intersection

The following recent works attempt to combine more sophisticated topologies with flow-level performance. [41] provides an analytical framework that calculates the stability of a Poisson Bipolar network. However, in a Poisson Bipolar network, each BS has a dedicated receiver at a random distance, so it cannot be used to examine the impact of users with different rates, associated to the same BS. Additionally, this work does not consider other flow-level metrics (e.g. delay) besides stability. In [42] the authors assume homogeneous PPP topologies for both BS and users in order to capture uplink performance, considering flow-level traffic dynamics. However, the authors assume a saturated interference scenario, not capturing the interplay between load-based interference and flow-level performance. In [43] and [44] the authors also model flow-level performance in a randomly placed network (h-PPP) using results from queuing theory as well. However, the BS and user spatial coordinates are assumed as input (rather than considering any specific model, stochastic or not). What is more, the user MCS distribution is further assumed to be another input to the problem, in order to avoid the key coupling problem between the MCS distribution and network load, which is at the core of this analytical problem. Hence, while useful, this framework can only be considered as a helpful tool that may accelerates the simulation process, rather than an analytical framework for an arbitrary, randomly placed network. Finally, in [44] and [45] the same authors perform a mean-cell analysis towards deriving analytically the mean BS load of a randomly placed network. The mean-cell approximation assumes that all BSs serve exactly the same number of users, so all BSs produce the same amount of interference. Due to the simplicity of the mean-cell approximation the framework is accurate only for the low load case. As a final remark, we note that all the aforementioned works [43–45] assume that the number of users in a cell is constant (uniform distribution of users) which does not allow to capture load distribution statistics and thus congestion probability for a BS.

Summarizing, the technical novelty of our work compared to the few related attempts to derive flow-level performance for large random networks consist of one or more of the followings: (i) Our framework models the number of users in a cell as a random variable, allowing us to derive the probability distributions for performance metrics of interest (load, utilization, congestion probability, delay) across BSs in the network, rather than just a “mean” BS. (ii) We

provide an explicit analysis and formula for both the mean delay and load of a base station, that is significantly more accurate even in the average sense, compared to mean-cell analysis. (iii) The state-of-the-art models of [45] assume that the MCS distribution, necessary to derive flow-level performance metrics, is actually given. We derive this MCS distribution analytically. (iv) Finally, unlike related works, we consider different queueing models for LTE and WiFi BSs to capture the respective MAC better.

As a final note, the seminal work of [46] addresses the optimal user association problem in a single tier from a flow-level dynamics point of view, proposing a load-based association algorithm. While optimal user association within a tier is not considered in this paper, we use a similar load-based approach for choosing between tiers, during our evaluation.

Chapter 3

PHY and MAC layer Modeling of LTE and WiFi RATs

In this chapter we consider our two main radio access technologies (LTE and WiFi), in order to mathematically model their PHY and MAC characteristics. In real implementations the PHY and the MAC layers of a telecommunication system are quite complex. In this chapter our goal is to do a proper system abstraction in order to end up to simple but still accurate expressions that capture the PHY and the MAC performance the radio access technologies of our interest. In the following chapters we will rely upon those expressions in order to construct and analyze the performance not of a single BS but of the entire network.

3.1 Introduction

The scope of this chapter is the proper modeling of both PHY and MAC layer of two widely used radio access technologies (RATs), LTE and WiFI. That being so, we have two main tasks, one for each layer (one for PHY and one for MAC).

- To match the user's SINR with the corresponding data rate. This tasks needs two internal steps: i) specify the SINR threshold of each MCS, ii) specify the rate of each MCS.

- To answer, how the resources are allocated with the presence of other users, and how the overall performance of the systems depends on the users' rate distribution. In other words, we should model each RAT scheduler with a proper queueing system in order to be able to analyze the dynamic behavior of a base station in terms of incoming load.

We suppose 20MHz eNodeB with a single antenna and a 802.11n single stream access point (AP), both of them operate with 20MHz bandwidth. This is a baseline scenario, but our results could be easily extended to other cases with proper abstraction.

3.2 PHY modeling

3.2.1 Introduction

Actual RATs do not provide an elegant way to calculate the user's rate, so it is common, when analyzing wireless networks to use the Shannon's theorem, as it constitutes a more simplified approach. When a single network is being analyzed, this assumption does not affect the validity of the qualitative results. However, in the case of modern HetNets, and especially when the heterogeneous networks operate with different RAT, this assumption does not hold. The user's rate of different RATs does not scale with the same way, with respect to SINR.

For instance, Fig. 3.1 presents the output of the PHY layer modeling procedure that will be presented in this section, every marker corresponds to an MCS and the x-coordinates of them are the SINR threshold τ_i and the y-coordinates are the corresponding rates.

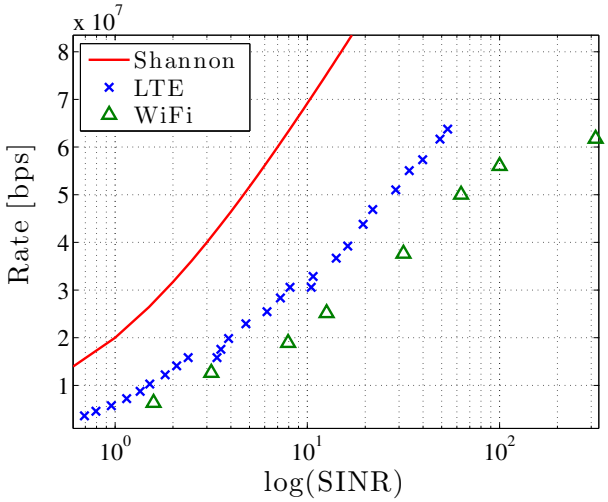


Figure 3.1 – Comparison between LTE and WiFi rates with respect to SINR. The solid line represents the Shannon's limit

LTE performs 37%, on average, closer to Shannon than the WiFi, at their common operating SINR range. Thus, for those HetNets, if the rates of both networks modeled according to Shannon's theorem, WiFi will be overestimated compared to LTE.

In order to produce Fig. 3.1 initially we need the characteristics (rate and SINR threshold) of supported MCS modes of each RAT. The rate of each MCS is always defined at the corresponding protocol description documents [15], [16]. Further, we need the operation threshold for each mode, this value is *not* always defined in the protocol since it heavily depends on the receiver implementation characteristics. So, we will need one SINR table for the LTE modes and one for the WiFi. The reference receiver of LTE will be the OpenAirInterface[1] platforms. For the WiFi we used a generic SINR threshold that is presented at [47].

[1] http://www.openairinterface.org/

3.2.2 LTE

From the OpenAirInterface LTE downlink simulator [48], Block Error Rate (BLER) vs SNR, for LTE Tx mode 1 (downlink use Single-antenna port, the port 0) [49], is generated for each MCS and shown in Fig. 3.2. Consequently, for a given BLER threshold (commonly at 10^{-1}) the SINR threshold (τ) for each MCS can be specified. Additionally, if we are interested for theoretical analysis, we can combine the knowledge for SINR distribution (or "coverage probability") with MCS threshold τ in order to end with MCS distribution.

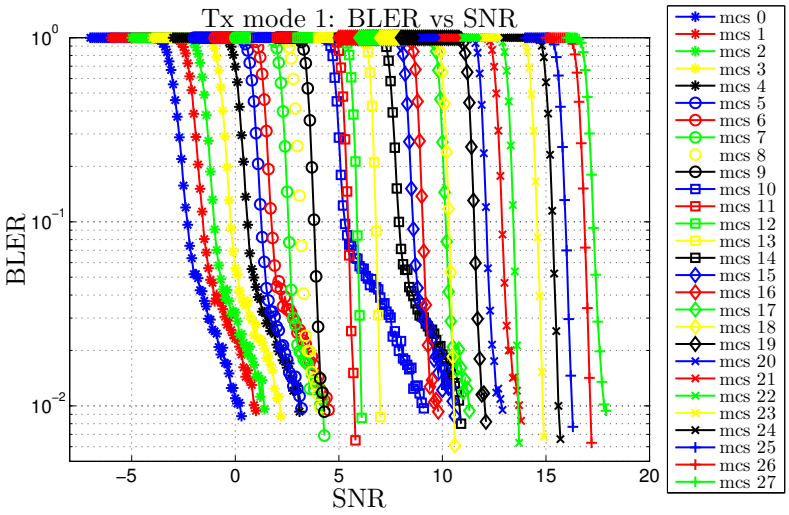


Figure 3.2 – BLER with respect to SINR for different MCS of LTE Tx mode 1, downlink use Single-antenna port

LTE uses orthogonal frequency-division multiplexing (OFDM) on the down link and divides the total frequency and time resources into resource blocks (RB) [49]. The size of a RB is 180kHz in the frequency domain and 0.5ms in the time domain, Fig. 3.3.

In the 20MHz bandwidth configuration there are 100 RB (also plus some white spaces for system's robustness to intra-cell interference). Each two RBs are grouped into one subframe with period one Transmission Time Interval (TTI), 1ms. We assume that all users will be allocated the same amount of subframes on average, so for a given number of associated users (n) at an eNodeB, each of them will be allocated $\frac{10^5}{n}$ subframes per second.

For a given *mcs*, LTE PHY specification 36.213, section 7.1.7.1 maps the index of the MCS to the index of the Transfer Block Size (I-TBS) for Downlink (part of this matrix is shown at table 3.1), which, together with the number of RBs, defines the amount of transferred bits per TTI, section 7.1.7.2.1 (part of this matrix is shown at table 3.2). Thus, by combining all the previous, for a given SINR value we are able to calculate the bit rate.

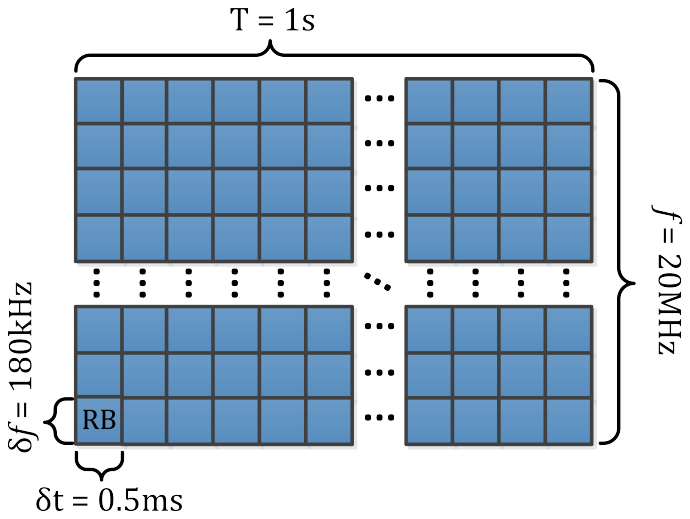


Figure 3.3 – Representation of LTE Resource Blocks

Table 3.1 – Mapping between mcs index and transfer block size (TBS) index

I_{mcs}	modulation order	I_{TBS}
1	2	1
2	2	2
3	2	3
4	2	4
5	2	5
6	2	6
7	2	7
8	2	8
9	2	9
10	4	9
11	4	10
⋮		

Table 3.2 – Number of transmitted bits with respect to TBS index and the number of RB

I_{TBS}	N_{RB}					
	1	2	3	4	5	6
1	16	32	56	88	120	156
2	24	56	88	144	176	208
3	32	72	144	176	208	256
4	40	104	176	208	256	328
5	56	120	208	256	328	424

...

⋮

3.2.3 WiFi

For WiFi 802.11n and ac, for each *mcs* we can extract the SINR threshold (τ) for each MCS from [47]. Part of this matrix could be seen at Fig. 3.4 and the physical data rate $Rate(msc)$ could be obtained from [50].

	SNR in dB	11	12	13	14	15	16	17	18	19	20
802.11b	20MHz	MCS 2	MCS 2	MCS 2	MCS 2	MCS 2	MCS 3	MCS 3	MCS 3	MCS 3	MCS 3
802.11a/g	20MHz	MCS 4	MCS 4	MCS 4	MCS 4	MCS 5	MCS 5	MCS 5	MCS 6	MCS 6	MCS 7
802.11n	20MHz	MCS 3	MCS 3	MCS 3	MCS 3	MCS 4	MCS 4	MCS 4	MCS 5	MCS 5	MCS 6
802.11n	40MHz	MCS 1	MCS 2	MCS 2	MCS 3	MCS 3	MCS 3	MCS 3	MCS 4	MCS 4	MCS 4
802.11ac	20MHz	MCS 3	MCS 3	MCS 3	MCS 3	MCS 4	MCS 4	MCS 4	MCS 5	MCS 5	MCS 6
802.11ac	40MHz	MCS 1	MCS 2	MCS 2	MCS 3	MCS 3	MCS 3	MCS 3	MCS 4	MCS 4	MCS 4
802.11ac	80MHz	MCS 1	MCS 1	MCS 1	MCS 1	MCS 2	MCS 2	MCS 3	MCS 3	MCS 3	MCS 3
802.11ac	160MHz	MCS 0	MCS 0	MCS 0	MCS 1	MCS 1	MCS 1	MCS 1	MCS 2	MCS 2	MCS 3

Figure 3.4 – WiFi's SINR thresholds for each MCS

The MAC performance the WiFi has been analyzed by Bianchi in [51]. In order to compute the impact of collisions, unused periods, overhead packets, RTS/CTS, etc., on the system's performance he modeled a system with a fixed number of n nodes (as nodes we assume both users and WiFi access point) were each node transmits according to an exponential back-off. For the aforementioned system Bianchi designed and solved its Markov chain representation in order to obtain its dynamic performance. It is worth mentioning that the users are assumed to be saturated, so they have always information to send, and if they do it or not depends only on the exponential back-off.

In our work we took into account an expansion of Bianchi's model, which is presented in [52] in order to include newer techniques of 802.11n and ac (frame aggregation and block of ACKs) that raise the utility of the MAC layer. Especially when the channel conditions are good, thus, the data rate is high, the overhead of idle periods caused by the back-off mechanism is too high compared with the transmission time of a packet, so the system performance increases considerably if we transmit more frames at once before doing the back-off process (frame aggregation). The percentage of the successful channel usage / normalized system throughput (% channel usage, P_s) w.r.t. the number of users n_u and different rates shown at Fig. 3.5. As we can notice,

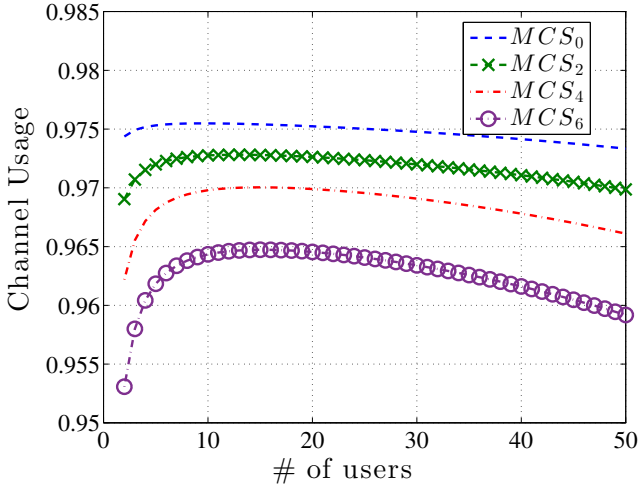


Figure 3.5 – Percentage of successful channel usage of WiFi 802.11n/ac systems with frame aggregation

for a reasonable number of connected users the performance of the MAC layer is roughly the same for a given MCS and does not depend on the number of connected users. Therefore, the average user throughput with a given MCS in the presence of n other users is given by

$$Rate(mcs, n) = P_s(n, mcs)\frac{Rate(msc)}{n} \ . \tag{3.1}$$

3.3 MAC modeling

When more than one users are served in parallel by a BS, the BS operates as a *queueing system*. The service rate depends on the number of associated users and their SINR (BS load). Additionally the service rate depends also on the centralized scheduler (e.g., in the case of 3G/4G) or distributed media access control (MAC) protocol (in the case of WiFi) which decide how the available resources will be distributed between users. While a number of different scheduling algorithms exist, the majority of them try to allocate the available resources between competing flows (e.g. LTE resource blocks, WiFi channel) in a fair or proportionally fair manner.

3.3.1 LTE

Assume the BS allocates the same amount of resources to all flows, and they are served simultaneously, e.g., with a round robin, TDMA-like algorithm. If the service time slot is small (e.g., of packet size) compared to the total size of a flow, the flow level performance at that BS can be approximated by a multi-class M/G/1 Processor Sharing (PS) system, as shown in Fig. 8.2. This model has already been used to analyzed 3G/3G+ BS performance [36, 37]. While each flow shares the channel for the same amount of time (hence "resource fair"), during that time it might transmit at a different rate, depending on its SINR and resulting MCS (hence the "multi-class" service).

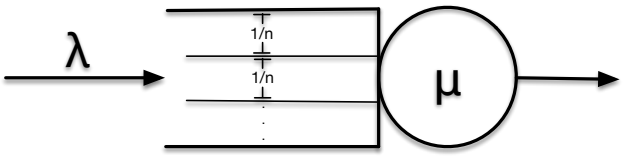


Figure 3.6 – M/G/1/PS Resource Fair

LTE schedulers are significantly complex, allocating competing flows both time and frequency resources (Resource Blocks), possibly taking into account the queue backlog of each flow and flow priority, and also attempting to take advantage of instantaneous SINR variations in time and frequency to achieve further multi-user diversity [49]. While a large number of algorithms have been proposed (see e.g., [53] for an extensive survey), in the lack of special priority traffic, most implemented schedulers lead to a proportionally fair allocation between flows [49] and can also be approximated by a similar multi-class M/G/1 PS queue. The following is a direct application of the multi-class M/G/1/PS result [54].

Lemma 3.3.1. For a BS with n users generating flows of mean size $\langle s \rangle$, with instantaneous transmission rates drawn from distribution $f_{\mathrm{R}}(r)$, and allocated resources by a resource fair scheduler, the effective service rate of the cell is

$$\langle \mu \rangle_{\mathrm{rf}} = \left(\sum_i \frac{f_R(r_i) \cdot \langle s \rangle}{r_i} \right)^{-1} \text{ flows/sec}, \tag{3.2}$$

and the mean flow delay is given by

$$E[T]_{\mathrm{rf}} = \frac{1}{\langle \mu \rangle_{\mathrm{rf}} - n\lambda_f} \ . \tag{3.3}$$

We further define the BS's load as

$$\rho = \frac{\text{input job rate}}{\text{service job rate}} = \frac{n\lambda_f}{\langle \mu \rangle} \tag{3.4}$$

when the system is stable $\rho < 1$.

Performance gains from opportunistic scheduling can be included in the above equation as a multiplicative factor in front of $\langle \mu \rangle_{\mathrm{rf}}$.

3.3.2 WiFi

Some schedulers attempt to achieve fairness more aggressively, by trying to equalize per flow throughput for all nodes. For example, if two concurrent flows experience different channel conditions (say one being "far" and one being "near" the BS) a throughput fair scheduler will attempt to give more resources to the flow with the worse channel (e.g., more resource blocks in the case of LTE, or schedule the far flow more often in the case of 3G). This can be seen as a Generalized or Discriminatory Processor Sharing system (a generalized version of the M/G/1/PS) [35], with different weights per flow that, for throughput-fair systems, can be taken as inversely proportional to the average rate experienced by that flow.

It is known that throughput fair schedulers perform poorly compared to proportionally fair ones, and thus are not often considered [34]. Nevertheless, throughput fair scheduling turns out to be a good approximation of how the 802.11 WiFi MAC allocates resources between flows [55]. In WiFi, all nodes compete for the channel and when they do get access, in the basic implementation, they send a single frame and then have to retry. WiFi like LTE supports rate adaptation, therefore each frame might be transmitted at a different rate, depending on the maximum MCS that can be offered to the respective node. Nevertheless, due to the random access MAC, each node gets access with equal chance, regardless of their distance from the AP. If each flow corresponds to a large number of frames (usually a good assumption given the small max size of a frame), this essentially equalizes the long-term throughput of each flow, regardless of its MCS. Hence, the WiFi scheduler for a single BS could be seen as throughput-fair, and can be modeled as a Discriminatory Processor Sharing (DPS) queue.

The following lemma presents the mean service time ($E[T]$) for such a throughput-fair scheduler in a system with rate adaptation.

Lemma 3.3.2. The mean per flow delay for a throughput fair system with input flow rate λ, and flows being served with rates r_k drawn from a pmf $f_R(r_k)$, can be calculated according to

$$E[T]_{\text{tf}} = \sum_k f_R(r_k) \left(\frac{\langle s \rangle / r_k}{1 - \lambda / \langle \mu \rangle_{\text{tf}}} + \frac{\sum_j f_R(r_j) \lambda (1 - \frac{r_j}{r_k}) (\langle s \rangle / r_j)^2]}{2(1 - \lambda / \langle \mu \rangle_{\text{tf}})^2} \right) , \tag{3.5}$$

where $\langle s \rangle$ is the mean flow size and $\langle \mu \rangle_{\text{tf}}$ is mean service rate of cell, equals with

$$\langle \mu \rangle_{\text{tf}} = \left(\sum_k \frac{f_R(r_k) \cdot \langle s \rangle}{r_k} \right)^{-1} . \tag{3.6}$$

Proof. Let us first consider a throughput fair system, and derive the mean service rate, Eq. (3.6). Consider a long time interval during which N packets get transmitted, corresponding to different flows. Assume each packet is of equal size S (e.g., the max WiFi frame size) but is transmitted with a possibly different rate r_k drawn from pmf $f_R(r_k)$ with K discrete values, depending on the MCS used for transmitting that packet. Assume that out of these N packets, N_k are transmitted with rate r_k, ($\sum_k N_k = N$). Hence, the *average* transmission rate in terms of bits/sec for these N packets is

$$\frac{\text{bits in } N \text{ pkts}}{\text{transmission time for } N \text{ pkts}} = \frac{N \cdot S}{N_1 \frac{S}{r_1} + \cdots + N_K \frac{S}{r_K}} . \tag{3.7}$$

However, as N goes to infinity, the N_k converges to its mean value $f_R(r_k) \cdot N$ by the law of large numbers, hence the denominator of Eq. (3.7) converges to

$$\lim_{N \to \infty} (N_1 \frac{S}{r_1} + N_2 \frac{S}{r_2} + \cdots + N_K \frac{S}{r_K}) = \sum_k f_R(r_k) \cdot N \cdot \frac{S}{r_k} . \tag{3.8}$$

Since $\frac{1}{x}$ is continuous and all $r_k > 0$, we can use the Continuous Mapping Theorem [56](Th. 5.23) to show that Eq. (3.7) converges to

$$\frac{1}{\sum_k f_R(r_k) \cdot \frac{1}{r_k}} . \tag{3.9}$$

where N and S canceled out. Eq. (3.9) thus gives the average transmission rate of the scheduler over a sufficiently long sample path of packets. Since the system is ergodic, we can divide with the mean flow size $\langle s \rangle$ to get Eq. (3.6).

To go beyond the mean load and derive the mean delay for this system, we use the approximation from Avrachenkov *et al.* [57] for DPS systems, which to our best knowledge, provides the most accurate solution, assuming large enough flow sizes. Specifically, for a given network load, the expected delay for flows of class k having size x, denoted as $E[T_k(x)]$, asymptotically converges to

$$\lim_{x\to\infty}\left(E[T_k(x)] - \frac{x}{1-\lambda/\langle\mu\rangle_{\text{tf}}}\right) = \frac{\sum_j \lambda_j(1-\frac{w_k}{w_j})E[X_j^2]}{2(1-\lambda/\langle\mu\rangle_{\text{tf}})^2}\ . \tag{3.10}$$

We applied the equation above to our system, by having classes corresponding to different MCS and weights $w_i = 1/r_i$ inversely proportional to the service rate. $E[X_i^2]$ is the second moment of service requirement (flow sizes normalized in seconds) for flows of class i, which we approximate with $E[X_i^2] \approx (\langle s \rangle / r_i)^2$. The incoming job rate λ_i of class i: assuming that the probability of an incoming job to be of class i is $f_R(r_i)$ then $\lambda_i = f_R(r_i) \cdot \lambda$.

Putting everything together gives us Eq. (3.5).

■

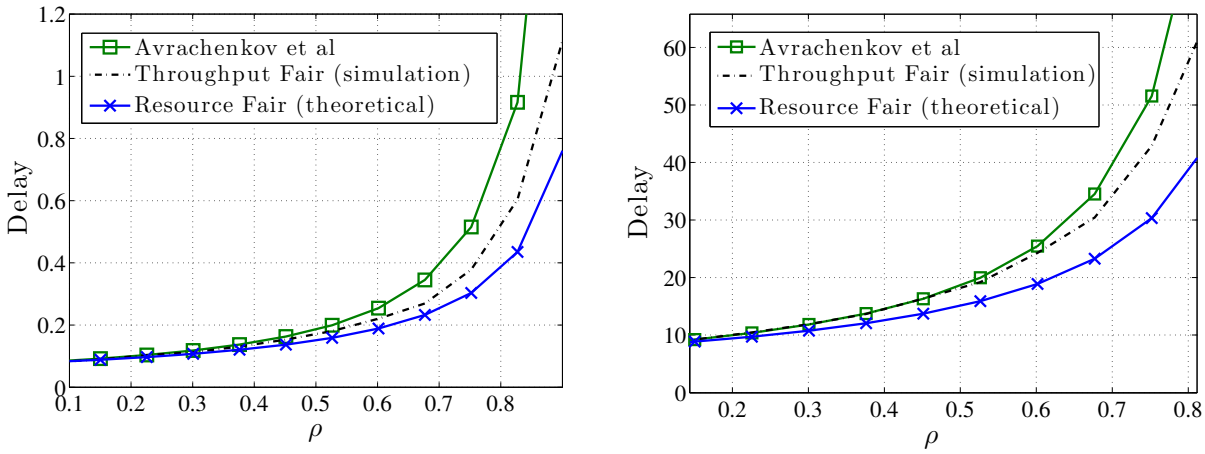
Note that the above analysis, when applied to 802.11, ignores the impact of collisions and RTS/CTS frames and thus is an upper bound. Nevertheless, as we saw in the previews section for the PHY layer, in light of the high speeds and features of 802.11n\ac, such as frame aggregation or block of ACK transmissions (by a single node), implies that the impact of such overhead can be safely ignored. Additionally, we include those overheads in the calculation of the data rate Eq. (3.1).

It is interesting to observe as we note in Chapter 2 that the above result implies that *the mean service rate, in the long run, for a WiFi system with rate adaptation, turns out to be the same as that of a resource-fair system (Eq.* (3.2)). Nevertheless, this does *not* imply that the mean flow delay is also the same, as the scheduling discipline is different (DPS instead of PS). Unfortunately, there does not exist a closed form solution for the mean flow delay of a throughput fair system. We will therefore consider the following two approximations to model the MAC performance of the WiFi.

Approximation 1: When the BS load is low (i.e., $\frac{\lambda}{\langle\mu\rangle} \to 0$), flows rarely "compete" with each other, and it is easy to see that the mean delay is approximately equal to the resource fair case, i.e., Eq. (3.3). This is also *a lower bound* on the delay, for higher load values. Furthermore, the observed poor performance of WiFi [55] has led researchers to propose slight modifications of 802.11, taking advance of the new feature of frame aggregation [52], in order for WiFi to operate closer to a resource fair scheduler.

Approximation 2: For general loads, we can use Avrachenkov's approximation as presented in Eq. (3.5). As we mentioned, this result is an asymptotic as flow sizes is going to infinity, but even for small flow sizes our simulation results show that the approximation is decent. Roughly, for small flow size the performance of a throughput fair system is equally spaced between the two approximations and as the size is increasing the performance of the system is approaching the approximation 2, Fig. 3.7 a) presents the comparison between the asymptotic approximation,

simulation results and resource fair scheduler for average $\langle s \rangle = 1$MBytes flow size. While the flow size is increasing the performance of the system is approaching the asymptotic approximation, Fig. 3.7 b) shows resource fair, approximation, and simulation results for flow size of $\langle s \rangle = 12.5$ MBytes.



a: Average flow size 1 Mb

b: Average flow size 12.5 Mb

Figure 3.7 – The line with square markers indicates Avrachencov's approximation for a throughput fair system, the dashed line in the result of the packet-level simulator and the line with $\times$ markers represents the best case of resource fair system

Chapter 4

Performance Analysis of Single tier Network

In this chapter, we use the results of a single BS modeling of Chapter 3 and we develop a flexible and accurate analytical model of large single-tier networks with random BS placement, in order to understand the impact of key network parameters like BS density and load on the network performance. The main goal is to understand the flow level dynamics of such a system, assuming non-saturated users and studying the congestion statistics for BSs and the per flow delay. To achieve this, we base our analysis on two main tools: (a) stochastic geometry, to understand the impact of topological randomness and coverage maps and (b) queueing theory, to model the competition between concurrent flows within the same BS. Our model is then applied to the popular radio access technologies, such as LTE and WIFi. Our results provide some interesting qualitative and quantitative insights about the performance of those networks and will be used in the following chapters in order to study the energy consumption and flow-level performance tradeoff as well as the performance of multi-tier networks.

4.1 Introduction

The trend of modern networks is to become denser, irregular placed, and more heterogeneous, due to the often unplanned and incremental deployment of new (small cell) BSs. As a result, analyzing such networks, e.g., for protocol comparison or network planning, becomes increasingly challenging. What is more, the usually considered metrics in such analyses, like SINR or capacity, often fail to capture the actual user experience, because flow-level performance (delay, congestion probability, etc.) strongly depends on the network's load and not only the channel conditions [9, 37]. A better metric is latency, which is one of the main performance indicators of 5G technologies [10, 11].

To this end, in this chapter we present a flexible and accurate model that analyses the performance of random placed networks, in order to understand the impact of important network parameters (BS density, load) on the network's performance. Our model consists of randomly located Base Stations as well as randomly placed users. Users are assumed to be non-saturated, randomly generating requests for new file/flow downloads of varying sizes and they perceive performance in terms of the average delay to finish such a download.

Our analysis is based on the combination of two key theoretical tools that have recently provided many insights on cellular network performance: (i) We use *queueing theory* to model the performance of dynamic flow arrival and service via the respective scheduler, at the level of a single BS; (ii) We utilize *stochastic geometry*, in order to understand the impact of topological randomness and interaction/competition between BSs at the network level, in order to derive statistics about the *number of users associated with a base station*, and the *modulation and coding schemes (MCS)* offered at each BS. Both these quantities serve as key inputs to the BS queueing model: the former to define the total traffic intensity (in terms of flow arrivals) a given BS has to serve, and the latter to define the average service rate (in terms of flow departures) that a BS is able to offer.

There is a number of works that examine the performance of a network using tools from stochastic geometry: [58] provides distribution of the coverage areas and [20] that derives the distribution of the interference assuming that all neighboring BS are saturated. Additionally, flow-level dynamics of cellular networks have been studied in [34,36,37,46,59], some of those focus on spectral efficiency and BS instantaneous throughput, while the rest assume simple cellular topologies (e.g., line networks, or small hexagonal topologies). Compared to these related works, to our best knowledge this is the first work jointly considering the stochastic geometry of the network and flow-level dynamics. Summarizing, the main contributions that presented in this chapter are:

I) We present a new analytical result deriving the probability mass function (pmf) of users' cardinality at an arbitrary BS, if both, users and BSs distributed as homogeneous Poisson point process;

II) We propose an analytical model that captures both physical and MAC layers performance, providing statistics for coverage maps and MCS distributions, as well as flow-level performance as perceived by the user (flow delay) and the network operator (congestion probability);

III) We derive a semi-analytical model that computes the coverage probability of a random placed network, considering the fact that neighboring BSs are not fully loaded (non-saturated) and thus create dynamic interference proportional to their load.

The rest of the chapter is organized as follows. In Section 4.2 we model performance at the BS level. In Section 4.3, we are modeling the PHY layer. In Section 4.4, we derive the users cardinality distribution for our topology and we compute the arrival rate. Section 4.5 presents the steps in order to specify the service rate, which includes both pure analytical formulas and technical details for each one of the chosen RAT. Section 4.6, validates our theoretical model and analyses the networks of interest. Section 4.7 presents the future steps of our work.

4.2 Performance at the BS level

We assume that each BS experiences a *dynamic* traffic load and we would like to study the performance at *flow-level*. We state here our assumptions regarding a single randomly chosen BS and comment where necessary.

A.1: Each *connected* user to a BS generates new *flow* requests randomly, and independently of other users, according to a Poisson Process with density λ_f.

A.2: A flow is a sequence of packets corresponding to the same user or application request (e.g., a file or web page download). Each flow has a random size, in terms of bits, drawn from a *generic* distribution with mean value $\langle s \rangle$.

A.3: The number of users n associated with a BS is a *random* variable with probability mass function (pmf) $f_N(n)$ that depends on the density of the BSs (λ_{BS}), the density of users (λ_u), and the association criteria. This pmf will be derived in Section 4.4.

The following Lemma follows easily, by using a simple Poisson merging argument [17].

Lemma 4.2.1. If n users are associated with a given BS, the aggregate flow arrival process to that BS is Poisson($n\lambda_f$).

Remark: While a Poisson arrival model is pretty standard in related literature, note that if the number of users n at a BS is relatively large, assumption (A.1) can be relaxed to more general traffic arrivals, and we can then use the Palm-Khintchine theorem [17] to support Lemma 4.2.1 as an approximation.

A.4: In the absence of other flows, *a single flow will be served at full rate*, with the maximum modulation and coding scheme that the BS can offer to that UE, which in turns depends on the SINR-BLER specifications for that RAT. The rate of the arbitrary user could be assumed as a random variable and the corresponding pmf, $f_{\mathrm{R}}(r)$, is presented in Section 4.4 and derived in the Appendix A.

We will assume a single MIMO layer and a single carrier in our analysis [49]. Increased rates due to spatial multiplexing and carrier aggregation can be easily included in the model with a proper physical abstraction models.

4.2.1 Queueing Model for BS Schedulers

When more than one flows are served in parallel by a BS, the BS operates as a *queueing system*. The service rate for a flow is generally smaller than what assumption (A.4) predicts. It depends on the number of active flows (BS load), and the centralized scheduler (e.g., in the case of 3G/4G) or distributed media access control (MAC) protocol (in the case of WiFi) which decides how the available resources will be distributed between flows. While a number of different scheduling algorithms exist, we assume for simplicity only the resource-fair one.

Resource Fair Scheduler

Regarding the LTE scheduler, as we mentioned Chapter 3 we can model the LTE scheduler a resource fair multi-class M/G/1 Processor Sharing (PS) system. As we mentioned before a multi-class, resource fair M/G/1/PS system, with n users that generates flows of mean size $\langle s \rangle$, with instantaneous transmission rates drawn from distribution $f_{\mathrm{R}}(r)$, has effective service rate

$$\langle \mu \rangle = \left(\sum_r \frac{f_R(r) \cdot \langle s \rangle}{r} \right)^{-1} \text{ flows/sec.} \tag{4.1}$$

The average users delay of this system is

$$Delay = \frac{1}{\langle \mu \rangle - n\lambda_f} \ , \tag{4.2}$$

The load of a system could be defined as $\rho = \frac{\text{input job rate}}{\text{service job rate}}$, for our case the average network load could be defined as

$$\rho = \frac{\zeta \cdot \lambda_f}{\langle \mu \rangle} \ , \tag{4.3}$$

where $\zeta = \lambda_u/\lambda_{BS}$. Additionally, the system is stable when $\rho < 1$. There are a lot of ways to express the network load, among them we chose the percentage of time that a BS is ON or due to ergodicity the percentage of the network that are ON at a random moment.

Another often studied scheduler (and good approximation for the 802.11 [55]) is the throughput fair, which equalizes the per flow throughput for all nodes. We ignore it here and we assume that 802.11 performs as resource fair scheduler, which asymptotically the best case, as the load goes to zero as we mention in Chapter 3, for the following reasons: i) assuming 802.11n characteristics the difference between those two schedulers is negligible, for small average flow size ($\approx$ 1 Mb) and utilization less than 70%, [60], ii) with minor modifications 802.11 is able to operates almost as a resource fair scheduler with proper selection of the amount of aggregated frames per MCS [52].

4.2.2 Network-wide Performance

Our goal in this chapter is to understand the network's performance considering the three following dimensions:

- *Stability (congestion probability)*: We would like to know the percentage of BS whose input load $n \cdot \lambda_f$ exceeds the available service capacity $\langle \mu \rangle$ (i.e., $\rho > 1$) thus exhibiting per flow delays that grow to infinity.

- *Utilization*: Network utilization expresses the probability of a randomly chosen BS to be active at a random time instance or the percentage of network's active BS at an arbitrary moment. It can be defined as the average utilization over all BSs, $\mathcal{U} = E[\mathcal{U}_{\mathcal{BS}i}]$, where $\mathcal{U}_{\mathcal{BS}i} = \min(\rho_i, 1)$ is the percentage of time that i-th BS is active, ρ_i according to Eq. (4.3).

- *Per flow delay*: we would like to know the expected network-wide delay for a randomly chosen user flow, when this flow is served by a stable BS.

Based on the previous discussion, for a single tier network these metrics depend on the same two key parameters:

1. The cardinality n of the users associated at a BS, which is a random variable with pmf $f_N(n)$ that depends on the topology of BS and user density.

2. The probability that each user is served with a given rate r, namely the rate distribution $f_R(r)$ for this BS that depends on the topology and interference between nearby BSs.

We derive $f_N(n)$ in Section 4.4 and then derive $f_R(r)$ in Section 4.5.

4.3 PHY Layer Modeling

Before we proceed with the derivation of the cardinality and rate probability distributions, we state here our assumptions about the network topology and physical layer model.

A.5: Users are distributed according to an independent homogenous Poisson Point Process with density λ_u.

A.6: The number of BSs inside an area S follows a homogeneous Poisson Point Process, Φ_{BS}, with density λ_{BS}. Therefore, in a given area S the number of BSs is a random variable according to

$$P(N = n \mid S) = \frac{(\lambda_{\text{BS}} S)^n e^{-\lambda_{\text{BS}} S}}{n!}, \quad n = 0, 1, \dots \quad . \tag{4.4}$$

It can be argued that the above model does not exactly capture current cellular networks, consisting mostly of macro eNodeBs that are usually carefully planned to maximize coverage, and could perhaps be better modeled by standard hexagonal or grid topologies. Nevertheless, in the case of WiFi access points or future, considerably more dense networks consisting mostly of pico- or femto-cells, topologies are expected to be considerably more random and uncoordinated, with BSs having a non-zero probability to be very close. We can consider the aforementioned two topologies as ideal and worst case scenarios respectively, in terms of interference. As shown in [20], the coverage probability in terms of the SINR threshold, in real BS deployments, lies in most cases roughly midway between the coverage probability in the two extreme cases above. Probably Hard core process could capture better the topology of the macro eNodeBs, but unfortunately it does not provide suitable analytical results.

A.7: A standard power loss propagation model is used. We assume a path loss exponent $\alpha > 2$ (for $\alpha \leq 2$ the denominator of SINR goes to infinity), Rayleigh fading at the channel with mean 1 and constant transmit power of P_{tx}. Thus, the received power at distance d from the BS is given by $P_{\text{rx}} = hd^{-\alpha}$ where h follows an exponential distribution, $h \sim \exp(P_{\text{tx}})$. Hence, the SINR is given by

$$SINR_i = \frac{P_{\text{rx}_i}}{\sum\limits_{n \neq i} P_{\text{rx}_n} + \sigma^2} \ , \tag{4.5}$$

where σ^2 is the thermal noise and calculated w.r.t the bandwidth (BW) from $\sigma^2_{\text{dBm}} = -174 + 10\log_{10}(BW)$ [61].

A.8: We assume that all BSs have equal transmit power and implement the same scheduling policy.

Assuming that on average, the received power is monotonic in respect to distance, our criterion is simplified to the closest distance criterion, so, the BSs's coverage areas could be represented by Voronoi Regions (Tessellations).

4.4 Cardinality of Associated Users

We are now ready to consider the pmf of the users' cardinality for an arbitrary BS, $f_N(n)$, which as explained earlier decides the total input traffic to each BS. It is observable that the size of an arbitrary cell is a random variable, depending on the random BS topology, and the number of users given a specific cell size is also a random variable. The proof for the following theorem as well as a useful and accurate asymptotic result can be found in Appendix A or to our technical report [62].

Theorem 4.4.1. Consider BSs distributed in 2D as a h-PPP with density λ_{BS}, and offering coverage to a set of users distributed as another h-PPP with density λ_{u}. Assume further that user association within this tier is done by using the closest-distance rule, as explained in Section 4.3.

Then, the probability of having exactly n users in an arbitrary cell, $f_N(n)$, is given by:

$$f_N(n) = \frac{343}{n!15}\sqrt{\frac{7}{2\pi}}\frac{\zeta^n}{(\zeta+\frac{7}{2})^{n+\frac{7}{2}}}\Gamma(n+\frac{7}{2}) \ , \tag{4.6}$$

where $\zeta = \frac{\lambda_u}{\lambda_{BS}}$ and $\Gamma(\cdot)$ is the gamma distribution. Figure 4.1 depicts the user cardinality pmf for different values of ratio ζ.

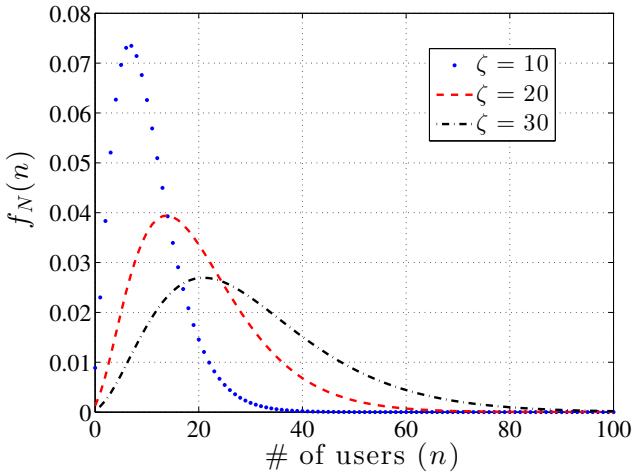


Figure 4.1 – Pmf of number of associated users per BS, were both the topology of BS and users is follows h-PPP for different values of ratio $\zeta = \frac{\lambda_u}{\lambda_{BS}}$

4.5 MCS Distribution for each RAT

We are interested in the maximum rate (or equivalently maximum MCS) that a user can receive from the BS that he is associated with, given a desired BLER. Our goal is to derive the rate distribution $f_R(r)$ in order to calculate the service rate $\langle\mu\rangle$ in terms of flows/sec for the average BS. This rate depends on the SINR for that user. A given SINR is mapped to an offered MCS [60]. The SINR in turn depends on both the distance of the user to the serving BS and the interference from other nearby BS. Furthermore, a nearby BS might not interfere if it is actually not transmitting at that time, which further complicates analysis. For this reason, we will first consider a "saturated" scenario where interfering BS are assumed to always be ON and interfering. We will then consider the case of load-based interference, where a BS only interferes if it is currently *active* serving at least one user.

For both interference cases (saturated, load-based), in order to calculate the rate distribution $f_R(r)$ we will use the coverage probability of the network and the SINR threshold for each MCS. Coverage probability is the probability that SINR of an arbitrary user is greater than a given threshold τ

$$p_c(\tau) = P\left[SINR > \tau | r\right] \ . \tag{4.7}$$

According to the previous discussion, the probability mass function of MCS is given by

$$f_{\text{MCS}}(mcs_i) = p_c(\tau_i) - p_c(\tau_{(i+1)}) \ . \tag{4.8}$$

In other words, the probability a user to operate with mcs_i is equal to the probability that his SINR is larger than the threshold of mcs_i minus the probability that his SINR is larger than the threshold of mcs_{i+1} (because then, he operates in higher MCS)

4.5.1 Rate Distribution for Always ON Interference

We will assume again that BSs and users are distributed according to independent homogeneous PPPs. In [20], the authors present an approach to derive the "coverage probability" of a randomly located user, i.e., the probability that the user's SINR is above a certain threshold. In doing so, it is assumed that interfering BSs always transmit with a power P_{tx}. This assumption is a good approximation when the load of the system is high, in which case the utilization of most BS is close to 1 (i.e., are serving users most of the time). It can also be a valid assumption if the SINR at the user is measured with respect to Reference Signals (i.e., "pilots") that are transmitted at specific times slots by all BS, regardless of whether a BS is serving users or not at that time [49]. Nevertheless, this is not always the case. As a result, in scenarios where BS utilization is low, this assumption might lead to fairly pessimistic results. We consider this case in following Section 4.5.2.

For the sake of completeness, we mention here again the main result from [20] that is applicable to our problem: Given a BS density λ_{BS}, and path loss constant α, the coverage probability for an SINR threshold T is

$$\begin{aligned} p_c\left(T, \lambda_{BS}, \alpha\right) &\triangleq \mathbb{P}\left[SINR > T\right] \\ &= \pi\lambda_{BS} \int_0^{\infty} e^{-\pi\lambda_{BS}u(1+\beta(T,\alpha))-\frac{1}{\mu}T\sigma^2u^{\alpha/2}} du \ , \end{aligned} \tag{4.9}$$

where $\beta\left(T,\alpha\right) = T^{2/\alpha}\int_{T^{-2/\alpha}}^{\infty}\frac{1}{1+u^{\alpha/2}}du$ and σ^2 is the additive noise.

If we assume that additive noise is negligible w.r.t. interference (a reasonable assumption for the dense modern networks) Eq. (4.9) can be significantly simplified as $p_c\left(T, \lambda_{BS}, \alpha\right) = 1/\left(1+\beta\left(T,\alpha\right)\right)$. Furthermore, if we assume that $\alpha = 4$, we obtain an elegant closed form solution

$$p_c\left(T, \lambda_{BS}, 4\right) = \frac{1}{1+\sqrt{T}\left(\pi/2 - \arctan\left(1/\sqrt{T}\right)\right)} \ . \tag{4.10}$$

Finally, assuming and SINR threshold τ_i for each MCS (mcs_i), the pmf of the MCS $f_{MCS}(mcs)$ can be obtained at Eq. (8.2) through the coverage probability.

$$f_{\text{MCS}}(mcs_i) = p_c\left(\tau_i, \lambda, \alpha\right) - p_c\left(\tau_{(i+1)}, \lambda, \alpha\right) \ . \tag{4.11}$$

Given the MCS, the actual rate can be easily calculated based on the total bandwidth of the system in question. Existence of multiple antenna ports and resulting MIMO layers can easily be added in this calculation. Similarly for independent carriers, by deriving the respective MCS for each.

Fig 4.2 use the LTE MCS thresholds as presented in Chapter 3 combined with Eq. (4.9) and Eq. (8.3) in order to derive the pfm of user's MCS distribution, and compare it with the

MCS distribution according to simulations. As we expected the theoretical results match the simulation results. The simulator predict slightly higher MCS, but this is explainable if we consider that we can not simulate an infinite topology, so there will be always edge BS that will suffer from less interference.

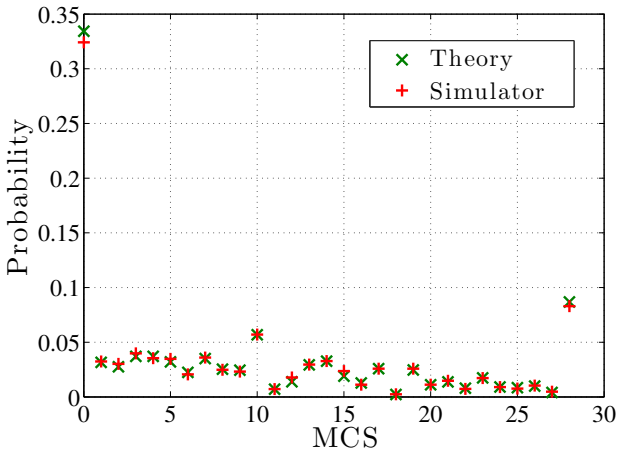


Figure 4.2 – User's MCS distribution in a homogenous PPP network

4.5.2 Rate Distribution for Load-based Interference

As mentioned earlier, the previous results assume that all BS are interfering all the time. In practice, when the load ρ of a BS A is low, e.g., $\rho = 0.5$, then BS A would be transmixtting and causing interference only 50% of the time [1]. This implies that another nearby BS B will be actually serving users at higher rates than the ones predicted in the saturated case. This, in turn, means that BS B will also have a higher $\langle\mu\rangle$ and thus lower utilization $\rho = \frac{\lambda}{\langle\mu\rangle}$ than the one predicted, which in turn creates less interference for BS A.

At flow level, this creates a system of dependent PS queues, which is notoriously hard to analyze at Markov chain level (see [38] for an attempt to derive some performance bounds). We choose to take here a different approach and we provide a semi-analytic approach in order to calculate $\langle\mu\rangle$ of those dependent BSs. Initially, we have to present the new coverage probability which takes into account the load of interfering BSs. The following lemma extends the previous analysis based on stochastic geometry, in order to approximate the coverage probability of the load-based interference scenario.

Lemma 4.5.1. The coverage probability of an arbitrary user in a random cellular network (assuming thermal noise negligible compared to interference), when the average utilization of BSs is $\mathcal{U}$, and each BS is interfering only for the amount of time that it is serving users (i.e., for a

[1] Even if the SINR estimate is based on the pilot signals, which are always transmitted at the designated LTE resource elements, the *actual* interference experienced during transmission will be lower in practice, leading to better effective rates (e.g., due to fewer HARQ retransmissions required).

percentage of time $\mathcal{U} \leq 1$) is given by

$$p_c^{lb}(\tau, \alpha, N_{max}) = \sum_{n=0}^{N_{max}-1} \left(f_N(n \mid \zeta) \frac{1}{1 + \mathcal{A}_{\mathcal{U}}} \right) + \overline{F_N}(N_{max} \mid \zeta) \frac{1}{1 + \mathcal{A}_{\mathcal{U}=1}} . \tag{4.12}$$

Where $N_{max} = \langle \mu \rangle / \lambda_f$ is the maximum amount of associated users per BS and

$$\mathcal{A}_{\mathcal{U}} = (\tau \mathcal{U})^{2/\alpha} \int_{(\tau \mathcal{U})^{2/\alpha}}^{\infty} \frac{1}{1 + u^{\alpha/2}} du . \tag{4.13}$$

We can consider the BS utility with respect to the number n of the associated users as $\mathcal{U} = \min(\frac{n}{N_{max}}, 1)$. Additionally, τ is the SINR threshold, α is the path loss exponent, f_N and $\overline{F_N}$ are the pdf and ccdf of users' cardinality and as previously $\zeta = \lambda_u / \lambda_{BS}$. The proof of Eq. (4.12) can be found in Appendix B.

Assuming $\alpha = 4$, Eq. (4.12) could further simplified by replacing $\mathcal{A}_{\mathcal{U}}$ and $\mathcal{A}_{\mathcal{U}=1}$ with

$$\begin{aligned} \mathcal{A}_{\mathcal{U}} &= \sqrt{\frac{\tau}{N_{max}} n} \cdot arccot \left(\frac{1}{\sqrt{\frac{\tau}{N_{max}} n}} \right) \\ \mathcal{A}_{\mathcal{U}=1} &= \sqrt{\tau} \cdot arccot \left(\frac{1}{\sqrt{\tau}} \right) . \end{aligned} \tag{4.14}$$

Service Rate

Taking into account that $N_{max} = \langle \mu \rangle / \lambda_f$, we can observe from Eq. (4.12) that the coverage probability depends on service rate $\langle \mu \rangle$. Thus, MCS distribution depends on the $\langle \mu \rangle$ as well (see Eq (8.2)). On the other hand, $\langle \mu \rangle$ is depending on MCS distribution as we can see at Eq. (4.1).

Due to the aforementioned dependencies we re-write Eq. (4.1) to its proper form

$$\langle \mu \rangle = \left(\sum_{mcs_i} \frac{f_R(mcs_i \mid \langle \mu \rangle) \cdot \langle s \rangle}{r(mcs_i)} \right)^{-1} . \tag{4.15}$$

To prove that Eq. (8.4) has maximum one solution with respect to $\langle \mu \rangle$ is not trivial at all. A study about the coupling problem between load and rate distribution was presented in [31] where authors studied the necessary conditions for the feasibility of the solution of a similar problem. As a remark we mention that the left part of equation Eq. (8.4) is a strictly increasing function with derivative equal to one, and the right part is again a strictly increasing function with respect to $\langle \mu \rangle$ but its derivative is strictly smaller than one for the cases of our interest, LTE or WiFi or Shannon's rates (calculated computationally).

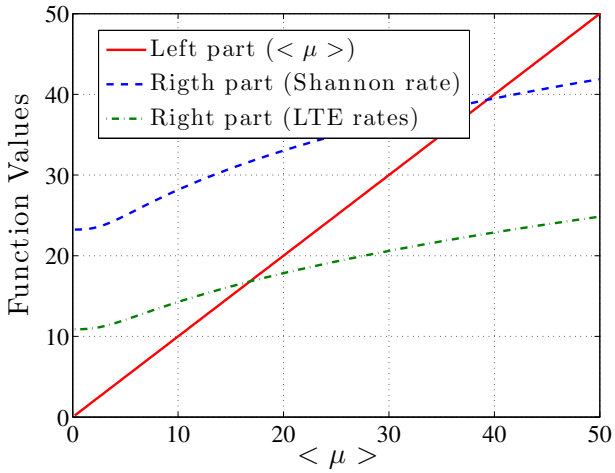


Figure 4.3 – Left and right part of the Eq. (8.4) for the case of LTE SINR thresholds and rates and Shannon's

Fig 4.3 depicts the left and the right part of Eq. (8.4) w.r.t. $\langle\mu\rangle$ for two cases a) the right part of the equation computed assuming LTE rate and b) assuming Shannon's rate. In both cases there is exactly one solution for the Eq. (8.4) and could be approached by simple gradient methods.

The Algorithm 1 presents a simple implementation that solves Eq. (8.4), if the solution exists. We initialize the process by set the service rate $\langle\mu\rangle$ equally to the service rate that we calculate previously for the saturated case (always ON) and based on it we compute the average utility of the BSs. With the given utility and the coverage probability as expressed in Eq. (4.12) we calculate the new MCS distribution and the new service rate. With the new service rate we re-calculate the new utility and following the same pattern iteratively we converge to our solution.

Algorithm 1 $\langle\mu\rangle$ calculator

1: initialize $\langle\mu\rangle_0 \leftarrow$ saturated case & $\langle\mu\rangle_{-1} = 0$ & $\varepsilon > 0$
2: **while** $|\langle\mu\rangle_i - \langle\mu\rangle_{i-1}| > \varepsilon$
3: calculate ρ_i according to $\langle\mu\rangle_i$.
4: calculate $\langle\mu\rangle_{i+1}$ according to Eq. (4.12) with $\rho = \rho_i$.
5: $i \leftarrow i + 1$.
6: **end**
7: **return** $\langle\mu\rangle_i$

Fig. 4.4 depicts the aforementioned algorithm procedure for an LTE network with $\lambda_{BS} = 1$, $\lambda_u = 100$, $\lambda_f = 0.005$ and mean flow size $\langle s\rangle = 5$MBs. It is observable that the iterative procedure converges after few iterations to the desired solution.

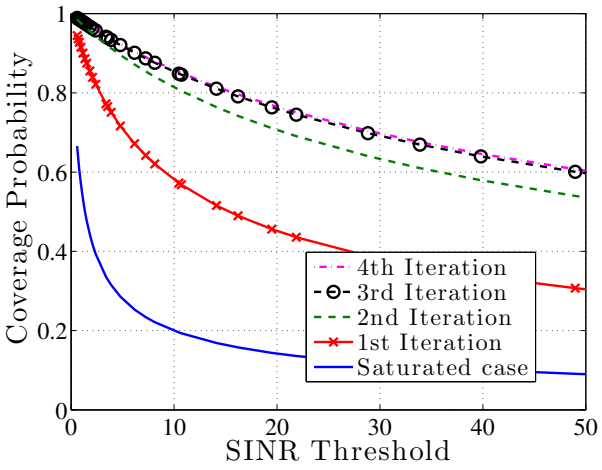


Figure 4.4 – Iterative convergence of coverage probability with respect to SINR using Algorithm 1

4.5.3 Rate for each RAT

Two parameters are missing in order to derive $\langle \mu \rangle$. Firstly, we need SINR thresholds τ_i for each MCS mode to calculate f_{MCS} from Eq. (8.2) and secondly, the corresponding rate of each MCS.

The supported MCS are RAT dependent and are always defined at the standard documents [15], [16]. On the other hand, operation threshold for each MCS is not always defined in the protocol since it depends on the receiver implementation characteristics. For the example that we demonstrate in this paper we will need one SINR-rate table for the LTE modes and one for WiFi. Those tables could be found either at Chapter 3 or at our technical report [60].

4.6 Results

The model parameters, for the rest of the simulation section are summarized as: (i) 5 Mbits average flow size, (ii) pathloss $\alpha = 4$, (iii) thermal noise $\sigma^2 = -100$dBm (iv) $BW_{LTE} = BW_{WiFi} =$ 20MHz, (v) one antenna per eNodeB and one spatial stream per WiFi AP.

Table 8.1 summarizes model parameters of the simulator.

We should mention that if the thermal noise is much smaller than the interference, the value of P_{tx} does not affect the results, as shown in [20].

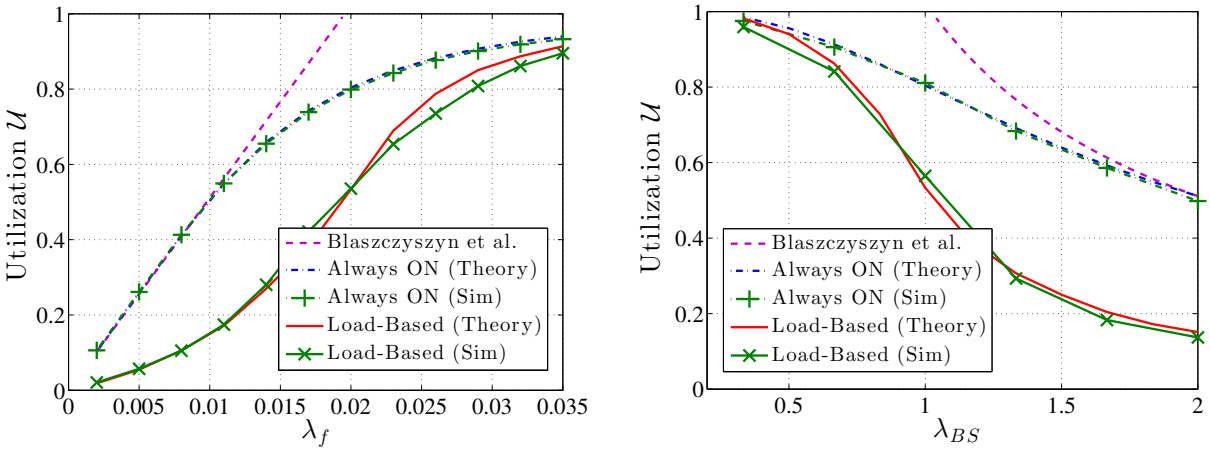
4.6.1 Validation / Performance Analysis

As a first step, we would like to validate our basic theoretical results for a single tier, against simulation results, in both saturated and load-based interference scenarios. Additionally, we compare our method with the analytical approach of [45] that provides closed form results about the network load for the saturated case. We kindly remind the reader that framework [45] assumes that the MCS distribution is somehow known and we fill this gap by using the MCS distribution as calculated in our framework. W.l.o.g an LTE network is considered for this purpose. The performance metrics from the simulated scenarios that are used for the comparison

Table 4.1 – Model Parameters

LTE density	$\lambda_{LTE} = 1$
WiFi density	$\lambda_{WiFi} = 1$
Users density	$\lambda_u = 100$
Flow size distribution	Generic
Mean flow size	5 Mbytes
α	4
BW_{LTE}	20MHz
BW_{WiFi}	20MHz
σ^2	-100dBm
# of antennas per eNodeB	1

are (i) network's utilization[2] and (ii) average flow delay of the median BS. The latter is computed by calculating the mean delay for each BS in the simulation and then taking the median among the BSs. We choose the simulated median rather than the average, as the latter grows to infinity even if a single BS is congested.



a: Load ρ w.r.t flow density λ_f, BS density $\lambda_{BS} = 1$

b: Load ρ w.r.t BS density λ_{BS}, flow density $\lambda_f = 0.02$

Figure 4.5 – Comparison of our theoretical prediction and the packet-level simulator results for both Interfering cases (Always ON and Load-based) for the case of single tier LTE network. With the purple dashed line we present the prediction using Blaszczyszyn's framework

Our packet-level simulator generates BSs and users randomly placed in a large surface with given densities (λ_{BS}, λ_u). Users are associated with the closest BS and generate flows according to a Poisson distribution with density λ_f and average flow size $\langle s \rangle = 5$ Mbits (625 Kbytes). The

[2]It turns out that congestion probability, average delay, interference, etc. depends more on utilization than on load.

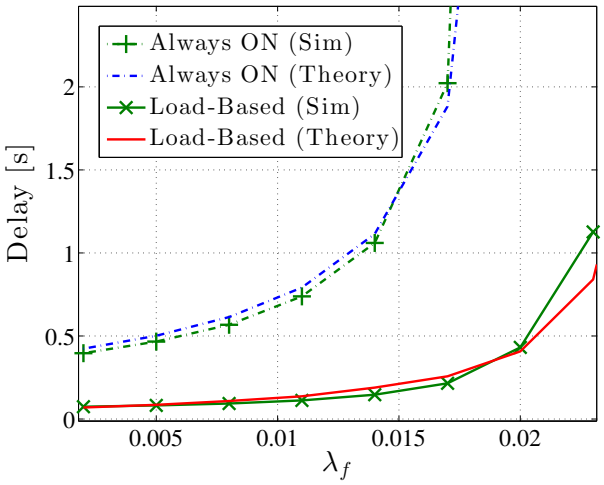


Figure 4.6 – Comparison of our theoretical prediction and the packet-level simulator results for both Interfering cases (Always ON and Load-based) for the case of single tier LTE network. Delay performance with respect to flow density λ_f, BS density $\lambda_{BS} = 1$

flows are forwarded to the corresponding BS which is modeled as a multi-class M/G/1/PS. The service rate of each flow for every time quantum is calculated via the SINR-MCS relation for LTE. We will consider two interference scenarios: (1) always ON case, where all the neighboring BS are contributing to the interference (corresponding to Section 4.5.1), (2) load-based case, where we calculate interference by taking into account only the base stations that are ON at this time quantum (corresponding to Section 4.5.2). We further consider only the users whose SINR is higher than the threshold of the lowest MCS for the always ON case (otherwise a user connected in one quantum might be outside of coverage in the next).

Fig. 8.3 (a) and (b), present the average load ρ (see Eq. (4.3)) of the system w.r.t. λ_f and λ_{BS} respectively, for both scenarios $\lambda_u = 200$. Three general comments from those plots are: (i) The Blaszczyszyn *et al* framework, due to the mean value approximation is able to predict accurately the performance of the system only when it is under-utilized. The prediction error depends on the BS congestion probability, which for low loads is $P(\rho > 1) \approx 0$. Mean value analysis essentially fails due to Jensen's inequality, using the mean load directly, and implicitly averaging some congested BSs (with load $\rho > 1$). Due to the concavity of the function, the utilization turns out to be higher than the average utilization of a cell. For example, in Fig. 8.3 (a) for a $\lambda_f = 0.2$ the prediction error of utilization according to [45] is 20% and our model is 0.5%, roughly 40 times less. While the utilization is a relatively simple metric, it is relatively easy to see that mean value analysis can have an equally important (if not bigger) impact on delay, especially in the case of load-based interference. Clearly, the amount of interference a neighboring BS contributes depends on the percentage of time it is active, i.e., its utilization. Hence, overestimating this utilization will overestimate the neighboring interference and underestimate the respective service rates (of the coupled queues), thus, further failing to predict delays. Consequently, mean value analysis like the one used in [45] could be seen as a first order approximation useful for low loads (and relatively large networks) only. (ii) Both of our theoretical results match the simulation results quite well. (iii) The gap between the always

ON and load-based interference scenarios are extremely high, underlining the importance of the latter.

In Fig. 8.3 (a), for $\lambda_f = 0.02$ the always ON prediction is that the network is 70% loaded instead of 30% of the load-based. That means that the network could be much more robust with respect to data traffic than the studies that assume saturated BSs predict.

In Fig. 8.3 (b), for high density of BS always ON model predicts 50% utilized network, while load-based only 15%. The gap between always ON and load-based prediction increases with respect to density of the network. This happens because saturated analysis is able to capture only the gain coming from the fact that an "arbitrary" BS on average serves less users at a denser network, but not the gain coming from the fact that surroundings BSs will be less loaded, and therefore will cause less interference. Thus, the gain to deploy a denser network is much higher than predicted by an analysis that does not take the load-depended interference into account.

Fig. 8.4, shows the median delay of the packet-level simulator as well as the theoretical predictions for saturated and load-based cases. Again it can be seen that the theoretical predictions are quite accurate, and that always ON interference over-estimates the delay by orders of magnitude.

4.6.2 Different RATs

Given that the validation of the theory worked well, in this section we will use directly the theoretical results, in order to avoid figures being too cluttered. Fig. 4.7 and 4.8 present compactly the performance (congested probability and delay) for the two networks of interest (LTE, WiFi) for the same density of connected users ($\lambda_u = 100$). Taking into account that we have assumed the MAC performance of WiFi equal with LTE (best case, valid for low load or with small modifications as discussed at Section 4.2) all differences between the RATs are due to the PHY characteristics of RATs (different MCS threshold and different rates).

Focusing now on the saturated (Always ON) case, it seems somewhat surprising, at first, that the LTE network performs worse than resource fair WiFi, if we take into account that for the same SINR, LTE tends to operate with higher rate (See also [63]). The reason for this is the edge users. A number of users with low SINR that are regarded as "out of service" and not taken into consideration for the WiFi network, are instead covered by a similar density LTE network. For example, for the always ON case, the coverage area was 0.67 and 0.47 for the LTE and WiFi networks respectively[3]. Hence, the "edge" users in a WiFi BS end up getting better rates than the "edge" users in LTE (This is also evident from Fig. 3.1, where the lowest MCS for WiFi - the lowest green triangle - provides higher rates than the lowest MCSs for LTE - the lowest blue crosses).

The above low values for the coverage area originate from two previous-mentioned worst case assumptions: (i) the random BS placement; in the PPP model, it is possible that a BS ends up asymptotically close to another; (ii) the interference is calculated assuming that neighboring BSs are saturated. By examining the load-based case, we notice how critical the second assumption is, as for low load values the coverage area was almost 1 for both RATs, while for a load around $\rho = 0.5$ coverage areas were 0.9 and 0.7 for LTE and WiFi, respectively.

Another interesting remark is that for low or middle-load scenarios in contradiction to always

[3]This is also the reason why we had to normalize the user densities to ensure that the *absolute* number of connected users per BS is the same for each network.

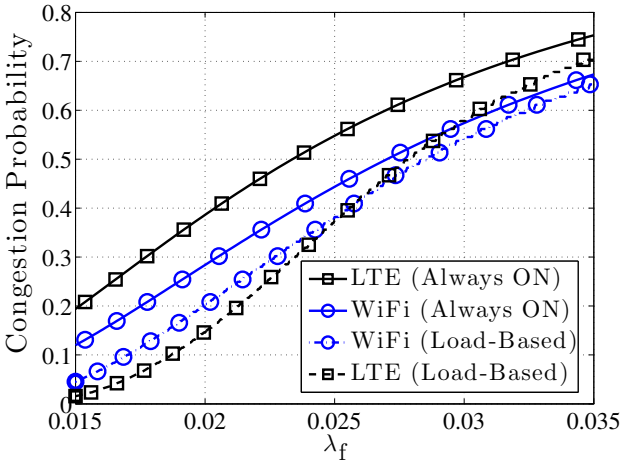


Figure 4.7 – Congestion probability with respect to flow density λ_f, for single tier LTE and WiFi networks for both interfering cases (Always ON and Load-based)

ON case the LTE operates better than WiFi. This happened due to LTE's smaller granularity between the MCS, so, the LTE attain higher SINR improvement. As the load increases the two networks approaching the always ON case which WiFi operates better.

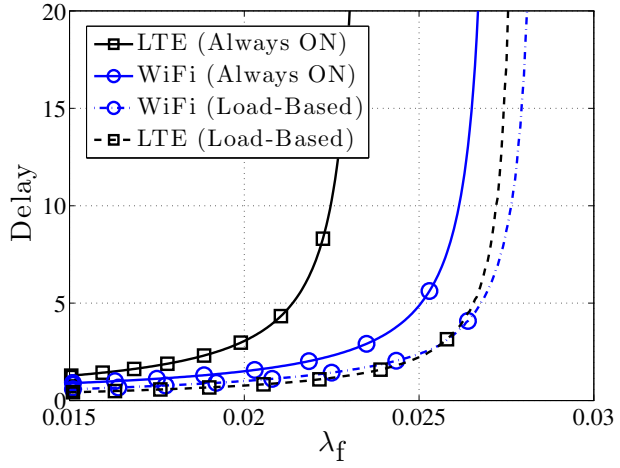


Figure 4.8 – Average user's Delay with respect to flow density λ_f, for single tier LTE and WiFi networks for both interfering cases (Always ON and Load-based)

4.7 Conclusions

We presented an analytical framework to model the flow-level performance of large randomly placed networks assuming saturated BS, as well as a semi-analytical model for the more realistic case of load-based interference. The gap between those two cases could be huge, leading to an

underestimation of the network performance. If the BSs do not interfere all the time, the network is much more robust to the total incoming load and the gain of denser deployment is much higher than the saturated case predicts. Additionally, which network's PHY characteristics are perform better turns out that is load-dependent.

Chapter 5

Energy Efficiency and User's QoE Tradeoff

Energy consumption is one of the primary concerns of modern dense small cell networks. One of the key concepts to improve the energy efficiency of a dense network is to turn off a part of its BSs when they are idle or only lightly loaded, since even then a considerable amount of energy is consumed. In this chapter we use the analytical framework of Chapter 4 to analytically investigate the tradeoff between energy efficiency and user experience, which is measured in terms of users' delay assuming a non-saturated traffic model. We provide an overall network performance with respect to the BS density and insights that can be valuable in terms of network design. Our analysis is based on a) a BS's linear energy consumption model b) stochastic geometry to model the topology of the network and the users and c) queuing theory in order to capture the flow-level performance. Our model is being applied to the popular LTE radio access technology but it can easily be extended to others. Our results provide guidelines and bounds that are able to predict the energy efficiency of the designed network.

5.1 Introduction

One way to increase area capacity in cellular networks is to add more BSs a process also known as densification. Especially the addition of small cells is expected to be one of the key solutions to tackle the exponential increase of traffic on the upcoming years [64–66]. On the other hand, as the data traffic and the density of the networks are increasing, the energy consumption is becoming more and more crucial for both environmental (reducing of the carbon footprint) and economical reasons [67].

Studies have shown that around 50%-70% of the total power consumption of telecommunications is taking place on the BSs [14]. A considerable amount of energy is consumed on the BS (for staying on or cooling) despite serving little or no traffic [13]. Furthermore, if we consider industrial zones, shopping roads, etc. those dense networks are deployed in order to serve the high amount of connected users on rush hours but for the rest of the day the network becomes under-utilized. Therefore, one of the key concepts for energy reduction is to turn off (or put in sleep mode) such base stations.

Depending on the radio access technology used, this can be achieved by various sleeping

techniques [68]. In 3GPP LTE-Advanced (release 10) for example, the carrier aggregation feature can be used to steer traffic to another cell and to power off cells with no traffic. This feature has been improved in release 12 using the discovery reference signal, which is sent by sleeping cells only in configurable intervals [69].

The problem of switching on/off BSs has been investigated in various works. In [40] authors solve the optimization problem of user association taking into account the energy consumption as well as the flow-level performance (delay), but this analysis does not provide analytical results for the overall network performance. In [39] authors take a queuing analytic approach to study the impact of turning off a BS on neighboring ones for different traffic models, but they only do a network wise performance analysis through numerical simulations. In [27] authors follow a stochastic geometry approach in order to provide the network performance while turning off BSs, this work assumes saturated BSs and does not provide impact about the flow-level performance of the network.

Our analysis is based on a common used energy cost model [13] combined with our recently developed framework that analyzes the flow-level performance of a random placed network [70, 71]. We combine tools from stochastic geometry and queuing theory in order to analyze the whole network performance and provide insights about the tradeoff between users' QoE (which is measured in terms of delay) and energy efficiency, without assuming saturated BSs.

We apply our results for the LTE radio access technology and mainly to the case of decreasing the network density by turning off (sleep) a part of the network, but the same framework can analyze the case of increasing the network density by adding BS (densify). In oder to provide close form expressions and to avoid complex system-level simulators we assume that BSs to be turned off are selected randomly, but as we will see this worst case approach is not so far from more sophisticated criteria such as minimum associated users.

Summarizing, our contributions are:

- We derive analytical and semi analytical formulas to study the tradeoff between energy efficiency and users' delay. There are cases where it is possible to have large energy gain and on the other hand affordable reduction of users' performance.

- We compare our assumption were we chose at random the BS that will be turned off with simulations of more sophisticated criteria such as minimum number of associated users and we will see that both approaches are converging as the network density increases.

- In off-peak hours the amount of users is significantly reduced and the network becomes under-utilized. For this scenario, we derive the maximum amount of BS that can be switched off without affecting the performance of the remaining users. Additionally, we provide a simple rule under which conditions this BS reduction leads to energy gain (is possible that the energy consumption increases despite turning off some BSs).

The rest of this chapter is organized as follows Section 5.2 presents our system model, including all of our assumptions on the topology, propagation model, scheduler etc. Section 5.3 derives the MCS distribution for the arbitrary BS. Section 5.4 presents the BS's energy cost model and modifies it to the more useful metric, energy per unit area, the scenario of users' density reduction and some theoretical results are presented as well. Finally, in Section 5.5 we present some interesting results of our analysis about the flow-level performance and the energy efficiency of the network.

5.2 Our Model

We should note that some of our assumptions have already been mentioned in Chapter 4 but we decided to shortly presented them here as well in order to avoid the extend use of cross-references which will make the chapter confusing and hard to read.

5.2.1 PHY Layer Modeling

The first step is to define the assumptions about the topology (both BS and users) as well as the PHY-layer characteristics.

A.1: Both BSs and users follow a homogeneous Poisson Point Processes with densities λ_{BS} and λ_u accordingly. Therefore, the number of BSs (or users) in an area S is

$$P(N = n \mid S) = \frac{(\lambda_{\mathrm{BS}} S)^n e^{-\lambda_{\mathrm{BS}} S}}{n!}, \;\; n = 0, 1, \dots \;\; , \tag{5.1}$$

and their placement is random.

A.2: A standard power loss propagation model is used, usually the path loss exponent is $2 < \alpha < 5$. Additionally, assume a Rayleigh fading channel with mean 1 and constant transmit power of P_{tx}. Thus, the received power at distance d from the BS is given by $P_{\mathrm{rx}} = hd^{-\alpha}$ where h follows an exponential distribution, $h \sim \exp(P_{\mathrm{tx}})$. Hence, the SINR if the user is associated with the i-th BS is given by

$$SINR_i = \frac{P_{\mathrm{rx}_i}}{\sum\limits_{n \neq i} P_{\mathrm{rx}_n} + \sigma^2} \;, \tag{5.2}$$

where σ^2 is the thermal noise. Usually, $\sigma^2_{\mathrm{dBm}} = -174 + 10\log_{10}(BW)$, where BW is the systems bandwidth [61].

A.3: We assume that all BSs have equal transmit power and implement the same scheduling policy. Additionally, we assume that each user gets connected to the closest BS, so, the BSs coverage area could be represented by Voronoi Regions (Tessellations).

Taking into account *A.1* and *A.3* the users' cardinality n for an arbitrary BS, is a *random* variable. Observe that the size of an arbitrary cell is a random variable, depending on the random BS topology, and the number of users given a specific cell size is also a random variable. The following lemma provides the probability mass function $f_N(n)$, of users cardinality on an arbitrary cell. The proof of it as well as a useful and accurate approximation could be found at our technical Appendix A or [62, 63].

Consider BSs distributed in 2D as a homogeneous PPP with density λ_{BS}, and offering coverage to a set of users distributed as another PPP with density λ_{u}, (*A.1*). Assume further that user association within this tier is done by using the closest-distance rule, (*A.3*). Then, the probability of having exactly n users in an arbitrary cell, $f_N(n)$, is given by:

$$f_N(n) = \frac{343}{n!15}\sqrt{\frac{7}{2\pi}}\frac{\zeta^n}{(\zeta + \frac{7}{2})^{n+\frac{7}{2}}}\Gamma(n + \frac{7}{2}) \;, \tag{5.3}$$

where $\zeta = \frac{\lambda_{\mathrm{u}}}{\lambda_{\mathrm{BS}}}$ and Γ is the gamma distribution.

A.4: In this chapter we assume that the part of the network that will be turned off is selected randomly. The random selection can serve as a worst case scenario (assuming that we do not

exploit our knowledge of the network in order to intentionally decrease its performance). There are more sophisticated criteria that can decide which BSs to turn off in order to improve the energy performance of the network (minimum amount of connected users, minimum providing service rate, etc.) but those criteria, in most cases cannot be modeled mathematically.

If we turn off randomly the 10% of the initial BSs, the remaining are again a Poisson process with density BS, $\lambda_{BS}^{'} = 0.9\lambda_{BS}$ due to the following lemma.

Lemma 5.2.1. Let a Poisson process with rate λ if we divide it randomly with probability p and $(1-p)$ to two processes, Then, the two outcome processes are again Poisson with new rates $\lambda_1^{'} = p\lambda$ and $\lambda_2^{'} = (p-1)\lambda$ respectively due to Poisson thinning property.

5.2.2 BS level Modeling

We assume that each BS experiences a *dynamic* traffic load and we would like to study the performance at *flow-level*. We now state our assumptions regarding a single randomly chosen BS, and comment where necessary.

A.5: Each *connected* user to a BS generates new *flow* requests randomly, and independently of other users, according to a Poisson Process with density λ_f.

A.6: A flow is a sequence of packets corresponding to the same user or application request (e.g., a file or web page download). Each flow has a random size, in terms of bits, drawn from a *generic* distribution with mean value $\langle s \rangle$.

We use the Lemma 4.2.1 in this chapters as well where If n users are associated with a given BS, the aggregate flow arrival process to that BS is Poisson($n\lambda_f$).

A.7: In the absence of other flows, *a single flow will be served at full rate*, with the maximum MCS that the BS can offer to that user, which in turns depends on the SINR-BLER specifications for that RAT. In this work we examine an LTE network, so, according to Chapter 3 or [63] the SINR thresholds and the corresponding rate for each MCS are depicted in Table 5.1.

Table 5.1 – LTE's SINR threshold (τ) in dB and the corresponding rate (MB/s) w.r.t. MCS index, for the case of 20MHz bandwidth and acceptable BLER 10^{-1}

#	τ	rate	#	τ	rate	#	τ	rate
0	-2.3	2.8	9	3.8	15.8	18	10.3	32.9
1	-1.6	3.6	10	5.3	16.0	19	11.5	36.7
2	-1.0	4.6	11	5.5	17.6	20	12.1	39.2
3	-0.2	5.7	12	5.9	19.8	21	12.9	43.8
4	0.6	7.2	13	6.8	22.9	22	13.4	46.9
5	1.3	8.8	14	7.9	25.5	23	14.6	51.0
6	1.8	10.3	15	8.6	28.3	24	15.3	55.1
7	2.6	12.2	16	9.1	30.6	25	16.0	57.3
8	3.2	14.1	17	10.2	30.8	26	16.9	61.7

The MCS probability mass function $f_{\text{MCS}}(mcs)$ is derived in Section 5.3.

We will assume a SISO system and a single carrier in our analysis [49]. Increased rates due to spatial multiplexing and carrier aggregation could be included with a proper physical abstraction models.

5.2.3 Queueing Model for BS Schedulers

When more than one flows are served in parallel by a BS, the BS operates as a *queueing system*. The service rate for a flow is generally smaller than what assumption (A.6) predicts. It actually depends on the number of active flows (BS load), and the BS scheduler or media access control (MAC) mechanisms which decide how the available resources will be distributed between flows.

Resource Fair Scheduler

The basic characteristics of resource fair scheduler as we have already mentioned in Chapter 4 such as *service rate*, average *user's delay* and average network's *load* can be expressed analytically according to Eq. (4.1), Eq. (4.2) and Eq. (4.3) respectively.

5.3 MCS Distribution

5.3.1 Coverage Probability / MCS distribution

In this chapter we are interested only in the case of load based interference. That being said, the coverage probability of this network, as well as the methodology in order to calculate user's MCS distribution discussed in Section 4.5.2. Worth to be mentioned again, that as we can see in Eq. (4.12) of Lemma 4.5.1 opposite to the always ON case the coverage probability as well as the user's MCS distribution are strongly depended on the ratio between BS's and users density $\zeta = \lambda_u/\lambda_{BS}$ and this dependence is not linear.

Once the MCS distribution is depended on the ratio ζ it resulted that the service rate $\langle\mu\rangle$ is depended as well. So, now if we re-write Eq. (4.3) taking into account the above dependance

$$\rho = \frac{\zeta \cdot \lambda_f}{\langle \mu_{(\zeta)} \rangle} \, , \tag{5.4}$$

It is quite obvious that in load based case, the dependance of network load with ratio ζ is not linear as well.

5.4 Energy Cost Model

The linear cost model is the most common energy consumption model [13]. The model consists of: a) a constant term that captures energy consumption of the BS in order to be ON and ready to operate and b) a linear term that is responsible for the energy consumption while the BS is operating (exchanging data). We set as E_{on} the amount of energy that the BS consumes in order to be ON for a period T. Additionally, E_{op} is the energy consumption while the BS is operating with a user and Δt is the amount of time that the BS is operating ($\Delta t \leqslant T$). Hence, the total energy consumption of the BS for period T is given by

$$E_{BS} = E_{on} + E_{op}\frac{\Delta t}{T} \, . \tag{5.5}$$

We are not interested in the energy consumption of a specific BS but for the total energy consumption per unit area. After some trivial calculations and expressing $\frac{\Delta t}{T}$ as network load ρ we end up with

$$\overline{E} = \lambda_{BS} \left[E_{on} + E_{op} \cdot \rho_{(\zeta)} \right] \ , \tag{5.6}$$

where $\rho_{(\zeta)}$ is the average utility of the network Eq. (5.4) and $\zeta = \frac{\lambda_u}{\lambda_{BS}}$.

Partial Reduction of Users Density (Off-Peak Hours)

A network is deployed in order to achieve a specific user performance. Some areas suffer for high users' variability in a day, shopping streets, industrial zones, etc. The network design aims to achieve a predefined users' performance even at the high traffic period. That means that the network is under-utilized on the low traffic period. The operator's goal is to turn off a part of the network in order to save energy, but without decreasing the performance of the remaining users.

Lemma 5.4.1. If we assume that for any reason the density of the users in a given area decreased according to a factor l_u the maximal decrease factor (l_{BS}) of the BS density but without affecting the average delay performance of the remaining users is

$$l_{bs} = l_u \ . \tag{5.7}$$

Proof. The users delay before and after the users and BSs reduction should be the same, so, we can write

$$Delay_1 = Delay_2 \ ,$$

where according to processor sharing scheduler we re-write

$$\frac{1}{\langle \mu \rangle_1 - \lambda_1} = \frac{1}{\langle \mu \rangle_2 - \lambda_2}$$
$$\langle \mu \rangle_1 - \lambda_1 = \langle \mu \rangle_2 - \lambda_2 \ ,$$

were the service rate $\langle \mu \rangle$ is a function of the densities ratio $\zeta = \frac{\lambda_u}{\lambda_{BS}}$. Assuming that the other network parameters constant we have

$$\left\langle \mu_{(\zeta)} \right\rangle - \zeta \cdot \lambda_f = \left\langle \mu_{(\frac{l_u}{l_{bs}}\zeta)} \right\rangle - \frac{l_u}{l_{bs}} \zeta \cdot \lambda_f \ . \tag{5.8}$$

Further, we define the function $g(\cdot)$

$$g(x) = \left\langle \mu_{(x)} \right\rangle - x \cdot \lambda_f \ . \tag{5.9}$$

Both $\left\langle \mu_{(x)} \right\rangle$ and $-x \cdot \lambda_f$ are strictly decreasing functions with respect to x. So $g(x)$ is a strictly decreasing function as well. If $g(\cdot)$ is a strictly monotonic function and $g(a) = g(b)$ then $a = b$. Thus, applying that to Eq. (5.8) we have

$$\zeta = \frac{l_u}{l_{bs}} \zeta \ . \tag{5.10}$$

■

Unfortunately, switching off some BS does not lead directly to energy improvement. When we switch off a part of the network, on the one hand we save some energy by E_{on} factor of the BSs that we turned off, but on the other hand, the network in total consumes higher amount of operational energy E_{op} than before, because the remaining BSs serve the users with worse channel conditions (therefore, for more time) than before.

Lemma 5.4.2. The BS reduction according to the reduction of users density that was presented in Lemma 5.4.1, surely reduces the energy consumption of the network if

$$\frac{E_{on}}{E_{op}} > \frac{1}{1 - l_{bs}} \, . \tag{5.11}$$

Proof. Regarding the energy consumption of this scenario the answer is not trivial. We aim the energy consumption of the thinned $\overline{E}_{th}$ network to be less than the consumption of the full network $\overline{E}_f$

$$\overline{E}_{th} < \overline{E}_f \, . \tag{5.12}$$

Assuming that for a portion of time p_t the users density has been reduced by a factor l_u, thus, according to Lemma 5.4.1 the BS's density will reduced by a factor $l_{BS} = l_u$ and by applying to the Eq. (5.6)

$$p_t l_{BS} \lambda_{BS} [E_{on} + \rho_{th} E_{op}] < p_t \lambda_{BS} [E_{on} + \rho_f E_{op}] \, , \tag{5.13}$$

after some trivial calculations we derive that in order to have some energy gain through the thinning of BS the following inequality should hold

$$\frac{E_{on}}{E_{op}} > \frac{l_{bs}\rho_{th} - \rho_f}{1 - l_{bs}} \, . \tag{5.14}$$

taking into account that the nominator of Eq. (5.14) is upper bounded by $l_{BS}\rho_{th} - \rho_f < 1$, we end up to the asymptotic rule of thumb of Eq. (5.11) ■

If Eq. (5.11) does not hold, we should calculate the load for each case by following the whole procedure of Section 4.5.2 and Eq. (4.3) combined with Eq. (5.6) in order to decide if the energy consumption of the network will be decreased by turning off a part of the network or not.

5.5 Results

5.5.1 Validation

The packet-level simulator generates both BSs and users randomly placed in a large surface with given densities (λ_{BS}, λ_u). Users are associated with the closest BS and generate flows according to Poisson distribution with density λ_f. The flows are forwarded to the corresponding BS which is modeled as a multi-class M/G/1/PS. The service rate of each flow for every time quantum is calculated via SINR. At the calculation of the interference we are taking into consideration only the base stations that are ON at this time quantum (load based case), for comparison with the most common assumption in the literature we also implement the case of taking into consideration, at the interference calculation, all the BS (saturated case). In order to compare

fairly both interference cases we consider only the users whose SINR is at least higher than the threshold of the lowest MCS at saturated case. We should mention that the packet-level simulator needs extremely high computational resources, so, the theoretical prediction is even more valuable.

We are interested in investigating the scalable performance of a large, random placed network while we turn off a percentage of BSs in order to improve the energy efficiency. In this section the *user's density does not change*, so by turning off BSs we know a priori that users' performance will be reduced, so we want to compare the energy gain with the performance reduction. In our theoretical analysis we select randomly which BSs will be turned off. As we have mentioned before, the random selection is a worst case scenario, there are more sophisticated criteria in order to decide which BSs to turn off (minimum associated users).

Figs. 8.5 and 5.2 shows how the flow level performance of the network (average flow delay and the average network load) scales with respect to the density of the BS (λ_{BS}) in four different cases i) our theoretical prediction that assumes load based interference (l-b) and the selection of which BSs to turn off is random (rnd) ii) simulation results for the case of load based interference and the selection of which BSs to turn off is random iii) simulation results assuming load based interference and the selection of which BSs to turn according to minimum associated users (min) criterion and iv) simulation results assuming saturated (sat) BSs and the selection of which BSs to turn off is made randomly. We can obtain three interesting conclusions from Figs 8.5 and 5.2: 1) our theoretical prediction is very accurate (for both load and delay) compared with the simulation results, 2) the assumption of saturated BSs (which is common at stochastic geometry works) changes totally the network's performance not only quantitatively but qualitatively as well and 3) the minimum associated users criterion does not differ a lot from the random one, especially when the network is very dense the difference is negligible. That means that for dense networks it is better to randomly switch off BS than using the centralized and more complex minimum associated users criterion to determine which BS to switch off.

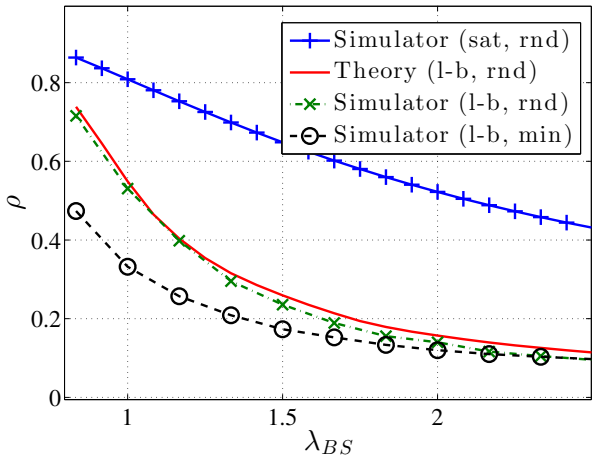


Figure 5.1 – Theoretical and simulation results for the average network load. *sat* indicates the saturated (Always ON) and *l-b* the load based cases of the interference. *rnd* indicates the random selection and *min* the minimum associated users criterion of turning off BS

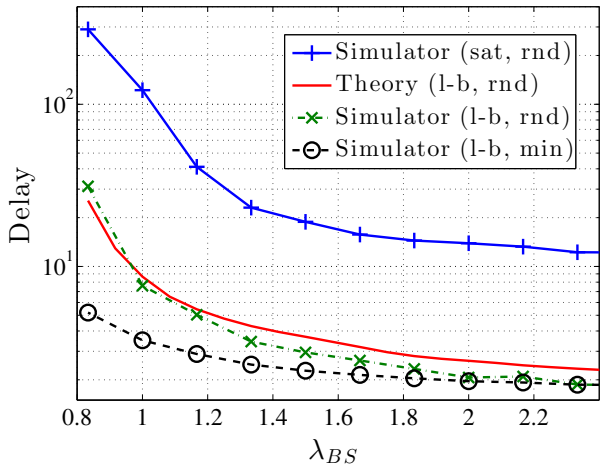


Figure 5.2 – Theoretical and simulation results for the average user's delay. *sat* indicates the saturated (Always ON) and *l-b* the load based cases of the interference. *rnd* indicates the random selection and *min* the minimum associated users criterion of turning off BS

5.5.2 Energy Vs Delay

In this subsection we are interested in investigating the tradeoff between energy efficiency and users delay. Regarding the interference we assume the more realistic case of load based. Initially, the network is under-utilized ($\rho \approx 0.1$), let E_0 and D_0 be the energy consumption and the average delay of this initial state. Then gradually we turn off a part of the network, we define as $\overline{E}/E_0$ the relative energy gain and as $\overline{D}/D_0$ the relative delay, for simplicity we will call those metrics energy gain and relative delay respectively.

Fig. 8.7 and 8.8 shows the energy gain w.r.t. relative delay for different $\frac{E_{on}}{E_{op}}$ ratios. Initially, by observing the case of $E_{on} = E_{op}$ we note that for low load the derivative of the energy gain is very high, so there is the capability of energy improvement without large delay cost. When the load of the network is $\rho \approx 0.5$ the derivative decreased dramatically, thus the delay cost is extremely high compared to the energy gain Fig. 8.7.

Furthermore, in Fig. 8.8 there are two more remarks

1. As the ratio between constant and operational energy becomes higher, the possible gain by turning off the BS is increasing. Considering the traditional small cells, the case that the constant energy term be much higher than the operational one seems unrealistic, but in future networks (e.g. drones) this could be the case

2. When the constant energy term is less than the operating one, there is a turning point in the performance curve. This means that after a point, as we turn off more BS, both the energy as well as the delay performance are getting worse. For the extreme case where the particular cost of a BS to be ON is negligible compared to the operational cost $\frac{E_{on}}{E_{op}} \to 0$ there is no capability for energy improvement, thus, the optimal strategy is simply to set all BSs ON Fig. 8.9.

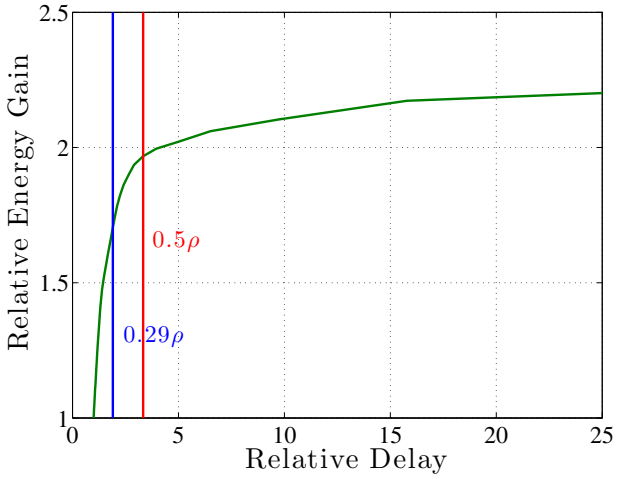


Figure 5.3 – Relative energy gain and relative delay for the case of $E_{on} = E_{op}$. The two vertical lines indicates the in which point the network load is ρ is equal with 0.29 and 0.5 respectively

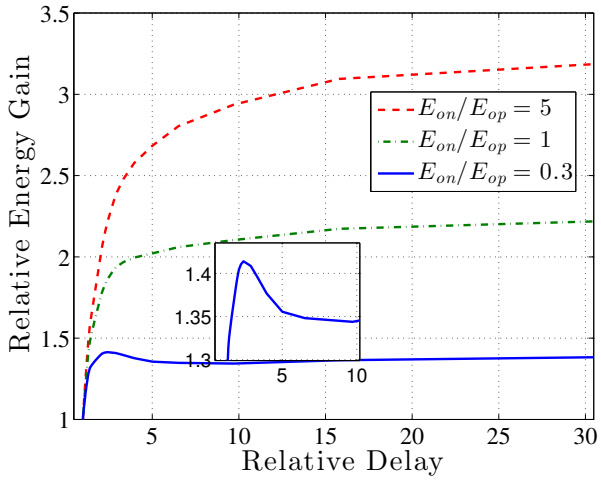


Figure 5.4 – Relative energy gain and relative delay for different $\frac{E_{on}}{E_{op}}$ ratios

One other interesting metric is the absolute energy consumption divided by the average user's delay $\frac{\overline{E}}{Delay}$ w.r.t. BS density λ_{BS}. Fig. 8.10 depicts the aforementioned metric for different values of the ratio $\frac{E_{on}}{E_{op}}$. If roughly $\frac{E_{on}}{E_{op}} > 0.1$ the ratio $\frac{\overline{E}}{Delay}$ scales linearly w.r.t. BS density λ_{BS}. This property could be a valuable empirical rule of thumb for network design if we consider that both $\overline{E}$ and $Delay$ are much more complex w.r.t. λ_{BS}.

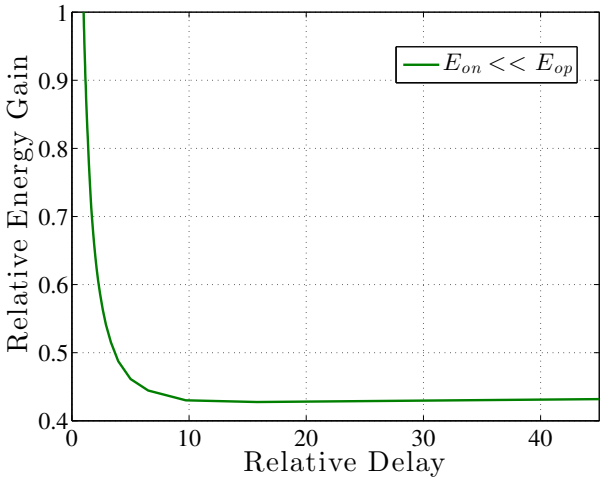


Figure 5.5 – Relative energy gain and relative delay for the case of $\frac{E_{on}}{E_{op}} \to 0$

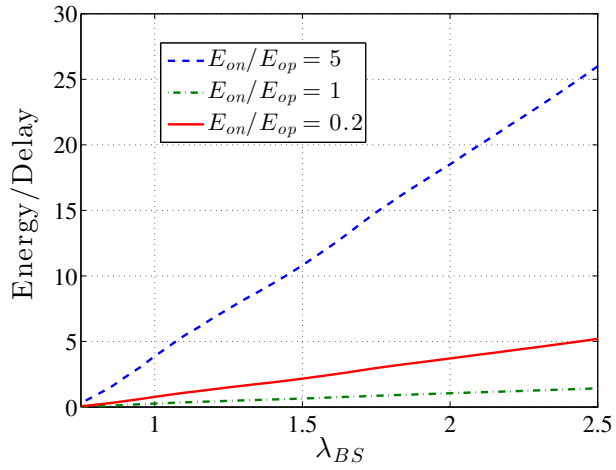
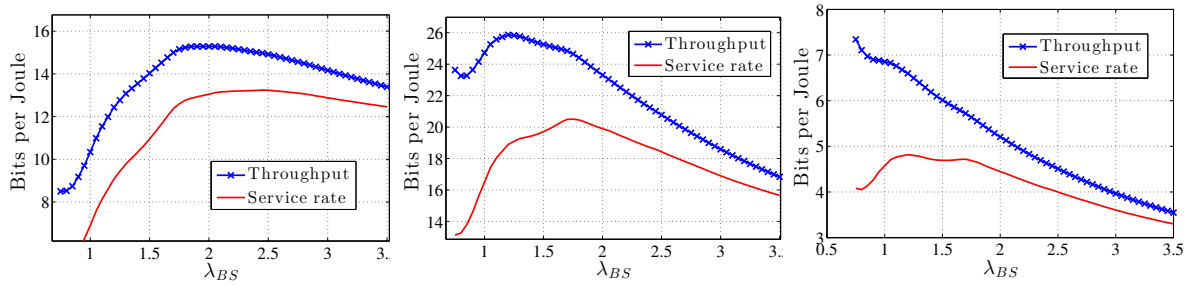


Figure 5.6 – Ratio of average energy consumption and average user's delay $\frac{\overline{E}}{Delay}$ with respect to BS density λBS

5.5.3 Bits per Joule

Another metric of networks energy efficiency is the amount of transmitting bits per joule $\overline{R}/\overline{E}$, where $\overline{R}$ usually represents the throughput of the the system $\sum f_{MCS}(mcs) \cdot r(mcs)$. Regarding the flow-level performance of the network the service rate $\langle mu \rangle = (\sum_{mcs} \frac{f_{MCS}(mcs)}{r(mcs)})^{-1}$ is more representative. So, respectively with the bits per joule we define the flows per joule metric of the system. Fig. 8.11 depicts bits per joule w.r.t. the network's density for different values of the ratio $\frac{E_{on}}{E_{op}}$. When $E_{on} \not\gg E_{op}$ there is an optimal network density that maximizes bits per joule and a different desity that maximizes the flows per joule. In terms of flows per joule the



a: $\frac{E_{on}}{E_{op}} = 0.2$ b: $\frac{E_{on}}{E_{op}} = 1$ c: $\frac{E_{on}}{E_{op}} = 5$

Figure 5.7 – Bits per Joule for different values of ratio $\frac{E_{on}}{E_{op}}$. *Throughout* indicates the statistical mean of the rate, *Service rate* shows the harmonic mean of the rate as this calculated in Eq. (4.1)

optimal density is higher than in bits per joule case.

5.6 Conclusions

In this Chapter we presented an analytical framework that provides the energy performance of the network and we investigated the tradeoffs between network's energy efficiency and user's QoE. We observe that the worst case assumption of random selection of which BS turn off the network performance is not much worse than the more complex and sophisticated solution of turning off the BS with the fewer amount of associated users if the deployed network is dense. We present some theoretical results about the reduction of BS density in the case of users reduction and we provide a rule of thumb, when this reduction will lead to energy gain. Additionally, if the amount of users will not be reduced, we saw that there is capability of energy improvement without affecting a lot user's QoE when the operational energy in not much greater than the constant energy term. Finally, we present some interesting metrics about the energy efficiency.

Chapter 6

Performance Analysis of Multi-tier Heterogeneous Networks

Modern cellular networks are becoming denser, less regularly planned, and increasingly heterogeneous, as a result of operators' efforts to deal with an unprecedented data crunch. This increased complexity however makes performance analysis challenging. In this chapter we use the results of Chapter 4 about the performance of a single tier network and we develop a flexible and accurate model in order to analyze the performance of large heterogeneous cellular networks (HetNets), and understand the impact of key network parameters. This model consists of K tiers of randomly located Base Stations (BSs), with different densities, transmit powers and Radio Access Technologies (RATs). Our main goal is to understand the impact of flow level dynamics on such a system for different inter-tier tier association criteria and assuming non-saturated users that randomly generate download requests ("flows"). Finally, we apply our model to the case of a popular 2-tier HetNet, based on LTE and WiFi, in order to understand the performance differences of popular user tier-association criteria, such as Off-load, Max-SINR association, and Min-Delay association. Our results provide some interesting qualitative and quantitative insights about the impact of these association policies and different traffic intensities.

6.1 Introduction

As mentioned in the thesis introduction, HetNets is one of the most promising solutions in order to tackle the exponentially increase of the mobile data traffic. To alleviate the overloaded macro-cell network, operators are additionally deploying small cells to capture traffic in hot spots. One promising scenario for such HetNet is the combination of LTE macro cells with WiFi small cells. Nowadays, it is possible to integrate WiFi access points into the core network of cellular systems, and perform off-loading of traffic from LTE to WiFi.

HetNet architectures offer numerous advantages, but they also lead to denser, irregular, and more heterogeneous deployments, due to the often unplanned and incremental deployment of new (small cell) BSs [72], as well as the potentially different Radio Access Technologies (RAT). As a result, analyzing such networks, e.g., for protocol comparison or network planning, becomes increasingly challenging. Taking into account that the usually considered metrics in such analyses, like SINR or capacity, often fail to capture the actual user experience, because

they do not take into account the heavy load of modern cellular networks [9,37], as we did in the previous chapters as well, we focus on other metrics such as the use delay and the congestion probability.

To this end, in this chapter we develop a flexible and accurate model for the performance of future HetNets, in order to understand the impact of important network parameters. Our model consists of K orthogonal tiers of randomly located BSs, with different densities, transmit powers and RATs, as well as randomly placed users. As per previous chapters users are assumed to be non-saturated, randomly generating requests for new file/flow downloads of varying sizes, and they perceive performance in terms of the average delay to finish such a download. In other words, we are interested in the flow-level dynamics or flow-level performance of this heterogenous network. Following our work so far we model the BSs as queueing systems, that schedule concurrently arriving user flows according to the respective RAT scheduler, and network-related performance is measured in terms of the stationary load imposed on each BS, and the probability (or percentage) of BS being congested.

Starting from our framework presented in chapter 4 that analyzes the flow-level performance of a single tier network, we extended it for the case of multiple tiers. Our goal is to provide an analytical framework that analyzes the flow level dynamics in large, random placed, multi-tier heterogeneous networks and to study the impact of different association criteria. Therefore, in this chapter we use our analytical framework to study the impact of popular user association policies like *Off-load* (all users within range of a WiFi AP are associated to the WiFi network), *Max-SINR* (a user is associated with the BS offers the best SINR, among any tier), and *Min-Delay* (a user is associated with the tier which offering the best combination of throughput and load in order to minimize the average delay [46]). Our results provide some interesting qualitative and quantitative insights.

The rest of the chapter is organized as follows. In Section 6.2 we present our model for performance at the BS level together with our PHY Layer model. In Section 6.3 we mathematically model the association rules, and we present the corresponding MCS distributions. Section 6.4 considers some scenarios of interest and applies our analytical results to obtain insights. Section 6.5 presents the future steps of our work.

6.2 Our Model

Most of our assumptions according the flow generation, the channel model, the topology or the intra-tier association are the same as presented in Chapter 4. For compactness we will present them here as briefly as possible together with our new assumptions for the multi-tier environment .

6.2.1 Performance at the BS level

We assume that each BS experiences a *dynamic* traffic load and we would like to study the performance at *flow-level*. Our assumptions regarding a single randomly chosen BS are

A.1: Each *connected* user to a BS generates new *flow* requests randomly, and independently of other users, according to a Poisson Process with density λ_f.

A.2: A flow is a sequence of packets corresponding to the same user or application request (e.g., a file or web page download). Each flow has a random size, in terms of bits, drawn from

a *generic* distribution with mean value $\langle s \rangle$.

A.3: The number of users n associated with a BS is a *random* variable with pmf $f_N(n)$ that depends on the density of the BSs, the density of users, and the association criteria. The pmf represented in Eq (5.3).

A.4: In the absence of other flows, *a single flow will be served at full rate*, with the maximum Modulation and Coding Scheme (MCS) that the BS can offer to that UE, which in turns depends on the SINR-BLER specifications for that RAT. The rate of the arbitrary user could be assumed as a random variable and the corresponding pmf, $f_{\mathrm{R}}(r)$.

6.2.2 Queueing Model for BS Schedulers

We model each BS scheduler with the proper queue model. As discussed in details in Chapter 3 for the two RATs of our interest the proper Queue models are:

1) For LTE scheduler can be modeled as a proportional fair, multi-class M/G/1 Processor Sharing queue. 2) regarding the WiFi scheduler we assume two different approaches i) *Approximation 1:* resource fair scheduler, this approximation is valid when the BS load is low and can be used as a lower bound on the delay, for higher load values. ii) *Approximation 2:* For general loads, we can model the wifi scheduler as throughput fair system and to use the approximation of Avrachenkov et all. For further details you can look back on Chapter 3, Section 8.3.2.

6.2.3 PHY Layer Modeling

Before we proceed with the derivation of the cardinality and rate probability distributions, we state here our assumptions about the network topology and physical layer model.

A.5: Users are distributed according to an independent Poisson Point Process with density λ_u.

A.6: We assume a network with k independent tiers of BSs. The number of BSs of tier j inside an area S follows a h-PPP, Φ_{BS_j}, with density λ_{BS_j}, and independently of other tiers. Hence, the number of BSs of tier j in an are S

$$P(N_j = n \mid S) = \frac{(\lambda_{\mathrm{BS}_j} S)^n e^{-\lambda_{\mathrm{BS}_j} S}}{n!}, \quad n = 0, 1, \dots \quad . \tag{6.1}$$

W.l.o.g., we assume that each tier operates at different frequency. Thus, tiers are orthogonal, i.e., interference at each BS originates only from BSs of the same tier. In the case of two tiers sharing the same frequency, we could model this as a single tier with possibly different transmission rates (e.g., for small cells and macro-cells).

A.7: A standard power loss propagation model is used Rayleigh fading (see Eq. (4.5))

A.8: We assume that all BS of the same tier have equal transmit power, but transmit power might differ between tiers. Similarly, all BSs at the same tier implement the same scheduling policy (either Resource Fair or Throughput Fair), but different tiers might have different policies.

A.9: When more than one tiers exist in the network, a user association policy decides which tier a given user will be sent to. We consider the following association policies:

- *Off-loading:* In the simplest scenario, the operator might steer to a preferred tier (e.g. WiFi or 4G) all users that can connect (i.e., receive a sufficiently high SINR) to this tier. Such offloading can be achieved by broadcasted *Absolute Priorities* to all nodes

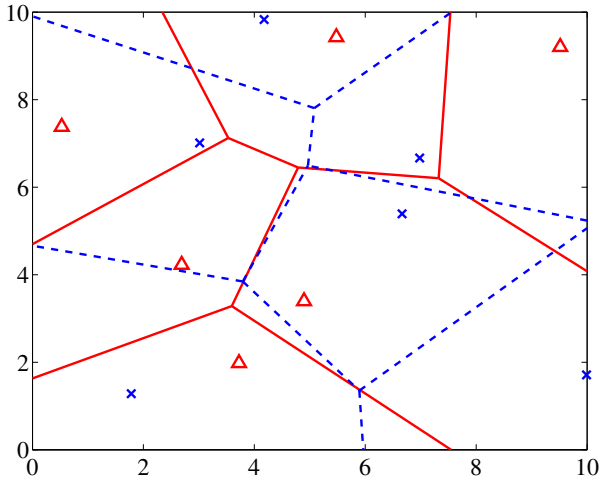


Figure 6.1 – Voronoi Tessellation example, 2-tiers with BS density $\frac{6}{100m^2}$ each

in RRC_IDLE state or dynamically through *Dedicated Priorities* indicated to nodes in RRC_CONNECTED state [49, 73].

- *Max-SINR:* A user choses to associate with the tier that provides the best SINR.
- *Min-Delay Association:* The load of each tier is also taken into account when associating, in order to minimize the average delay of the system. While a number of load-based association algorithms have been considered, here we assume a simple version of the association rule proposed in [46].

Within a given tier, we assume the user association criterion is maximum SINR, which is standard. A number of recent works [46, 74] have shown that this criterion is sub-optimal and more sophisticated criteria (e.g., load-based, as in the case of inter-tier association) could be applied for intra-tier association, in order to improve performance. Nevertheless, these results are equally applicable to every tier, and our focus in this paper is the relative impact of using multiple tiers, rather than the optimal performance of each tier itself.

Assuming all of the above and additionally, that on average, the received power is monotonic in respect to distance, our criterion is simplified to the closest distance criterion, so, the BSs's coverage areas could be represented by Voronoi Regions (Tessellations), Fig. 8.12 shows two orthogonal networks and their voronoi regions, solid lines correspond to tessellations in respect to $\bigtriangleup$ network and dash lines to $\times$ network.

6.3 MCS Distribution for each Association Criterion

As we saw in Chapter 4 the pmf distribution of each MCS $f_{MCS}(mcs)$ derived through the coverage probability p_c. Coverage probability denotes the probability that the SINR of a randomly located user grater than a given threshold (in our case, is SINR threshold for each MCS). We define this coverage probability in two different cases i) Always ON interference, where we

assumed that interfering BS always transmitting ii) Load Base Interference, were the BS are interfering only for the amount of time that serves a user.

Now using the MCS distribution $f_{MCS}(mcs)$ and the coverage probability p_c of each tier, we want to study different association criteria and to derive the new MCS distribution and the percent of the associated users to each of them.

6.3.1 Multi-tier Association

In the case of a multi-tier network, the user density λ_u and pmf of MCS $f_{MCS}(mcs)$ of each tier also depend on the inter-tier association policy. We present here how to calculate those two parameters for the basics association schemes considered. We should clarify that the association rules below denote the tier that the user will be associated with and not the BS. Given the tier, the user is associated with the closest BS of it. Let's assume there are two tiers, with $f^i_{\text{MCS}}(mcs)$ and $p^i_c\left(T, \lambda_{BS}, \alpha\right)$ be the MCS distribution and the coverage probability of each tier $i = \{1, 2\}$.

Lemma 6.3.1 (Offload). For the Offload case, the user is associated with tier-2, if the achieved SINR for that tier is higher than a coverage threshold τ_0. Then, the pmf of MCS and the density of users for tier-1 are given by

$$p_1^{'}(mcs) = f^1_{\text{MCS}}(mcs)\left(1 - p_c^2\left(\tau_0, \lambda_{BS}, \alpha\right)\right) \tag{6.2}$$

$$\lambda_u^1 = \lambda_u\left(1 - p_c^2\left(\tau_0, \lambda_{BS}, \alpha\right)\right) \ . \tag{6.3}$$

The term $\left(1 - p_c^2\left(\tau_0, \lambda_{BS}, \alpha\right)\right)$ denotes the non-coverage probability, meaning the probability that the user's SINR in tier 2 is less than the threshold τ_0. Due to the tiers' orthogonality, the probability of achieving MCS_i at one tier and the non-coverage probability at the second are independent. Finally, due to Poisson thinning, the density of first tier is the initial density λ_u thinned by the non-coverage probability at tier two.

Lemma 6.3.2 (Max SINR). For the Max-SINR case, the user is associated with the tier providing the maximum SINR. Thus, pmf of the MCS and the corresponding density of users for tier-1 are as follows

$$p_1^{'}(mcs) = \int_{\tau_i}^{\tau_{i+1}} p_c^1(\tau)\left(1 - p_c^2\left(\tau_0, \lambda_{BS}, \alpha\right)\right) d\tau \tag{6.4}$$

$$\lambda_u^1 = \lambda_u \int_0^{\infty} p_c^1(\tau)\left(1 - p_c^2\left(\tau_0, \lambda_{BS}, \alpha\right)\right) d\tau \ . \tag{6.5}$$

Similar for tier-2.

The above tier-association rules are the two most common ones considered. However, these rules purely depend on coverage statistics and ignore the load factor which plays an equally, if not more, important role on performance. We therefore propose a third tier-association rule that takes this load into account. This rule is slightly more involved, and is a modified version for multi-tiers of the iterative distributed association algorithm of [46] (originally proposed for a single tier), where we have used $\alpha = 2$ in the α-function that leads to minimized delay.

Lemma 6.3.3 (Min Delay). Let us assume that the arbitrary user could operate with any of N different MCS at tier 1, according to $f^1_{\text{MCS}}(mcs)$, and M different MCS at tier 2, according to $f^2_{\text{MCS}}(mcs)$. We can form a new set $\mathcal{L}$ of $N \times M$ values of the combinations of the two initial pmfs. For example, users of "type" $(2,3) \in \mathcal{L}$ corresponds to the sub-set of users who can receive MCS 2 from tier 1 and MCS 3 from tier 2. Each of those $N \times M$ possible states of the arbitrary user will associate to the tier i according to the following criterion:

$$i(x) = \underset{j \in \mathcal{B}}{\text{argmax}} \; c_j \left(1 - \mathcal{U}_j\right)^2 \; , \forall x \in \mathcal{L} \; , \tag{6.6}$$

where $\mathcal{B}$ is the set of all tiers (in our case $\{1,2\}$), c_j is the rate that tier j is able to provide at users of type $x \in \mathcal{L}$ and $\mathcal{U}_j$ is the utilization of each tier. This association rule is applied iterative among all classes, until convergence.

It is easy to see that the above lemmas can be easily extended to more than two tiers.

6.4 Results

Already today it is possible to integrate WiFi networks into the core networks of cellular systems, and perform offloading of traffic to WiFi access points. In future releases of 3GPP, a tighter integration of WiFi and LTE technologies is expected. For this reason, we choose a heterogeneous RAT scenario consisting of LTE and WiFi orthogonal tiers, as a case study. We will consider the following "fixed" parameters for the two networks: (i) pathloss $\alpha = 4$, (ii) thermal noise $\sigma^2 = -100$ dBm (iii) $BW_{LTE} = BW_{WiFi} = 20$ MHz, (iv) one antenna per eNodeB and one spatial stream per WiFi AP. Finally, we should mention that if the thermal noise is much smaller than the interference (which is the case in our system), the value of P_{tx} does not affect the results, as shown in [20]. The rest of the parameters will act as variables, and we'll discuss their value range per scenario.

For the WiFi network, a number of different setups and 802.11 standards could be considered. The traditional WiFi protocol is tuned to a roughly 20 MHz channel. This channel get chosen from a number of partially overlapping channels, usually with the criterion of maximum SINR, and this number of channels differs between countries. Additionally, the newer versions of WiFi (n/ac) have the capability of channel bonding in order to operate with 40 to 160 MHz. Larger bandwidths could also be considered via carrier aggregation in LTE. All of those additional channel capabilities are orthogonal to our model and out of our scope, so we assume for simplicity and fairness that both networks operate with 20 MHz.

Finally, as explained earlier, current WiFi implementation operates closer to a throughput fair scheduler. However, as mentioned in Chapter 3, the WiFi scheduler could be modified to avoid the "WiFi anomaly" problem and operate as resource fair [55]. We will therefore consider WiFi with both types of schedulers, in order to better understand their impact, individually and in a 2-tier setup.

6.4.1 Comparing Different RATs

Having validated our theoretical results about the coverage probability and the prediction of network utilization and delay in Chapter 4, we proceed now with their direct application to

different scenario of interest, in order to obtain insights regarding the congestion probability of a BS (the probability that a BS's load is $\rho > 1$) and flow delay statistics in large random topologies. Firstly we study the impact of the following elements on flow-level performance: (a) the MCS-SINR relation (which differs between WiFi and LTE), (b) the scheduler (throughput fair and resource fair), and (c) the type of interference (always-on or load-based).

We do this initially for single-tier systems, before moving on to multi-tier systems. This will also facilitate our subsequent discussion of 2-tier systems, where multiple factors affect performance concurrently.

At this point, it is useful to introduce the following abbreviations for the legends in all figures: (*ON*) or (*LB*) refer to the always ON or load-based interference case respectively. Additionally, for WiFi tiers (*app1*) refers to the resource-fair and (*app2*) to the throughput-fair scheduling policy approximation (see Section 5.2.2).

Fig. 6.2 and 6.3 present compactly the performance (congestion probability and delay) for the two networks of interest (LTE, WiFi) for the same density of connected users ($\lambda_u = 100$) and for average flow size $\langle s \rangle = 12.5$ Mbytes. As already mentioned at Section 5.2.2, two different approximations for the MAC performance of WiFi are assumed: (i) similar MAC performance with LTE (i.e., resource fair scheduler), which is the best case scenario for WiFi; this is a valid approximation for very low loads or assuming a modified WiFi scheduler, and (ii) an asymptotic approximation which is accurate for real, throughput-fair WiFi schedulers, as flow sizes increase. We stress here that the respective congestion probabilities are the same for each WiFi scheduler, as this only depends on the incoming job rate and the average service rate $\langle \mu \rangle$, which are the same in both cases, as we showed in Section 5.2.2. We therefore only show 2 WiFi curves on Fig. 6.2 (one for the Always ON and one for the Load-Based cases).

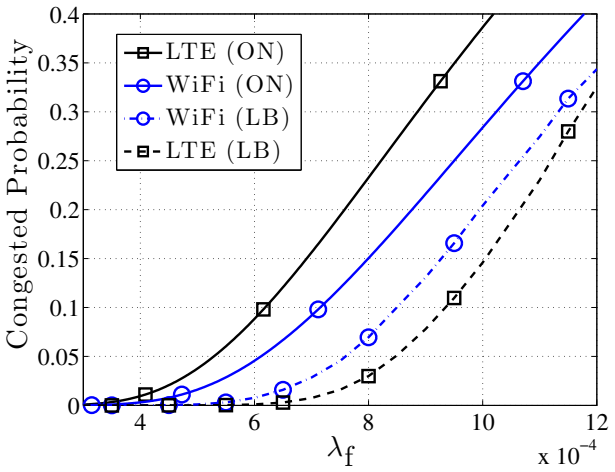


Figure 6.2 – Congestion probability w.r.t flow density λ_f

Looking at the mean delay, in Fig. 6.3, and comparing the two different cases of the WiFi schedulers, it is clear that resource-fair version of WiFi outperforms the throughput fair one, as expected. For the load-based case, when the network load causes 10% of congestion probability, the average flow delay of the throughput fair WiFi system is 25% higher compared to the resource fair one.

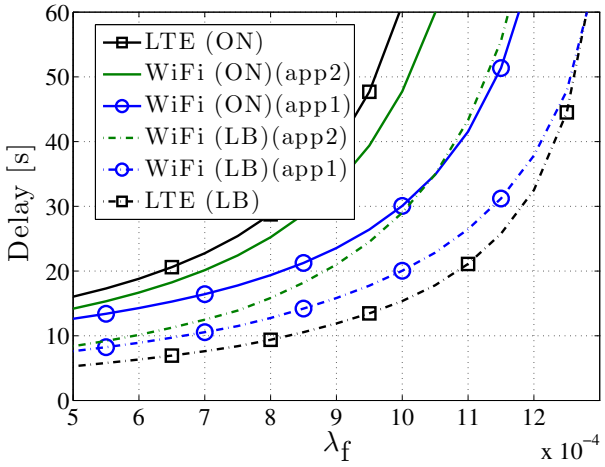


Figure 6.3 – Delay of each network w.r.t flow density λ_f

Extended discussion about comparisons between LTE and WiFi RAT can be found in Section 4.6.2.

6.4.2 Cooperative 2-tier HetNets

In this last section, we move on to multi-tier HetNets which is the main focus of this chapter. We are interested in understanding the impact of coexistence of different RATs in orthogonal frequencies (the case of coexistence in the same band could also be handled with some modifications by our model, but this is part of future work). Particularly, our goal is to capture the impact of different types of association criteria between different tiers, on the performance gains by introducing a 2nd tier. As mentioned before, the tier-association criteria of interest are:

Off-load: This is the simplest (and most aggressive offload) policy, where the user, if able to establish connection with the WiFi network, does it without any further criterion.

Max-SINR: Here, the user choses to associate with the tier that provides the best SINR, thus attempting to improve

Min-Delay: In this scenario, the user choses to associate with load related criteria in order to minimize the average delay of the system. In our case the association criterion, between tiers, is the algorithm which proposed at [46].

Firstly, we examine the simple Off-load policy. We assume that LTE is the primary network and WiFi the secondary one with the same density. Fig. 8.13 presents two different cases of this 2-tier HetNet. The difference between those two cases is about the WiFi scheduler: one were the WiFi AP operate as an "ideal", resource-fair scheduler and one as throughput-fair. Additionally, for comparison reasons, we include as "baseline" plot a single-tier LTE network with double BS density (i.e. the total number of LTE BS is equal to the sum of LTE BS and WiFi AP in the other two scenarios). Interestingly, for the saturated case almost all scenarios perform the same. What is particularly surprising is that the Off-loading case performs almost the same as the single-tier LTE case: while the total density of the BSs is the same for both cases, the Off-loading scenario uses double the spectrum than the single-tier LTE scenario. In order to sketch the explanation we mention: 1) Off-load association does not affect the MCS

distribution of each tier, 2) on the always ON interference the MCS distribution of each tier does not depend on the BS's density (the gain of the probabilistic higher received power is equal with the loss of higher interference). 3) In always ON case, WiFi captures slightly less than the 50% of traffic. Taking into account the aforementioned, the gain of the second tier is only that the cardinality of the users to the BS decreasing, just like the single-tier with higher BS density. The picture totally changes in the load based case, because the extra spectrum means less users per band and therefore, less interference.

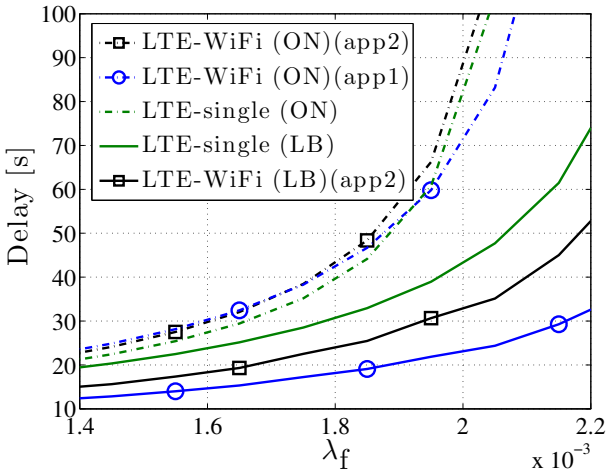


Figure 6.4 – Off-load policy for different 2-tier network cases, w.r.t. flow density λ_f

Nevertheless, if we turn our attention to the load-based interference cases, we see that: (i) the WiFi scheduler highly affects the overall performance; (ii) the 2-tier network outperforms the LTE-only for both schedulers, which is more in-line with what we would have expected. This further underlines the importance of load-based analysis, which in this case not only has a quantitative, but also a clear qualitative impact.

For the rest of this section, we only consider a best case WiFi network (i.e., resource fair), in order to focus our attention on association policies, and understand the limits of performance improvements by introducing a WiFi tier. To be more realistic we assume now denser secondary network than the primary one. More precise, a secondary WiFi network with $\lambda_{WiFi} = 5$, and a primary LTE network with $\lambda_{LTE} = 1$. Congestion probability and per flow delay for different traffic input rates λ_f, are depicted in Fig. 8.14 and 8.15, respectively.

Looking at the Off-load and Max-SINR criteria for the saturated case, the congestion probability of both cases is almost equal. However, Max-SINR performs better, with respect to mean delay. However, considering the load-based interference cases, the Off-load policy is much more robust with respect to congestion probability and outperforms Max-SINR with respect to delay, as well.

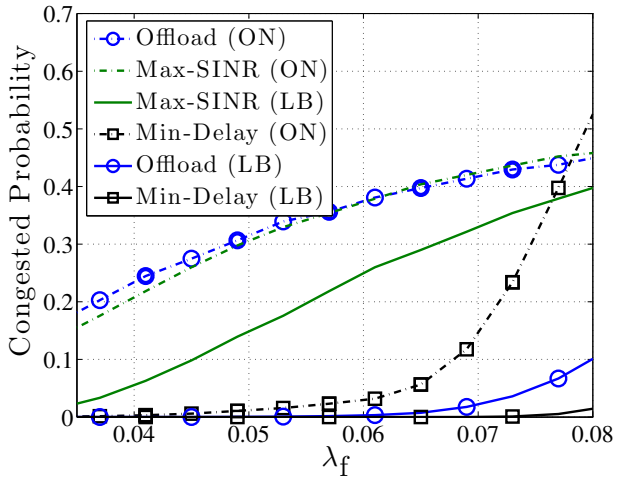


Figure 6.5 – Congested probability comparison of different association schemes, w.r.t. flow density λ_f

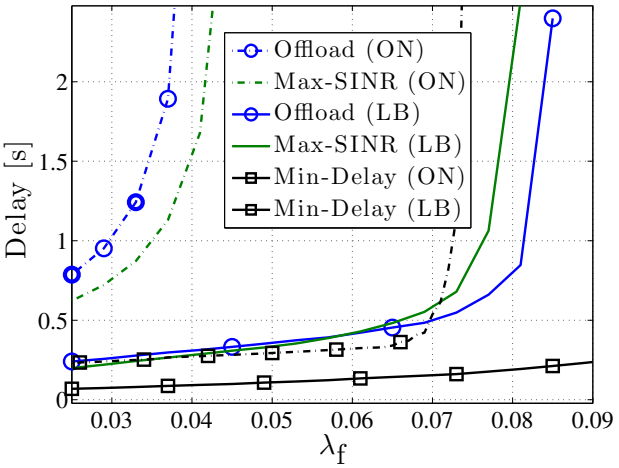


Figure 6.6 – Delay comparison of different association schemes, w.r.t. flow density λ_f

This discrepancy between the saturated and load-based cases originates from fact that saturated analysis is able to capture only one side of the gain stemming from increasing network density. On the one hand, the saturated case correctly captures the fact that an "arbitrary" BS on average has to serve fewer users, in a denser network (thus dealing with a smaller ρ due to a decrease in the numerator, i.e., the input traffic). On the other hand, it fails to capture that the surrounding BSs will be less loaded as well, and therefore causes less interference, which in turn, leads to even better performance for the (fewer) users served (due to an increase in $\langle\mu\rangle$ and a resulting further decrease in ρ). As a consequence, the saturated model underestimates association schemes that tend to utilize the denser (WiFi) network more.

Last but not least, it is clear from Fig. 8.15 that the Min-Delay association policy significantly

outperforms the other two, by up to an order of magnitude or more, for high loads. Unlike the other two policies that only consider PHY layer performance (MAX-SINR) or naively try to reduce the load of the primary network (Off-load), Min-Delay directly takes into account the actual load experienced, which plays the key role on the per-flow performance. It is also interesting to note that, especially for low loads, the Min-Delay policy is also quite stable in terms of congestion probability (Fig. 8.14). While the considered association policies are admittedly abstracted version of real detailed policies, we believe these results make a strong case for sophisticated, load-based association mechanisms in future HetNets, in order to better balance the loads between tiers and ensure the best user experience.

6.5 Conclusion

In this chapter we presented an analytical frame work that analyze the flow-level performance of a multi-tier HetNet and compares some common used association criteria. Three main conclusions came up from this chapter: i) The scheduling policy could strongly effect system's flow-level performance, even if the PHY characteristics are the same as we saw when comparing the two different cases of the WiFi scheduler (resource fair, throughput fair). ii) The two different interference approaches, always ON and load-based, change totally the performance of the system (single-tier or multi-tier), so we should be very careful about this assumption when model a system. iii) The gain of the load related association policy (min-Delay) is surprisingly high comparing to the more traditional ones (Off-load or Max-SINR).

Chapter 7

Conclusions and Future Research

This thesis contributes to the general problem of modeling and performance analysis of randomly placed wireless networks. We follow a not so well investigated but very interesting approach that focuses on the analysis of the flow-level performance of the network. We strongly believe that as the network is becoming more and more loaded and packet oriented those approaches will provide a lot of benefits and insights in the network design.

Our goal was to develop a framework that exports some extra information about the network, such as the users delay and the BS congested probability that traditional approaches are not able to capture . In order to achieve that, we combine ideas from queueing theory to model the dynamic performance of a BS scheduler and stochastic geometry in order to model the network's topology. We utilized those two mathematical tools with our PHY layer abstraction in order to create a solid analytical framework for an accurate prediction of the flow-level performance of a large randomly placed network.

The aforementioned analysis considers both the case of always ON interfering neighboring BSs, as well as the case of load-depended interference. It turns out that the performance gap between the aforementioned cases could be rather high, affecting not only quantitative insights, but often qualitative conclusions as well, and thus should be carefully taken into account during network design. Another parameter that affects the systems performance is the selection of the scheduler; we saw that even if the PHY characteristics remain the same, the user performance (delay) strongly depends on the BS scheduler, even if the operator-wise performance (congestion probability) does not change. Some other interesting general concussions are: 1) The flow-level performance of the network depends on the service rate that usually calculated according to the harmonic mean of the rates and not on the statistical mean; that means that the edge users affect much more on the network performance. 2) The relationship between users delay and network load is not linear, the delay explodes when the load is over a limit. This limit depends on the scheduling policy but we can set it roughly to 0.9.

Additionally, we compared our work with the SoA frameworks that use mean value approaches in order to estimate the performance of the network and we saw that this approach outperforms significantly from ours especially in the high load regime (up to 40 time higher estimation error in network load).

Furthermore we analyzed theoretically the energy cost of such a network and we studied the trade-off between energy efficiency and network's flow level performance. The main network parameter of our study was the density of the BS, and we show that in general turning off BS

does not necessarily lead to energy savings as often believed. We presented some theoretical results about the reduction of BS density in the case of users density reduction and we provided a rule of thumb, when this reduction leads to energy gain. On the other hand we show that if the amount of users is not reduced there is a possibility of energy improvement without affecting a lot user's QoE when the operational energy is not much greater than the constant energy consumption term.

Additionally, we expanded our framework for the case of heterogenous networks considered multi-tier topologies and different radio access technologies. We model some common association criteria, and evaluating their impact on both user- and network-centric performance. This study shows that if we consider the load in our tier-association the flow-level performance gain could be extremely high.

This study given the large range of parameters and degrees of freedom could never be complete, but we provide representative insights, and demonstrate the utility of our proposed analytical framework.

Future Research

Still there are a lot of open problems in our framework that future work could probably fulfill. The most important of that is the theoretical study of the convergence between MCS distribution and service rate in the load based interference case. This study should focus to derive bounds about when this convergence is feasible and if is possible to end up with a closed form solution of the service rate. We know that this is not an easy task at all, but we believe that this result could help a lot of researchers to study analytically the dynamic network performance in the more realistic scenario of load based interference.

As further future work, it will be interesting to apply our framework, together with different association criteria, in Carrier Aggregation scenarios. We should examine scenarios were the users schedule their traffic to deferent tiers according to their flow size and to investigate for possible improvements, e.g., a user will send all of his sort contents like "text" through his cellular connection and the large contents like photos though his WiFi connection.

Additionally, we believe that our framework could be modified to analyze scenarios of LTE and WiFi coexistence in the same bands and to study mechanisms in order to be this coexistence as fair as possible such as Almost Blank Subframes or carrier sense LTE or optimal tune of the MCS thresholds and transmitting power levels for both RATs.

Furthermore, we should study the network performance for the case were HetNets operates in the same frequency bands and the small cells are able to do Cell Range Expansion in order to off-load a part of the traffic from the macro cell. On the one hand the small cell will off-load some users on the other will cause more interference.

Finally, an interesting study will be to expand our framework for the case of priority users. This could happen by assuming that a part of the network will serve only the primary users or by using some of the already developed results of queueing theory that provide analytical expressions for queues with different class of users, where under a proper modeling we can utilize those results in wireless systems as well.

Appendices

Appendix A

Distribution of the Number of Poisson Points in Poisson Voronoi Tessellation

A.1 Introduction

Voronoi Tessellation is a widely used mathematical framework for space subdivision in random partitions. It is used in a variety of fields such us statistical mechanics, quantum field theory, astrophysics, telecommunications, social networks, biology etc. On the other hand, Poisson Point Process (PPP) is one of the most common tools at stochastic geometry, since it provides suitable mathematical models and appropriate statistical methods to analyze macroscopic properties by averaging all possible micro-states. In this Appendix we study and solve the following problem: Let two independent sets Φ_1 and Φ_2 follow homogeneous PPP having different densities in two dimensional space assuming that Voronoi Tessellations are generated with respect to Φ_1. What is the probability distribution of Φ_2 cardinality at an arbitrary tessellation of Φ_1? Furthermore, we calculate the first and the second moments of the aforementioned probability distribution. In our problem, Φ_1 represents the positions of telecommunication Base Stations (BSs) and Φ_2 the position of the Users.

A.2 Base Stations and Users topology

We model positions of the BSs and Users as a homogeneous Poisson Point Process. So, if the density of users (or BSs) at a certain area A is λ then the Number N of them is a random variable which is given from

$$P(N = k \mid A) = \frac{(\lambda A)^k e^{-\lambda A}}{k!}, \quad k = 0, 1, \dots \tag{A.1}$$

Homogeneous, means that after the chosen of the number of BSs at certain aria A, their locations follows uniform distribution at $2D$ space.

A.3 Distribution of the Cell Size

A given set of centers can divide the space to specific regions, known as Voronoi Tessellations (or Voronoi Regions). Each of them contains those points of space that are closest to the same center. In our case, the centers is the location of the BSs so Voronoi tessellation represents the aria of coverage of each one of them (we assume that all the BS have the same transmit power). At the particular case, where the centers are randomly and uncorrelated distributed, is called Poisson Voronoi Tessellation (PVT), see Figure A.1.

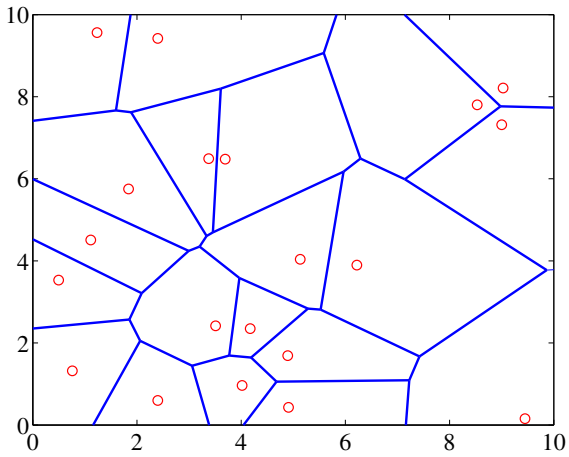


Figure A.1 – The red cycles represent the points and the blue lines the corresponding Voronoi regions

We are interested in the PDF of Voronoi cells size (2D), if the number BSs follows homogeneous PPP. Unfortunately this is an open mathematical problem and does not exist any close form solution until today. However there exist several PDFs that provide an approximate numerical solution, based of the Gamma distribution.

$$g(x;a,b,c) = \frac{ab^{\frac{c}{a}}x^{c-1}e^{-bx^a}}{\Gamma(\frac{c}{a})} \tag{A.2}$$

Note that $x = \frac{S}{\langle S\rangle}$, S is the specific size of a certain cell and $\langle S\rangle$ is the average size of all the Voronoi Regions. For more compact notation for our problem we can write $x = S\lambda_{BS}$, where now λ_{BS} is the density of the BS. a, b, c are fitting parameters. At [20], we take the results of the fitting parameters $a = 1.07950,\ b = 3.03226,\ c = 3.31122$. In [75] the same author present us the best fitting results if we fix $a = 1$, which is $b = 3.52418$ and $c = 3.52440$. We calculate the difference between them $\delta^{3,2} = |g(x;a,b,c) - g(x;a=1,b,c)|$, the average absolute difference is $\delta^{3,2}_{av} = 9.8977e-04$ and the maximum difference is $\delta^{3,2}_{max} = 0.0103$.To have an intuition about the result, means that the maximum difference if roughly 1% of the $\langle S\rangle$. See Figure A.2. The superscript 3,2 denotes the comparison between models with 3 and 2 and fitting parameters respectively.

At [76], a simpler distribution has been proposed taking into account only the dimensions of the space.

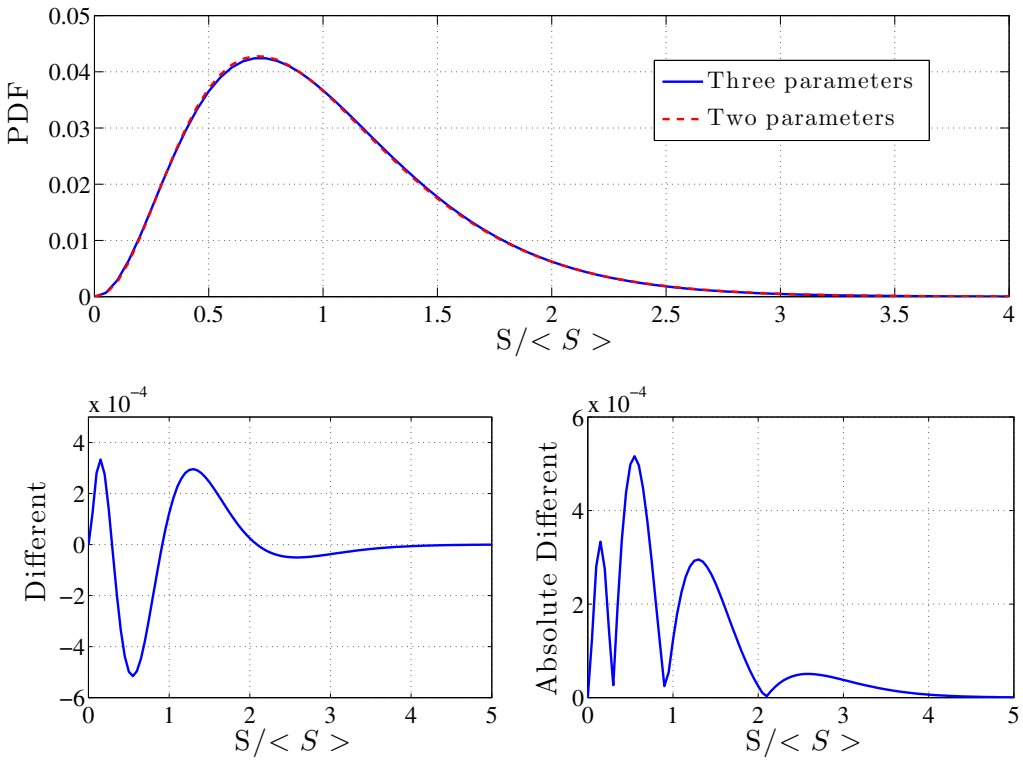


Figure A.2 – difference between three and two parameters fitting models

$$f(x;d) = \frac{[\frac{3d+1}{2}]^{(3d+1)/2} x^{\frac{3d-1}{2}} e^{\frac{-(3d+1)x}{2}}}{\Gamma(\frac{3d+1}{2})} \tag{A.3}$$

Where d is number of dimensions of the space. In our case $d = 2$ and again $x = S\lambda_{BS}$. So the result for our case is the same if you set at equation A.2, $a = 1,\ b = c = 3.5$. So we get

$$f(S) = \frac{343}{15}\sqrt{\frac{7}{2\pi}}(\lambda_{BS}S)^{\frac{5}{2}} e^{-\frac{7}{2}\lambda_{BS}S}\lambda_{BS} \tag{A.4}$$

Which is simpler than above. It is obvious that there is a trade off between accuracy and complexity. We compute the difference $\delta^{3,1} = |g(x;a,b,c) - f(x;d=2)|$. The average difference is $\delta^{3,1}_{av} = 0.0011$ and the maximum difference is $\delta^{3,1}_{max} = 0.0108$, Figure A.3. Which are almost the same as the case of tow parameters. For higher dimension spaces this accuracy does not hold.

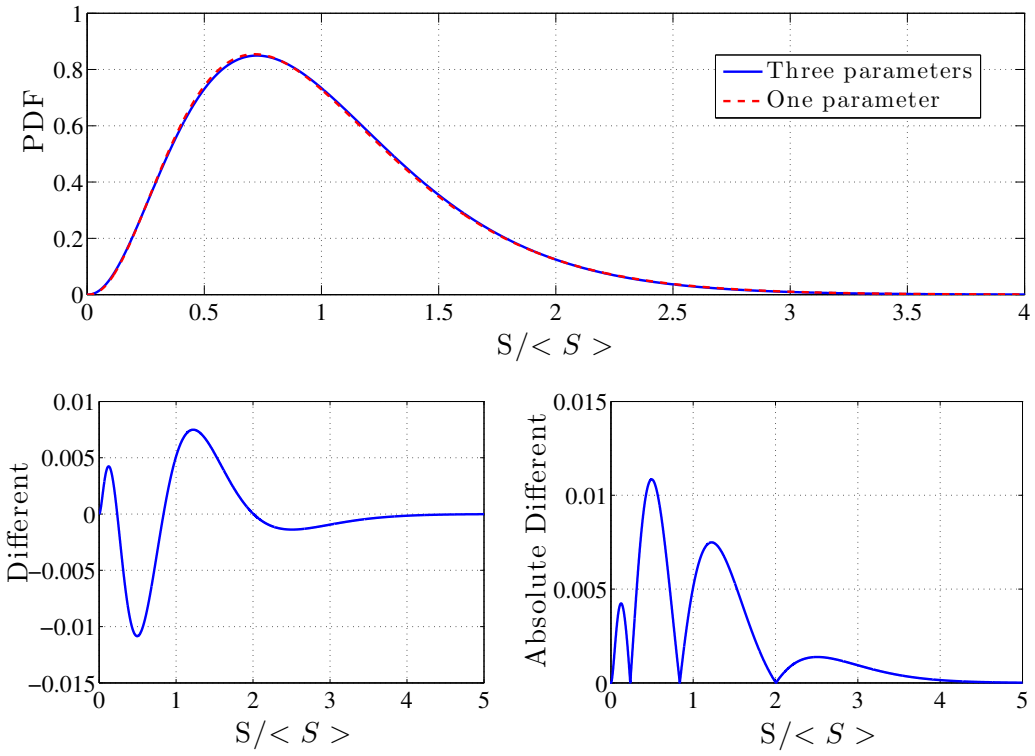


Figure A.3 – difference between three and one parameters fitting models

A.4 PDF of Number of users in a Cell

As the users are distributed as a homogeneous PPP (the following results holds and for the non-homogeneous case), as we see above

$$P(N_u = k \mid S) = \frac{(\lambda_u S)^k e^{-\lambda_u S}}{k!}, \quad k = 0, 1, \ldots \tag{A.5}$$

Thur $P(N_u = k)$ is calculating from

$$P(N_u = k) = \int_0^\infty P(N_u = k \mid S) f(S) dS \tag{A.6}$$

$$P(N_u = k) = \int_0^\infty \frac{(\lambda_u S)^k e^{-\lambda_u S}}{k!} \frac{343}{15} \sqrt{\frac{7}{2\pi}} (\lambda_{BS} S)^{\frac{5}{2}} e^{-\frac{7}{2}\lambda_{BS} S} \lambda_{BS} dS \tag{A.7}$$

$$P(N_u = k) = \frac{\lambda_u^k}{k!} \frac{343}{15} \sqrt{\frac{7}{2\pi}} \lambda_{BS}^{\frac{7}{2}} \int_0^\infty S^{k+\frac{5}{2}} e^{-(\lambda_u + \frac{7}{2}\lambda_{BS})S} dS \tag{A.8}$$

We set $\mathcal{A}(k) = \frac{\lambda_u^k}{k!} \frac{343}{15} \sqrt{\frac{7}{2\pi}} \lambda_{BS}^{\frac{7}{2}}$.

$$P(N_u = k) = \mathcal{A}(k) \int_0^\infty S^{k+\frac{5}{2}} e^{-(\lambda_u + \frac{7}{2}\lambda_{BS})S} dS \tag{A.9}$$

The calculation of the integral is not so trivial, so we will present the basic steps. Which give us a better understanding of the solution. First of all we set $\beta = \lambda_u + \frac{7}{2}\lambda_{BS}$ After $k+2$ integrations by parts we get

$$\int_0^\infty S^{k+\frac{5}{2}} e^{-\beta S} dS = \frac{\left(\frac{1}{\beta}\right)^{k+2} \Gamma(k+\frac{7}{2})}{\sqrt{\pi}} \int_0^\infty S^{\frac{1}{2}} e^{-\beta S} dS \; . \tag{A.10}$$

We continue with the calculation of the new integral by set a new variable $u^2 = S$ and $dS = 2udu$

$$\int_0^\infty u e^{-\beta u^2} 2u du = -2 \int_0^\infty -u^2 e^{-\beta u^2} du \; , \tag{A.11}$$

note that we can write the integral argument as derivative of constant β, $-u^2 e^{-\beta u^2} = \frac{\partial}{\partial \beta} e^{-\beta u^2}$ so we get

$$\int_0^\infty u e^{-\beta u^2} 2u du = -2 \int_0^\infty \frac{\partial}{\partial \beta} e^{-\beta u^2} du \tag{A.12}$$

$$= -2 \frac{\partial}{\partial \beta} \int_0^\infty e^{-\beta u^2} du \; , \tag{A.13}$$

where is a Gaussian integral so

$$\int_0^\infty S^{\frac{1}{2}} e^{-\beta S} dS = -2 \frac{\partial}{\partial \beta} \frac{\sqrt{\pi}}{2\sqrt{\beta}} \tag{A.14}$$

$$= \frac{1}{2} \sqrt{\pi} \beta^{-3/2} \; , \tag{A.15}$$

finally, if we combine them

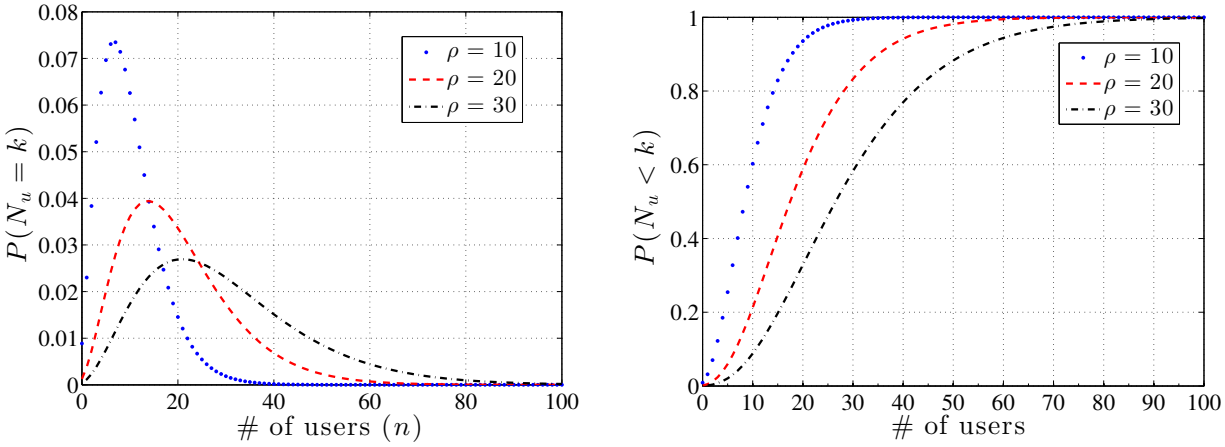
$$P(N = k) = \mathcal{A}(k) \left(\frac{1}{\beta}\right)^{k+\frac{7}{2}} \Gamma(k + \frac{7}{2}) \; . \tag{A.16}$$

or

$$P(N_u = k) = \frac{343}{k!15} \sqrt{\frac{7}{2\pi}} \frac{\lambda_{BS}^{\frac{7}{2}} \lambda_u^k}{(\lambda_u + \frac{7}{2}\lambda_{BS})^{k+\frac{7}{2}}} \Gamma(k + \frac{7}{2}) \tag{A.17}$$

If we derive both numerator and denominator by $\lambda_{BS}^{k+\frac{7}{2}}$, we take an expression which depends only at the ratio between of the users and BS density $\rho = \frac{\lambda_u}{\lambda_{BS}}$, which is nice!

$$P(N_u = k) = \frac{343}{k!15} \sqrt{\frac{7}{2\pi}} \frac{\rho^k}{(\rho + \frac{7}{2})^{k+\frac{7}{2}}} \Gamma(k + \frac{7}{2}) \tag{A.18}$$



a: PDF for different values of ratio $\rho = \frac{\lambda_u}{\lambda_{BS}}$

b: CDF for different values of ratio $\rho = \frac{\lambda_u}{\lambda_{BS}}$

Figure A.4 – Moments of user's cardinality

From "reference of equation" and after some calculations, the probability of having "zero" points at one PVR is $P(N_u = 0) = P_0 = \frac{343}{8\alpha^{7/2}}\sqrt{\frac{7}{2}}$. Thus we re-write the last result in a more intuitive way. At the Figures 4.1 and A.4 we see PDF and CDF for different values of ratio $\rho = \frac{\lambda_u}{\lambda_{BS}}$.

$$P(N_u = k) = \left(\frac{\rho^k}{k!(\rho + 7/2)^k} \prod_{n=1}^{k} (n + \frac{5}{2}) \right) P_0 \tag{A.19}$$

A.4.1 First and Second Moments of the Distribution

We calculate the Mean and the Variance of the Number of Poisson Points in Poisson Voronoi Tessellation. By equation A.17 we calculate the average number of points and their variance in a random PVT by $\langle k \rangle = \sum_{k=0}^{\inf} k \cdot P(N_u = k)$ and $Var_k = \langle k^2 \rangle - \langle k \rangle^2$. Hopefully the series converge, so first and second moments of the distribution are

$$\langle k \rangle = \rho \qquad \text{and} \qquad \text{var}_k = \rho + \frac{2}{7}\rho^2 \, . \tag{A.20}$$

From equation A.20, we observe that the variance of the number of users within a cell drops quadratically w.r.t the density of deployed BSs, but the mean drops to. The coefficient variation is greater than 1. So the relative variance of points in a voronoi cell is not decreasing by the rising of BS density.

A.5 Approximation of the result

If we take into account the asymptotic of gamma function $\lim_{n\to 0} \frac{\Gamma(n+\alpha)}{\Gamma(n)n^\alpha} = 1$ and the definition $\Gamma(k) = (k-1)!$, eq. A.18 is significant simplified to

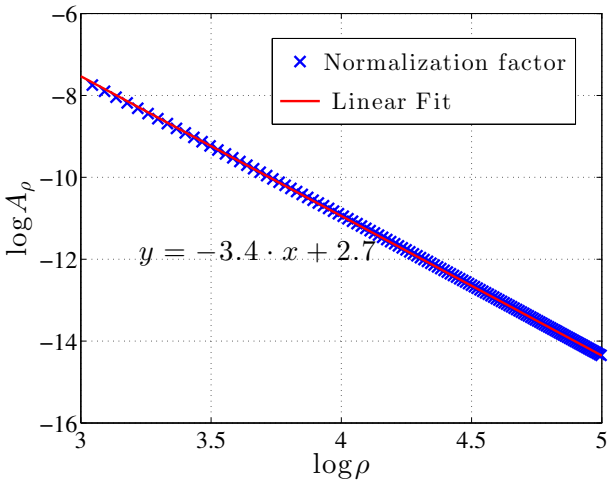


Figure A.5 – Fit to normalization factor w.r.t ρ

$$P(N = k) = A_\rho u_\rho^k k^{5/2} \,, \tag{A.21}$$

where $u_\rho = \frac{\rho}{\rho+7/2}$. Due to the asymptotic approach of gamma function the pdf lost its normalization. Therefore, A_ρ is the new normalization factor which depends on ρ. Fig. A.5 provides an analytical fit to the normalization factor and Fig. A.6 shows the square error between the Approximation and the distribution of users cardinality for $\rho = 30$.

The simpler analytical form eq. A.21 allows not only further theoretical use of the result but also provides much wider computing operability range instead of eq. A.18. e.x. eq. A.18 in the process to calculate probability having $k = 169$ users in a cell (independent of ρ) exceeds $1.7977e+308$ digits (largest finite floating-point number in IEEE double precision), at the same time eq. A.21 for the worst case $\rho = 1$ does not exceed $4.0466e+111$ and for more reasonable values e.x. $\rho = 20$ the needed floating-points is $7.9601e+09$.

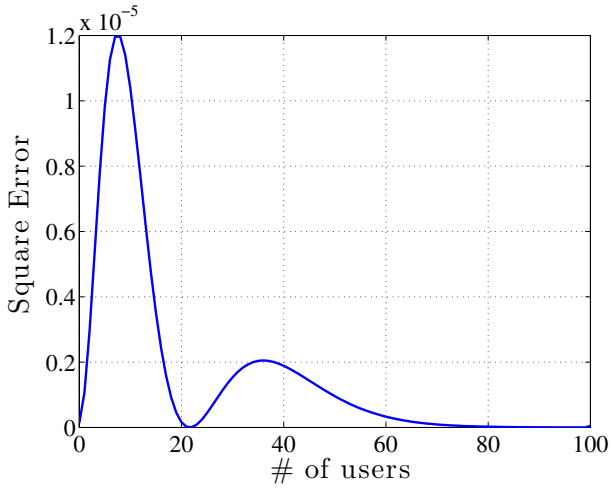


Figure A.6 – Square Error of Approximation for $\rho = 30$

Appendix B

Derivation of Load-Based Coverage Probability

B.1 Introduction

In this Appendix we present the derivation of the coverage probability of an arbitrary user in a random cellular network (assuming thermal noise negligible compared to interference), when the average utilization of BSs is $\mathcal{U}$, and each BS is interfering only for the amount of time that it is serving users (i.e., for a percentage of time $\mathcal{U} \leq 1$) as presented in *Lemma* (4.5.1)

$$p_c^{lb}(\tau, \alpha, N_{max}) = \sum_{n=0}^{N_{max}-1} \left(f_N(n \mid \zeta) \frac{1}{1+\mathcal{A}_{\mathcal{U}}} \right) + \overline{F_N}(N_{max} \mid \zeta) \frac{1}{1+\mathcal{A}_{\mathcal{U}=1}} . \tag{B.1}$$

Where $N_{max} = \langle \mu \rangle / \lambda_f$ is the maximum amount of associated users per BS and

$$\mathcal{A}_{\mathcal{U}} = (\tau \mathcal{U})^{2/\alpha} \int_{(\tau\mathcal{U})^{2/\alpha}}^{\infty} \frac{1}{1+u^{\alpha/2}} du . \tag{B.2}$$

B.2 Proof

Lemma B.2.1. Distance of an arbitrary user to the nearest BS is a random variable r with pdf

$$f_r(r) = e^{-\lambda \pi r^2} 2\pi\lambda r . \tag{B.3}$$

Positions of BS are described by a 2-D homogenous Poisson process, so the cdf is given by $P[r \leq R] = F_r(R) = 1 - e^{-\lambda\pi R^2}$ and the pdf can be found as $f_r(r) = \frac{dF_r(r)}{dr}$.

Coverage Probability

Coverage probability is the probability that SINR of an arbitrary user is greater than a given threshold T

$$\begin{aligned} p_c(T,\lambda,\alpha) &= E_r\left[P\left[SINR > T|r\right]\right] \\ &= \int_{r>0} P\left[SINR > T|r\right] f_r(r)\, dr \\ &= \int_{r>0} P\left[h > Tr^{\alpha}\left(\sigma^2 + I_r\right)|r\right] e^{-\lambda\pi r^2} 2\pi\lambda r dr\ . \end{aligned} \tag{B.4}$$

Where I_r is the mean Interference at distance r (the rest of parameters have been defined at $A.7$). Taking into account the channel model $h \sim \exp(P_{\text{tx}})$, The probability $P\left[h > Tr^{\alpha}\left(\sigma^2 + I_r\right)|r\right]$ could be re-defined

$$\begin{aligned} P\left[h > Tr^{\alpha}\left(\sigma^2 + I_r\right)|r\right] &= E_{I_r}\left[e^{\left(-P_{\text{tx}}Tr^{\alpha}\left(\sigma^2+I_r\right)\right)}|r\right] \\ &= e^{-P_{\text{tx}}Tr^{\alpha}\sigma^2} E_{I_r}\left[e^{\left(-P_{\text{tx}}Tr^{\alpha}I_r\right)}|r\right]\ . \end{aligned} \tag{B.5}$$

In the non-saturated case interference is given by $I = \sum_{i\in\Phi\setminus\{b_0\}} \mathcal{U}_i h_i R_i^{-\alpha}$, where $\mathcal{U}_i$ is the utility of the i-th BS which is equal to the probability to be ON. So, setting $s = P_{\text{tx}}Tr^{\alpha}$, the expectation of Eq.(B.5), $E_{I_r}\left[\exp\left(-P_{\text{tx}}Tr^{\alpha}I_r\right)|r\right]$, could be re-written as

$$\begin{aligned} E_{I_r}\left[\exp\left(-sI_r\right)|r\right] &= E_{\mathcal{U},\Phi,h}\left[\exp\left(-s\sum_{i\in\Phi\setminus\{b_0\}} h_i\mathcal{U}_i R_i^{-\alpha}\right)\right] \\ &= E_{\mathcal{U},\Phi}\left[\prod_{i\in\Phi\setminus\{b_0\}} \frac{P_{\text{tx}}}{P_{\text{tx}} + s\mathcal{U}_i R_i^{-\alpha}}\right]\ , \end{aligned} \tag{B.6}$$

Lemma B.2.2. As the cardinality of BSs is raising, the distribution of $\mathcal{U}$ becomes independent from Φ's realizations. Thus, $E_{\mathcal{U},\Phi}[\cdot]$ could be treated as two independent expectations $E_{\mathcal{U}}[E_{\Phi}[\cdot]]$. This happens because of the law of large numbers and the ergodicity of the process and can be illustrated at Fig. B.1, where the "variance" of the cdf $\mathcal{U}$ is decreasing w.r.t. BS cardinality.

Additionally, due to the properties of exponential distribution,

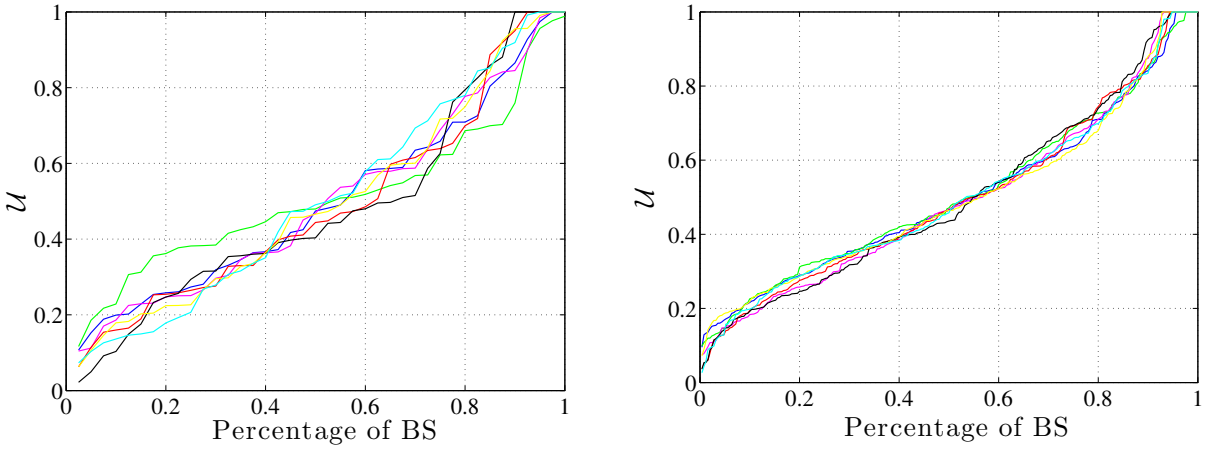
$$E_{\Phi}\left[\prod_{x\in\Phi} f(x)\right] = \exp\left(-\lambda\int_{R^2}\left(1 - f(x)\right)dx\right)\ , \tag{B.7}$$

and after some trivial calculations Eq.(B.6) is equal to

$$E_{\mathcal{U}}\left[\exp\left(-2\pi\lambda\int_r^{\infty}\left(1 - \frac{1}{1 + T\mathcal{U}\left(\frac{r}{R}\right)^{\alpha}}\right)RdR\right)\right].$$

Finally, by setting $u = \left(\frac{R}{r(\mathcal{U}T)^{1/\alpha}}\right)^2$ the initial expectation of eq. B.5 is equal to

$$E_{I_r}\left[\exp\left(-sI_r\right)|r\right] = E_{\mathcal{U}}\left[\exp\left(-2\pi\lambda\mathcal{A}_{\mathcal{U}}\right)\right]\ . \tag{B.8}$$



a: $\#BSs = 90$

b: $\#BSs = 250$

Figure B.1 – CDF of utilization for different realization for two different number of BSs

Where $\mathcal{A}_{\mathcal{U}} = (T\mathcal{U})^{2/\alpha} \int_{(T\mathcal{U})^{2/\alpha}}^{\infty} \frac{1}{1+u^{\alpha}/2} du$.

So, by replacing Eq.(B.5), (B.8) to Eq.(B.4) the coverage probability becomes

$$\begin{aligned} p_c(T,\lambda,\alpha) &= \int_{r>0} 2\pi\lambda r e^{-\lambda\pi r^2} e^{-P_{\text{tx}} T r^{\alpha} \sigma^2} E_{\mathcal{U}}\left[e^{-2\pi\lambda\mathcal{A}_{\mathcal{U}}}\right] dr \\ &= E_{\mathcal{U}}\left[\int_{r>0} 2\pi\lambda r e^{-\lambda\pi r^2(1+\mathcal{A}_{\mathcal{U}})} e^{-P_{\text{tx}} T r^{\alpha} \sigma^2} dr\right] . \end{aligned}$$

The above equation can be significantly simplified under the assumption that $\sigma^2 << I$ so $\sigma^2 = 0$.

$$p_c(T,\lambda,\alpha) = E_{\mathcal{U}}\left[\frac{1}{1+\mathcal{A}_{\mathcal{U}}}\right] . \tag{B.9}$$

Lemma B.2.3. We assume that cell's average service rate $\langle\mu\rangle$ is independent from the users' cardinality. There is a dependency between cell size and users' cardinality as well as between cell's size and $\langle\mu\rangle$. We state that the dependent of those dependencies is negligible.

A large scale topology is presented in Fig. B.2. Each dot represents a BS of a given number of users and average cell rate. We can observe that the linear fit is almost constant w.r.t. the cardinality of users; the linear term is 5 orders of magnitude less than the constant term. So, our assumption that the $\langle\mu\rangle$ and the number of associated users could be treated as independent variables is confirmed.

Applying Lemma B.2.3 we assume that $\langle\mu\rangle$ is the equal to all cells. Thus, we define utility's distribution via users' cardinality $f_N(n)$

$$\mathcal{U} = \begin{cases} \frac{n\cdot\lambda_f}{\langle\mu\rangle} & , n < N_{max} \\ 1 & , n \geq N_{max} , \end{cases} \tag{B.10}$$

where $N_{max} = \frac{\langle\mu\rangle}{\lambda_f}$ is the maximum number of users that a BS could serve. Applying Eq.(B.10) to Eq.(B.9) we end up to our solution

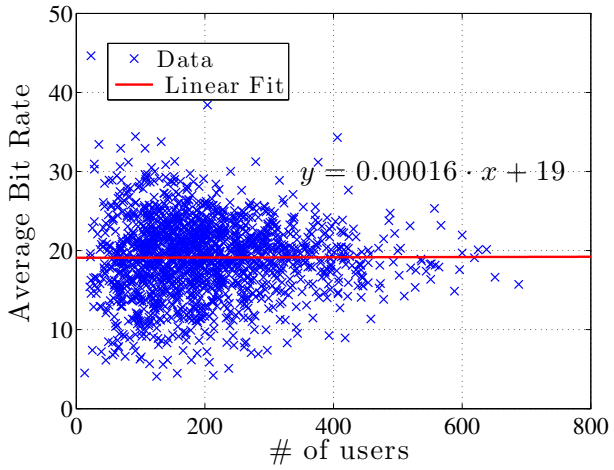


Figure B.2 – Linear Interpolation of Average Rate w.r.t. users' cardinality

$$p_c^{lb}(\tau, \alpha, N_{max}) = \sum_{n=0}^{N_{max}-1} \left(f_N(n \mid \zeta) \frac{1}{1 + \mathcal{A}_{\mathcal{U}}} \right) + \overline{F_N}(N_{max} \mid \zeta) \frac{1}{1 + \mathcal{A}_{\mathcal{U}=1}} . \tag{B.11}$$

Where $N_{max} = \langle \mu \rangle / \lambda_f$ is the maximum amount of associated users per BS and

$$\mathcal{A}_{\mathcal{U}} = (\tau \mathcal{U})^{2/\alpha} \int_{(\tau \mathcal{U})^{2/\alpha}}^{\infty} \frac{1}{1 + u^{\alpha/2}} du . \tag{B.12}$$

As mentioned before, we can consider the BS utility as $\mathcal{U} = \min(\frac{n}{N_{max}}, 1)$. Additionally, α is the path loss exponent, f_N and $\overline{F_N}$ are the pdf and ccdf of users' cardinality and as previously $\zeta = \lambda_u / \lambda_{BS}$.

Additionally, assuming $\alpha = 4$, Eq. (B.11) could be further simplified by replacing $\mathcal{A}_{\mathcal{U}}$ and $\mathcal{A}_{\mathcal{U}=1}$ with

$$\begin{aligned} \mathcal{A}_{\mathcal{U}} &= \sqrt{\frac{\tau}{N_{max}} n} \cdot arccot \left(\frac{1}{\sqrt{\frac{\tau}{N_{max}} n}} \right) \\ \mathcal{A}_{\mathcal{U}=1} &= \sqrt{\tau} \cdot arccot \left(\frac{1}{\sqrt{\tau}} \right) . \end{aligned} \tag{B.13}$$

Chapter 8

Resume [Francais]

8.1 Résumé / Introduction

L' évolution récente des communications mobiles et l' utilisation généralisée des appareils mobiles "intelligents" ont radicalement changé le comportement et les besoins de l' utilisateur mobile. Dans le monde développé, les personnes qui utilisent leurs appareils mobiles juste pour faire des appels et écrire des SMS sont la minorité. L' accès omniprésent à l' Internet, la diffusion en continu des vidéos / chansons et le téléchargement des flux de données sont quelques exigences modernes de l' utilisateur mobile. Il ne serait pas exagéré de dire que les utilisateurs modernes ont presque les mêmes exigences, indépendamment s' ils sont connectés via un réseau téléphonique depuis leur ordinateur personnel ou s' ils ont établi une connexion sans fil à travers leurs appareils cellulaires.

D' une part, la charge globale du réseau augmente, d' autre part, les stations de base ont des ressources limitées du fait qu' elles ne peuvent fonctionner que dans une partie limitée du spectre électromagnétique. Certaines solutions (temporelles) pour le problème susmentionné sont l' expansion des bandes opérationnelles en fréquence (onde mm) ou l' utilisation de plus d' antennes (MIMO massive) ou le déploiement plus dense de petites cellules. Dans cette thèse, nous sommes intéressés par le cas prometteur de petites cellules plus denses qui seront plus étroites intégrées à la macro cellule. Malheureusement, le déploiement d' un petit réseau de cellules de manière optimale, n' est pas trivial. Un petit réseau cellulaire est généralement déployé dans un style ad hoc et pas complet, dont une partie du réseau existe déjà et ne peut pas être planifiée. En outre, il existe des obstacles naturels et des limitations physiques qui ne permettent pas de déployer des stations de base partout où nous voulons. Ainsi, la topologie du réseau de petites cellules est tout à fait différente de celle de la structure traditionnelle (macro cellulaire).

Selon nous, l' avenir des communications mobiles abordera les éléments suivants: une masse d' utilisateurs entraînera un trafic de données non-descriptif qui sera diffusé à partir d' un réseau irrégulier et hétérogène. Cependant, cette image chaotique rend le problème de la modélisation du réseau et de l' analyse des performances extrêmement difficile.

Dans ce but, l' objectif principal de cette thèse est de construire un cadre analytique afin d' analyser les performances d' un réseau placé au hasard, qui sert des utilisateurs aussi placé au hasard. Pour cela, nous avons basé notre analyse sur deux outils principaux: (a) la géométrie

stochastique, pour comprendre l' impact de l' aléa topologique et des cartes de couverture et (b) la théorie des files d' attente, pour modéliser la concurrence entre les flux simultanés au sein de la même BS. Les utilisateurs sont supposé d' être non saturé; chacun génère des flux selon un processus de point de Poisson (PPP). En outre, notre cadre ne suppose pas que les stations de base voisines sont saturées ou qu' elles contribuent constamment à l' interférence afin d' obtenir des résultats de forme fermée, mais traite le problème du couplage entre la charge du réseau et la qualité de service de l' utilisateur. Le deuxième objectif est de proposer, sur la base de cette analyse, quelques directives générales de conception et / ou des idées sur des scénarios de communication spécifiques.

Plus précisément, le chapitre 1 décrit une brève introduction de la performance au niveau des flux d'un réseau placé au hasard, la motivation de notre travail ainsi que et les principales contributions de notre travail. Par la suite, le chapitre 2 présente une brève introduction aux deux principaux outils mathématiques que nous avons utilisés dans cette thèse (géométrie stochastique et théorie des files d'attente) et un aperçu du travail connexe.

Au chapitre 3, nous modélisons et étudions la couche PHY et la couche MAC des techniques d'accès radio LTE et WIFI (RAT). Les résultats de ce chapitre tels que les seuils de taux et la politique du planificateur pour chaque RAT seront largement utilisés dans les chapitres suivants.

Notre modèle n'assume pas de BS ou d'Utilisateurs saturés et dérive des mesures de performance réseau telles que le retard moyen des utilisateurs, la charge du réseau et la probabilité encombrée de BS. Enfin, nous avons appliqué notre cadre aux très populaires LAT et WIFI RAT et fourni des informations utiles sur la manière dont les caractéristiques PHY affectent les performances au niveau des utilisateurs et les tensions entre l'expérience des utilisateurs (par exemple le délai d'écoulement) et les paramètres du réseau. densité).

Au chapitre 5, nous étudions la consommation d'énergie des réseaux placés au hasard. Nous insistons principalement sur la façon dont la consommation d'énergie du réseau évolue par rapport à la densité de la BS. De plus, nous fournissons des informations sur le compromis entre la consommation d'énergie et les retards des utilisateurs. De plus, dans ce chapitre nous dérivons une règle simple qui indique que la diminution de la BS conduit à une diminution de la consommation d'énergie. En outre, nous considérons le cas où la densité des utilisateurs change radicalement dans une zone donnée, et nous répondons à la question, comment la densité de BS devrait-elle s'adapter à la nouvelle situation.

Le chapitre 6 analyse les performances des réseaux hétérogènes orthogonaux (HetNet) fournissant des métriques de niveau de flux pour de tels systèmes (délai, charge, probabilité encombrée). Nous modélisons mathématiquement les critères d'association d'utilisateurs populaires, tels que l'association à vide, Max-SINR et Min-Delay, et nous les appliquons au cas d'un scénario HetNet populaire à deux niveaux, basé sur LTE et WiFi, afin de comprendre leurs différences de performance.

Enfin, nous concluons nos constatations et discutons de l'orientation future de la recherche au chapitre 7.

8.2 Chapitre 2: Contexte Mathématique

Dans cette thèse, deux outils mathématiques principaux ont été utilisés, la géométrie stochastique et la théorie de la file d'attente.

8.2.1 Stochastique

La géométrie stochastique (connue comme probabilité géométrique) étudie des modèles géométriques complexes afin de générer des modèles mathématiques appropriés et des outils statistiques utiles. Initialement, la géométrie stochastique considérait les problèmes du nombre fini d'objets placés au hasard. La théorie moderne initiée par D. G. Kendall, K. Krickeberg et R. Miles considère des distributions plus complexes et suppose généralement que le modèle est répandu dans le plan infini.

Il vise également à décrire et à modéliser mathématiquement une collection aléatoire de points, en une, deux, trois dimensions ou plus. De plus, il a l'intention d'étudier les propriétés statistiques du modèle susmentionné et de dériver des moyennes statistiques sur toutes les réalisations possibles d'une telle collection aléatoire.

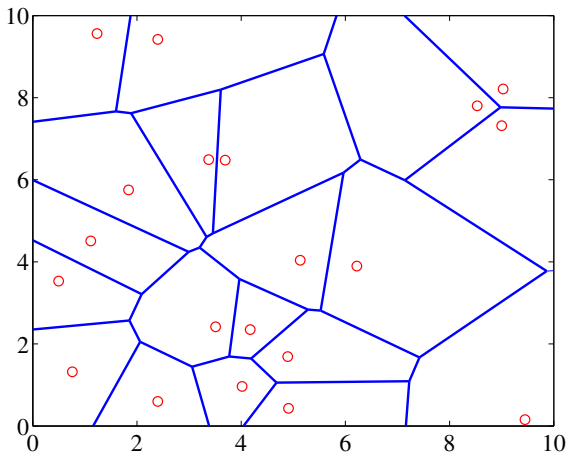


Figure 8.1 – Les cycles rouges représentent les points et les lignes bleues les régions correspondantes de Voronoï

Dans cette thèse, nous utilisons les processus ponctuels pour modéliser la distribution des BS ou l'emplacement des utilisateurs dans le réseau 8.1. Cette hypothèse nous offre la possibilité de caractériser la performance d'un réseau sans supposer une réalisation spécifique et de tirer des conclusions plus générales.

Les définitions de base de la géométrie stochastique ont été présentées dans cette section.

8.2.2 La théorie des files d'attente

Le premier article sur ce que nous appelons maintenant la théorie des files d'attente n'a été publié par Agner Krarup Erlang qu'en 1909. De nos jours, la théorie des files d'attente apporte des

solutions dans presque tous les domaines: sciences, physique, informatique, électrotechnique, biologie, etc. Le concept de base des problèmes de la théorie des files d'attente ainsi que la notation habituelle évitant les définitions et les dérivations mathématiques sont présentés dans ce chapitre. Notre but est d'aider le lecteur non-versé à suivre le reste de notre travail sans avoir à se livrer à la théorie des files d'attente, mais nous le recommandons fortement [17] .Pour cette raison, un exemple d'introduction et les définitions de base de la théorie des files d'attente sont présenté dans cette section. En outre, en appliquant la théorie de la file d'attente aux télécommunications sans fil, nous avons réduit la quantité totale d'informations.

Enfin, nous avons présenté les travaux connexes dans le domaine des communications sans fil avec l'intention de couvrir tous les problèmes concernant la théorie des files d'attente et la géométrie stochastique mais de rendre cette thèse aussi cohérente que possible et d'aider les lecteurs avec ces sujets, pour comprendre plus facilement notre travail.

8.3 Chapitre 3: Modélisation des couches PHY et MAC des LAT et WiFi RAT

La portée de ce chapitre est la modélisation appropriée des couches PHY et MAC de deux technologies d'accès radio (RAT), LTE et WiFI largement utilisées. Cela étant, nous avons deux tâches principales, une pour chaque couche (une pour PHY et une pour MAC).

- Pour faire correspondre le SINR de l'utilisateur avec le débit de données correspondant. Cette tâche nécessite deux étapes internes: i) spécifier le seuil SINR de chaque MCS, ii) spécifier le débit de chaque MCS.
- Pour répondre, comment les ressources sont allouées avec la présence d'autres utilisateurs, et comment la performance globale des systèmes dépend de la distribution des tarifs des utilisateurs. En d'autres termes, nous devrions modéliser chaque planificateur RAT avec un système de mise en file d'attente approprié afin de pouvoir analyser le comportement dynamique d'une station de base en termes de charge entrante.

Nous supposons que l'eNodeB à $20M$ avec une seule antenne et un point d'accès à un seul flux (AP) de $802,11$ US fonctionnent tous les deux avec une bande passante de $20M$. Ceci est un scénario de base, mais nos résultats pourraient facilement être étendus à d'autres cas avec une abstraction appropriée.

8.3.1 Modélisation PHY

Les RAT réels ne fournissent pas un moyen élégant de calculer le taux de l'utilisateur, il est donc courant, lors de l'analyse des réseaux sans fil, d'utiliser le théorème de Shannon, car il constitue une approche plus simplifiée. Lorsqu'un seul réseau est analysé, cette hypothèse n'affecte pas la validité des résultats qualitatifs. Cependant, dans le cas des HetNets modernes, et en particulier lorsque les réseaux hétérogènes fonctionnent avec des RAT différents, cette hypothèse ne tient pas. Le taux de différentes RAT de l'utilisateur ne varie pas de la même manière, par rapport à SINR.

Le LTE offre en moyenne 37%, plus près de Shannon que du WiFi, à leur portée SINR d'exploitation commune. Ainsi, pour ces HetNets, si les taux des deux réseaux sont modélisés selon le théorème de Shannon, le WiFi sera surestimé par rapport à LTE.

8.3.2 Modélisation MAC

Lorsque plus d'un utilisateur est servi en parallèle par une BS, la BS fonctionne comme un *système de files d'attente*. Le taux de service dépend du nombre d'utilisateurs associés et de leur SINR (charge BS). De plus, le taux de service dépend également du planificateur centralisé (par exemple, dans le cas de 3G / 4G) ou du protocole de contrôle d'accès aux médias distribués (MAC) (dans le cas du WiFi) qui décident comment les ressources disponibles seront réparties entre les utilisateurs. Bien qu'il existe un certain nombre d'algorithmes d'ordonnancement différents, la majorité d'entre eux essaie d'allouer les ressources disponibles entre des flux concurrents (par exemple des blocs de ressources LTE, un canal WiFi) d'une manière équitable ou proportionnelle.

Pour comprendre les blocs de ressources LTE, nous pouvons supposer que la station de base alloue la même quantité de ressources à tous les flux, et ils sont servis simultanément, par exemple, avec un algorithme de type TDMA à tour de rôle. Si la tranche de temps de service est petite (par exemple, de taille de paquet) comparée à la taille totale d'un flux, les performances de niveau de flux à cette BS peuvent être approximées par un système de partage de processeur $M/G/1$, comme indiqué dans la Fig. 8.2. Ce modèle a déjà été utilisé pour analyser les performances $3G/3G+BS$ [36,37]. Alors que chaque flux partage le canal pendant le même laps de temps (d'où le terme «ressource juste»), il peut transmettre pendant cette période à un débit différent, en fonction de son SINR et du MCS résultant (d'où le «multi-classe» un service).

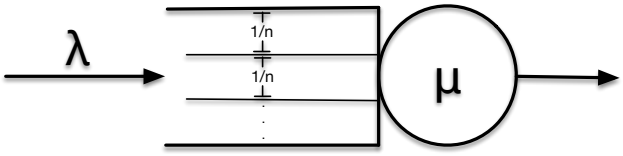


Figure 8.2 – M/G/1/PS Resource Fair

Les ordonnanceurs LTE sont très complexes, allouant des flux concurrents (ressources et temps) (Resource Blocks), prenant éventuellement en compte le backlog de chaque flux et priorité de flux, et essayant de tirer parti des variations instantanées de SINR dans le temps et la fréquence. diversité multi-utilisateur [49]. Alors qu'un grand nombre d'algorithmes ont été proposés (voir par exemple [53] pour une enquête approfondie), en l'absence de trafic prioritaire spécial, la plupart des planificateurs implémentés conduisent à une allocation proportionnellement équitable entre les flux [49] et peut également être approchée par une file d'attente PS multi-classes $M/G/1$ similaire.

8.4 Chapitre 4: Analyse des performances du réseau à un seul niveau

Dans ce chapitre, nous utilisons les résultats d'une seule modélisation BS du chapitre 3 et nous développons un modèle analytique flexible et précis de grands réseaux à un seul niveau avec placement BS aléatoire, afin de comprendre l'impact de paramètres de réseau clés comme la densité BS et la charge sur les performances du réseau. L'objectif principal est de comprendre la dynamique du niveau de flux d'un tel système, en supposant que les utilisateurs non saturés et en étudiant les statistiques d'encombrement pour les BS et le délai par flux. Pour y parvenir, nous basons notre analyse sur deux outils principaux: (a) la géométrie stochastique, pour comprendre l'impact des cartes de couverture et aléatoire aléatoires et (b) la théorie des files d'attente, pour modéliser la concurrence entre flux simultanés dans la même BS. Notre modèle est ensuite appliqué aux technologies d'accès radio populaires, telles que LTE et WIFi. Nos résultats fournissent des aperçus qualitatifs et quantitatifs intéressants sur les performances de ces réseaux et seront utilisés dans les chapitres suivants afin d'étudier la consommation d'énergie et le compromis de performance au niveau des flux ainsi que la performance des réseaux multi-tiers.

Plus précisément, la tendance des réseaux modernes est de se densifier, de se placer de manière irrégulière et de devenir plus hétérogène, en raison du déploiement souvent non planifié et incrémental de nouvelles BS (petites cellules). Par conséquent, l'analyse de tels réseaux, par

exemple pour la comparaison de protocoles ou la planification de réseau, devient de plus en plus difficile. Les métriques habituellement considérées dans ces analyses, comme SINR ou capacité, échouent souvent à capturer l'expérience utilisateur réelle, car les performances au niveau du flux (délai, probabilité d'encombrement, etc.) dépendent fortement de la charge du réseau et pas seulement des conditions du canal [9, 37]. Une meilleure métrique est la latence, qui est l'un des principaux indicateurs de performance des technologies 5G [10, 11].

Ainsi, nous présentons dans ce chapitre un modèle flexible et précis qui analyse les performances des réseaux placés au hasard, afin de comprendre l'impact des paramètres réseau importants (densité BS, charge) sur les performances du réseau. Notre modèle se compose de stations de base localisées au hasard ainsi que des utilisateurs placés au hasard. Les utilisateurs sont supposés être non saturés, générant de manière aléatoire des demandes pour de nouveaux téléchargements de fichiers - flux de tailles variables et ils perçoivent les performances en termes de délai moyen pour terminer un tel téléchargement.

Notre analyse est basée sur la combinaison de deux outils théoriques clés qui ont récemment fourni de nombreuses informations sur les performances des réseaux cellulaires: (i) Nous utilisons la théorie des files d'attente pour modéliser les performances de l'arrivée et du service du flux dynamique via le planificateur respectif , au niveau d'une seule BS; (ii) Nous utilisons la *géométrie stochastique*, afin de comprendre l'impact du hasard topologique et de l'interaction / compétition entre les BS au niveau du réseau, afin de dériver des statistiques sur le *nombre d'utilisateurs associés à une station de base* , et les *schémas de modulation et de codage (MCS)* offerts à chaque BS. Ces deux grandeurs servent d'entrées clés au modèle de file d'attente BS: le premier pour définir l'intensité totale du trafic (en termes d'arrivées de flux) qu'une BS donnée doit desservir, et la seconde pour définir le débit moyen de service (en termes de débit départs) qu'un BS est capable d'offrir.

Il existe un certain nombre d'études qui examinent les performances d'un réseau en utilisant des outils de géométrie stochastique: [58] fournit la distribution des zones de couverture et [20] qui dérive la distribution de l'interférence en supposant que toutes les BS voisines sont saturés. De plus, la dynamique des réseaux cellulaires a été étudiée dans [**?**, 34, 36, 37, 46], certains d'entre eux se concentrant sur l'efficacité spectrale et le débit instantané BS, tandis que les autres supposent des topologies cellulaires simples (p.ex., réseaux linéaires ou petites topologies hexagonales). Par rapport à ces travaux connexes, à notre connaissance, il s'agit du premier travail qui considère conjointement la géométrie stochastique du réseau et la dynamique des flux. Pour résumer, les principales contributions qui ont été présentées dans ce chapitre sont:

I) Nous présentons un nouveau résultat analytique dérivant la fonction de masse de probabilité (pmf) de la cardinalité des utilisateurs à une BS arbitraire, si les deux, les utilisateurs et les BS sont distribués en tant que processus de point de Poisson homogène;

II) Nous proposons un modèle analytique qui capture les performances des couches physiques et MAC, fournissant des statistiques pour les cartes de couverture et les distributions MCS, ainsi que les performances au niveau du flux perçues par l'utilisateur (délai) et l'opérateur réseau (probabilité d'encombrement);

III) Nous dérivons un modèle semi-analytique qui calcule la probabilité de couverture d'un réseau placé au hasard, compte tenu du fait que les BS voisines ne sont pas complètement chargées (non saturées) et créent ainsi une interférence dynamique proportionnelle à leur charge.

Le reste du chapitre est organisé comme décrit ci-dessous. Dans la section 4.2, nous modélisons les performances au niveau BS. Dans la section 4.3, nous modélisons la couche PHY. Dans la section 4.4, nous dérivons la distribution de cardinalité des utilisateurs pour notre

topologie et nous calculons le taux d'arrivée. La section 4.5 présente les étapes afin de spécifier le taux de service, qui inclut à la fois des formules analytiques pures et des détails techniques pour chacune des RAT choisies. La section 4.6, valide notre modèle théorique et analyse les réseaux d'intérêt. La section 4.7 présente les étapes futures de notre travail. Plus précisément,

pour modéliser les performances au niveau BS, nous avons supposé que chaque BS subissait une charge de trafic dynamique et nous avons étudié les performances à *flux-niveau* et nous avons énoncé nos hypothèses concernant une seule BS choisie aléatoirement. De plus, notre objectif était de comprendre les performances du réseau en considérant *stabilité*, *utilisation* et *par délai de flux*. Pour la modélisation PHY, nous avons initialement énoncé nos hypothèses sur la topologie du réseau et le modèle de couche physique et nous sommes arrivés à la conclusion que la puissance reçue est monotone par rapport à la distance, donc les zones de couverture pourraient être représentées par les régions Voronoï. De plus, nous avons dérivé la distribution de cardinalité des utilisateurs pour notre topologie et nous avons calculé le taux d'arrivée en considérant la pmf de la cardinalité des utilisateurs pour une BS arbitraire, $f_N(n)$, qui détermine le trafic total d'entrée BS. Il est observable que la taille d'une cellule arbitraire est une variable aléatoire, en fonction de la topologie BS aléatoire, et le nombre d'utilisateurs donné une taille de cellule spécifique est également une variable aléatoire. La preuve du théorème suivant ainsi qu'un résultat asymptotique utile et précis peuvent être trouvés dans l'Annexe A ou dans notre rapport technique [62].

Pour spécifier le débit maximal qu'un utilisateur peut recevoir de la station de base à laquelle il est associé, étant donné un BLER désiré, nous avons dérivé la distribution de débit $f_R(r)$ afin de calculer le taux de service μ en termes de flux / sec pour la moyenne BS. De plus, une BS à proximité pourrait ne pas interférer si elle ne transmet pas à ce moment-là, ce qui complique encore l'analyse. Pour cette raison, nous avons considéré un scénario " saturé " où les BS interférents sont supposés toujours actifs et interférents. Nous avons considéré le cas de l'interférence basée sur la charge, où une BS interfère seulement si elle est actuellement *active* au service d'au moins un utilisateur.

Pour les deux cas de brouillage (saturés, basés sur la charge), afin de calculer la distribution de débit $f_R(r)$, nous avons utilisé la probabilité de couverture du réseau et le seuil SINR pour chaque MCS. La probabilité de couverture est la probabilité que SINR d'un utilisateur arbitraire soit supérieur à un seuil donné τ

$$p_c(\tau) = P\left[SINR > \tau | r\right] \ . \tag{8.1}$$

Basé sur la discussion précédente, la fonction de masse de probabilité de MCS est donnée par:

$$f_{\mathrm{MCS}}(mcs_i) = p_c(\tau_i) - p_c(\tau_{(i+1)}) \ . \tag{8.2}$$

En d'autres termes, la probabilité qu'un utilisateur opère avec mcs_i est égale à la probabilité que son SINR soit supérieur au seuil de mcs_i moins la probabilité que son SINR soit supérieur au seuil de mcs_{i+1} (car alors, il opère dans un MCS supérieur).

Pour les performances au niveau du flux (toujours ON), nous avons supposé à nouveau que les BS et les utilisateurs sont distribués selon des PPP homogènes indépendants et en supposant que le seuil SINR τ_i pour chaque MCS (mcs_i), pmf du MCS $f_{MCS}(mcs)$ peut être obtenu à l'équation (8.2) à travers la probabilité de couverture.

$$f_{\mathrm{MCS}}(mcs_i) = p_c\left(\tau_i, \lambda, \alpha\right) - p_c\left(\tau_{(i+1)}, \lambda, \alpha\right) \ . \tag{8.3}$$

Compte tenu du MCS, le débit réel peut être facilement calculé sur la base de la bande passante totale du système en question. L'existence de plusieurs ports d'antenne et de couches MIMO résultantes peut facilement être ajoutée dans ce calcul. De même pour les transporteurs indépendants, en dérivant les MCS respectifs pour chacun.

Enfin, en prenant en compte que $N_{max} = \mu/\lambda_f$, nous avons observé que la probabilité de couverture dépend du taux de service μ. Ainsi, la distribution MCS dépend aussi du μ. D'un autre côté, μ dépend de la distribution MCS.

En raison des dépendances mentionnées ci-dessus, nous réécrivons Eq. (4.1) à sa forme correcte:

$$\mu = \left(\sum_{mcs_i} \frac{f_R(mcs_i|\mu) \cdot s}{r(mcs_i)} \right)^{-1} \ . \tag{8.4}$$

Les paramètres du modèle, pour le reste de la section de simulation, sont résumés comme suit: (i) taille de flux moyenne de 5 Mbits, (ii) pathloss $\alpha = 4$, (iii) bruit thermique $\sigma^2 = -100$ dBm (iv) $BW_{LTE} = BW_{WiFi} = 20$ MHz, (v) une antenne par eNodeB et un flux spatial par point d'accès WiFi.

Table 8.1 résume les paramètres du modèle du simulateur.

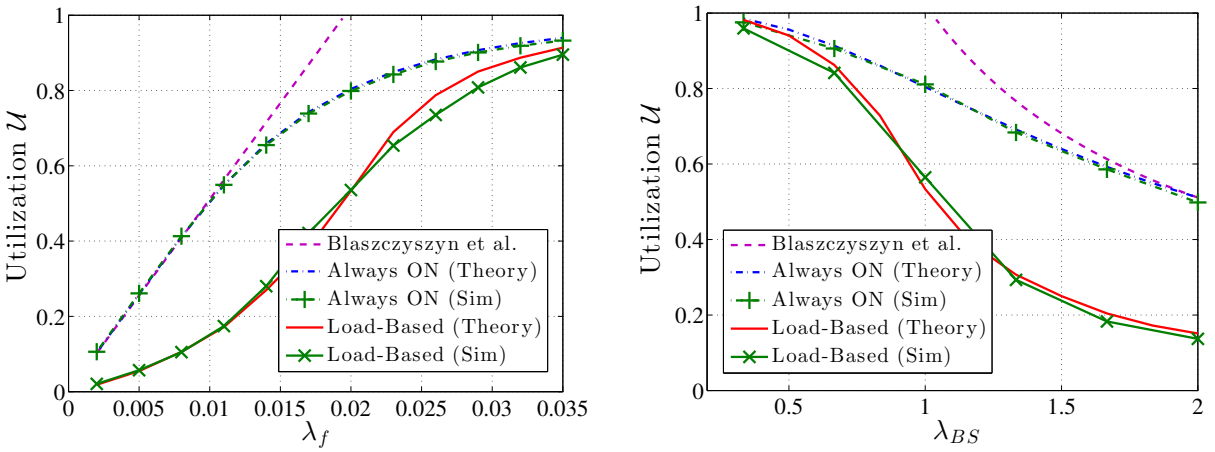
Table 8.1 – Paramètres du modèle

Densité LTE	$\lambda_{LTE} = 1$
Densité Wifi	$\lambda_{WiFi} = 1$
Densité utilisateurs	$\lambda_u = 100$
Distribution de la taille du flux	Generique
Taille moyenne du débit	5 Mbytes
α	4
BW_{LTE}	20MHz
BW_{WiFi}	20MHz
σ^2	-100dBm
# d'antennes par eNodeB	1

Nous devrions mentionner que si le bruit thermique est beaucoup plus petit que l'interférence, la valeur de P_{tx} n'affecte pas les résultats, comme indiqué dans [20].

Pour la validation du modèle et l'analyse du réseau, nous l'avons fait dans un premier temps pour un seul niveau dans les scénarios d'interférence à la fois saturés et basés sur la charge. De plus, nous comparons notre méthode avec l'approche analytique de [45] qui fournit des résultats de forme fermée sur la charge du réseau pour le cas saturé. Nous rappelons au lecteur que framework [45] suppose que la distribution MCS est en quelque sorte connue et nous comblons cet écart en utilisant la distribution MCS telle que calculée dans notre cadre. W.l.o.g un réseau LTE est considéré à cette fin. Les paramètres de performance des scénarios simulés utilisés pour

la comparaison sont (i) l'utilisation du réseau [1] et le retard de flux moyen de la BS médiane. Ce dernier est calculé en calculant le retard moyen pour chaque BS dans la simulation, puis en prenant la médiane parmi les BS. Nous choisissons la médiane simulée plutôt que la moyenne, car cette dernière croît à l'infini même si une seule BS est congestionnée.



a: Load ρ w.r.t flow density λ_f, BS density $\lambda_{BS} = 1$ b: Load ρ w.r.t BS density λ_{BS}, flow density $\lambda_f = 0.02$

Figure 8.3 – Comparaison de notre prédiction théorique et des résultats du simulateur de paquets pour les deux cas d'interférence (Always ON et Load-based) dans le cas d'un réseau LTE à un seul niveau. Avec la ligne pointillée violette, nous présentons la prédiction en utilisant le cadre de Blaszczyszyn.

Notre simulateur de paquets génère des BS et des utilisateurs placés aléatoirement sur une grande surface avec des densités données (λ_{BS}, λ_u). Les utilisateurs sont associés à la BS la plus proche et génèrent des flux selon une distribution de Poisson de densité λ_f et de taille de flux moyen $s = 5$ Mbits (625 Kbytes). Les flux sont transmis à la BS correspondante qui est modélisée comme une multi-classe M/G/1/PS. Le taux de service de chaque flux pour chaque temps est calculé via la relation SINR-MCS pour LTE. Nous considérerons deux scénarios d'interférence: (1) toujours ON, où toutes les BS voisines contribuent à l'interférence, (2) cas basé sur la charge, où nous calculons l'interférence en ne prenant en compte que les stations de base qui sont activées à ce moment quantum (correspondant à la section 4.5.2). Nous considérons en outre seulement les utilisateurs dont le SINR est supérieur au seuil du MCS le plus bas pour le cas toujours ON (sinon un utilisateur connecté dans un quantum pourrait être en dehors de la couverture dans le suivant).

Fig. 8.3 (a) et (b), présente la charge moyenne ρ (voir Eq. (4.3)) du système par rapport à λ_f et λ_{BS} respectivement, pour les deux scénarios $\lambda_u = 200$. Trois commentaires généraux de ces graphiques sont: (i) Le cadre de Blaszczyszyn *et al*, en raison de l'approximation de la valeur moyenne, est capable de prédire avec précision la performance du système seulement quand il est sous-utilisé. L'erreur de prédiction dépend de la probabilité de congestion de BS, qui pour

[1] Il s'avère que la probabilité d'encombrement, le retard moyen, le brouillage, etc. dépendent plus de l'utilisation que de la charge.

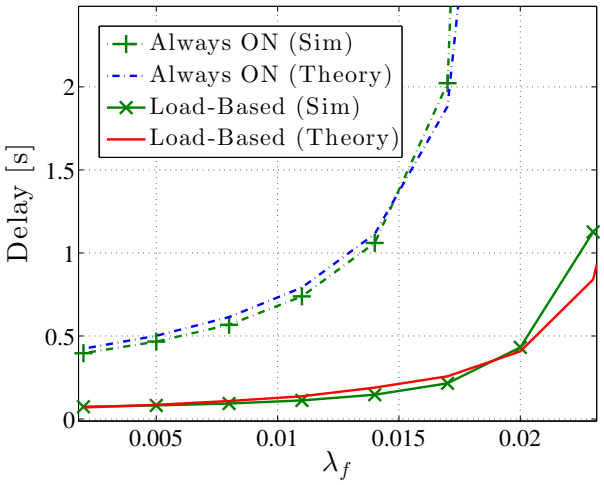


Figure 8.4 – Comparaison de notre prédiction théorique et des résultats du simulateur de paquets pour les deux cas d'interférence (Always ON et Load-based) dans le cas d'un réseau LTE à un seul niveau. Retarder la performance par rapport à la densité de flux λ_f, BS density $\lambda_{BS} = 1$

les faibles charges est $P(\rho > 1) \approx 0$. L'analyse de la valeur moyenne échoue essentiellement en raison de l'inégalité de Jensen, en utilisant directement la charge moyenne et en faisant implicitement la moyenne de certaines BS encombrées (avec la charge $\rho > 1$). En raison de la concavité de la fonction, l'utilisation s'avère supérieure à l'utilisation moyenne d'une cellule. Par exemple, dans Fig. 8.3 (a) pour un $\lambda_f = 0.2$ l'erreur de prédiction d'utilisation selon [45] est 20% et notre modèle est 0.5%, à peu près 40 fois moins. Alors que l'utilisation est une métrique relativement simple, il est relativement facile de voir que l'analyse de valeur moyenne peut avoir un impact tout aussi important (sinon plus) sur le retard, en particulier dans le cas d'interférence basée sur la charge. Clairement, la quantité d'interférence qu'une BS voisine contribue dépend du pourcentage de temps pendant lequel elle est active, c'est-à-dire de son utilisation. Par conséquent, une surestimation de cette utilisation surestimera le brouillage voisin et sous-estimera les débits de service respectifs (des files d'attente couplées), ce qui ne permettra pas davantage de prévoir les retards. Par conséquent, une analyse de valeur moyenne comme celle utilisée dans [45] peut être considérée comme une approximation de premier ordre utile pour les faibles charges (et les réseaux relativement grands) seulement. (ii) Nos deux résultats théoriques correspondent assez bien aux résultats de la simulation. (iii) L'écart entre les scénarios de brouillage toujours ON et load-based est extrêmement élevé, ce qui souligne l'importance de ces derniers.

Dans la figure 8.3 (a), pour $\lambda_f = 0.02$ la prédiction toujours ON est que le réseau est chargé à 70%au lieu de 30% de la charge. Cela signifie que le réseau pourrait être beaucoup plus robuste en ce qui concerne le trafic de données que les études qui supposent que les stations de base saturées prédisent.
Dans la figure 8.3 (b), pour une densité élevée de BS toujours sur le modèle prédit 50% réseau utilisé, tout en charge seulement 15%. L'écart entre la prédiction toujours active et la prédiction basée sur la charge augmente en fonction de la densité du réseau. Cela arrive parce que l'analyse saturée est capable de capturer seulement le gain provenant du fait qu'une BS "arbitraire" sert

en moyenne moins d'utilisateurs sur un réseau plus dense, mais pas le gain provenant du fait que les BS environnantes seront moins chargées, et causera donc moins d'interférences. Ainsi, le gain pour déployer un réseau plus dense est beaucoup plus élevé que prévu par une analyse qui ne tient pas compte de l'interférence dépendante de la charge.

Fig. 8.4, montre le retard médian du simulateur de paquets ainsi que les prédictions théoriques pour les cas saturés et basés sur la charge. Encore une fois, on peut voir que les prédictions théoriques sont assez précises, et que toujours l'interférence sur-surestime le retard par ordre de grandeur.

De ce qui précède, nous pouvons supposer que l'écart entre le modèle que nous présentons pour les performances de grands réseaux placés au hasard en supposant une BS saturée et le modèle semi-analytique pour le cas plus réaliste d'interférence basée sur la charge pourrait être énorme. une sous-estimation de la performance du réseau. Si les BS n'interfèrent pas tout le temps, le réseau est beaucoup plus robuste à la charge entrante totale et le gain de déploiement plus dense est beaucoup plus élevé que les prédictions de cas saturés. En outre, les caractéristiques PHY du réseau sont meilleures, ce qui dépend de la charge.

8.5 Chapitre 5: Efficacité énergétique et compromis QoE de l'utilisateur

Au chapitre 5, nous avons étudié la consommation d'énergie des réseaux placés au hasard. La consommation d'énergie est l'une des principales préoccupations des réseaux modernes de petites cellules denses. L'un des concepts clés pour améliorer l'efficacité énergétique d'un réseau dense consiste à éteindre une partie de ses BS lorsqu'ils sont inactifs ou faiblement chargés, car même dans ce cas, une quantité considérable d'énergie est consommée. Dans ce chapitre, nous utilisons le cadre analytique du chapitre 4 pour analyser analytiquement le compromis entre l'efficacité énergétique et l'expérience utilisateur, qui est mesurée en termes de retard des utilisateurs en supposant un modèle de trafic non saturé. Nous fournissons une performance globale du réseau en ce qui concerne la densité de BS et les idées qui peuvent être utiles en termes de conception de réseau. Notre analyse est basée sur a) le modèle de consommation d'énergie linéaire d'une BS b) la géométrie stochastique pour modéliser la topologie du réseau et les utilisateurs et c) la théorie de la file d'attente afin de capturer les performances au niveau du flux. Notre modèle est appliqué à la technologie d'accès radio populaire LTE mais il peut facilement être étendu à d'autres. Nos résultats fournissent des lignes directrices et des limites qui permettent de prédire l'efficacité énergétique du réseau conçu.

Une façon d'augmenter la capacité de la zone dans les réseaux cellulaires est d'ajouter plus de BS un processus également connu sous le nom de densification. En particulier, l'ajout de petites cellules devrait être l'une des solutions clés pour lutter contre l'augmentation exponentielle du trafic sur les prochaines années [64–66]. D'autre part, à mesure que le trafic de données et la densité des réseaux augmentent, la consommation d'énergie devient de plus en plus cruciale tant pour l'environnement (réduction de l'empreinte carbone) que pour des raisons économiques [67].

Des études ont montré qu'environ 50% - 70% de la consommation totale d'énergie des télécommunications est en cours sur BSs [14]. Une quantité considérable d'énergie est consommée sur la BS (pour rester allumé ou se refroidir) malgré le peu ou pas de trafic [13]. En outre, si l'on considère les zones industrielles, les routes commerciales, etc., ces réseaux denses sont

déployés pour desservir le nombre élevé d'utilisateurs connectés aux heures de pointe, mais pour le reste de la journée, le réseau devient sous-utilisé. Par conséquent, l'un des concepts clés pour la réduction d'énergie est de désactiver (ou mettre en mode veille) de telles stations de base. Depending on the radio access technology used, this can be achieved by various sleeping techniques [68]. In 3GPP LTE-Advanced (release 10) for example, the carrier aggregation feature can be used to steer traffic to another cell and to power off cells with no traffic. This feature has been improved in release 12 using the discovery reference signal, which is sent by sleeping cells only in configurable intervals [69].

Le problème de l'activation / désactivation des BS a été étudié dans divers travaux. Dans [40] les auteurs résolvent le problème d'optimisation de l'association des utilisateurs en tenant compte de la consommation d'énergie ainsi que des performances au niveau du flux (délai), mais cette analyse ne fournit pas de résultats analytiques pour la performance globale du réseau. Dans [39], les auteurs adoptent une approche analytique de la file d'attente pour étudier l'impact de la désactivation d'une BS sur des BS voisins pour différents modèles de trafic, mais ils effectuent seulement une analyse de performance réseau par des simulations numériques. Dans [27] les auteurs suivent une approche de la géométrie stochastique afin de fournir les performances du réseau tout en désactivant les BS, ce travail suppose des BS saturées et n'influe pas sur les performances au niveau du flux du réseau.

Notre analyse est basée sur un modèle de coût de l'énergie utilisé commun [13] combiné avec notre cadre récemment développé qui analyse les performances au niveau du flux d'un réseau placé au hasard [70, 71]. Nous combinons des outils de la géométrie stochastique et de la théorie des files d'attente pour analyser la performance globale du réseau et fournir des informations sur le compromis entre la QoE des utilisateurs (mesurée en termes de délai) et l'efficacité énergétique sans supposer des BS saturées.

Nous appliquons nos résultats pour la technologie d'accès radio LTE et principalement au cas de décroissance de la densité du réseau en éteignant une partie du réseau, mais le même cadre peut analyser le cas de l'augmentation de la densité du réseau en ajoutant BS (densifier). Afin de fournir des expressions proches et d'éviter les simulateurs complexes au niveau du système, nous supposons que les BS à désactiver sont sélectionnées aléatoirement, mais comme nous le verrons, cette approche n'est pas si éloignée des critères plus sophistiqués.

En résumé, nos contributions sont:

- Nous dérivons des formules analytiques et semi-analytiques pour étudier le compromis entre l'efficacité énergétique et le retard des utilisateurs. Il y a des cas où il est possible d'avoir un gain énergétique important et d'autre part une réduction abordable des performances des utilisateurs.

- Nous comparons notre hypothèse si nous avons choisi au hasard les BS qui seront désactivées avec des simulations de critères plus sophistiqués tels que le nombre minimum d'utilisateurs associés et nous verrons que les deux approches convergent à mesure que la densité du réseau augmente.

- Pendant les heures creuses, le nombre d'utilisateurs est considérablement réduit et le réseau devient sous-utilisé. Pour ce scénario, nous dérivons la quantité maximale de BS qui peut être désactivée sans affecter les performances des utilisateurs restants. En outre, nous fournissons une règle simple dans laquelle les conditions de cette réduction BS conduit

à un gain d'énergie (il est possible que la consommation d'énergie augmente malgré la désactivation de certains BS).

Le reste de ce chapitre est organisé comme suit: Section 5.2 présente notre modèle de système, incluant toutes nos hypothèses sur la topologie, le modèle de propagation, l' ordonnanceur, etc. La section 5.3 dérive la distribution MCS pour l'arbitraire BS. La section 5.4 présente le modèle du coût énergétique de la BS et le modifie en fonction de la métrique, de l'énergie par unité de surface, du scénario de réduction de la densité des utilisateurs et de certains résultats théoriques. Enfin, dans la section 5.5, nous présentons quelques résultats intéressants de notre analyse de la performance au niveau du flux et de l'efficacité énergétique du réseau.

Nous devons noter que certaines de nos hypothèses ont déjà été mentionnées dans le chapitre 4, mais nous avons décidé de les présenter brièvement ici aussi, afin d'éviter l'utilisation étendue de références croisées qui rendront le chapitre confus et difficile à lire.

8.5.1 Validation

Le simulateur de niveau paquet génère à la fois des BS et des utilisateurs placés aléatoirement dans une grande surface avec des densités données (λ_{BS}, λ_u). Les utilisateurs sont associés à la BS la plus proche et génèrent des flux selon la distribution de Poisson avec la densité λ_f. Les flux sont transmis à la BS correspondante qui est modélisée comme une multi-classe M / G / 1 / PS. Le taux de service de chaque flux pour chaque temps quantique est calculé via SINR. Au moment du calcul de l'interférence, nous ne prenons en compte que les stations de base qui sont activées à ce moment (cas basé sur la charge), pour la comparaison avec l'hypothèse la plus courante dans la littérature, nous prenons également en calcul d'interférence, tous les BS (cas saturé). Afin de comparer équitablement les deux cas de brouillage, nous ne considérons que les utilisateurs dont le SINR est au moins supérieur au seuil du MCS le plus bas en cas de saturation. Nous devons mentionner que le simulateur de niveau de paquet a besoin de ressources de calcul extrêmement élevées, de sorte que la prédiction théorique est encore plus précieuse.

Nous sommes intéressés à étudier les performances évolutives d'un grand réseau placé au hasard alors que nous désactivons un pourcentage de BS afin d'améliorer l'efficacité énergétique. Dans cette section, la *densité de l'utilisateur ne change pas*, donc en désactivant les BS nous savons a priori que les performances des utilisateurs seront réduites, donc nous voulons comparer le gain d'énergie avec la réduction de performance. Dans notre analyse théorique, nous sélectionnons au hasard les BS qui seront désactivées. Comme nous l'avons mentionné précédemment, la sélection aléatoire est un scénario catastrophe, il existe des critères plus sophistiqués pour décider quelles BSs désactiver (utilisateurs associés minimum).

Les figures 8.5 et 8.6 montrent comment la performance au niveau du flux du réseau (délai moyen et charge moyenne du réseau) varie en fonction de la densité de la BS (λ_{BS}) dans quatre cas différents i) notre prédiction théorique qui suppose une interférence basée sur la charge (lb) et la sélection des BS à désactiver est aléatoire (rnd) ii) des résultats de simulation pour le cas de brouillage basé sur la charge et la sélection de Les BSs à désactiver sont aléatoires iii) les résultats de simulation supposant une interférence basée sur la charge et la sélection des BS à tourner selon le critère minimum d'utilisateurs associés (min) et iv) les résultats de simulation supposant des BS saturés (sat) et la sélection des BS à tourner off est fait au hasard. Nous pouvons obtenir trois conclusions intéressantes à partir de Figs 8.5 et de 8.6: 1) notre prédiction théorique est très précise (pour la charge et le retard) par rapport aux résultats de simulation, 2)

l'hypothèse de saturation Les BS (ce qui est courant dans les travaux de géométrie stochastique) changent totalement non seulement quantitativement mais qualitativement les performances du réseau et 3) le critère minimal des utilisateurs associés ne diffère pas beaucoup de celui aléatoire, surtout quand le réseau est très dense. négligeable. Cela signifie que pour les réseaux denses, il est préférable de désactiver BS de manière aléatoire que d'utiliser le critère d'utilisateurs associés minimum centralisé et plus complexe pour déterminer quelle BS doit être désactivée.

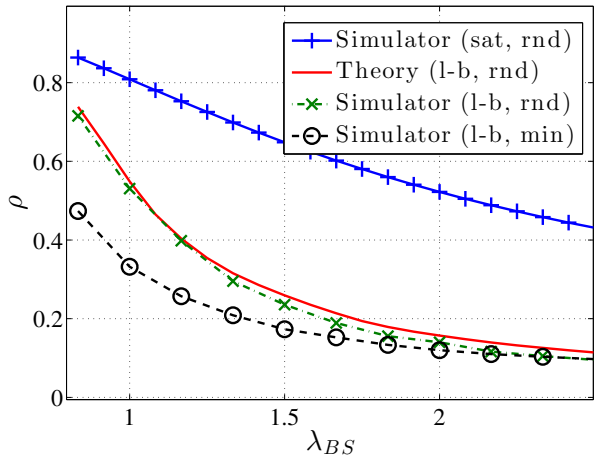


Figure 8.5 – Résultats théoriques et de simulation pour la charge moyenne du réseau. *sat* indique le saturé (Always ON) et *l-b* les cas basés sur la charge de l'interférence. *rnd* indique la sélection aléatoire et *min* le critère des utilisateurs associés minimum de désactivation de BS.

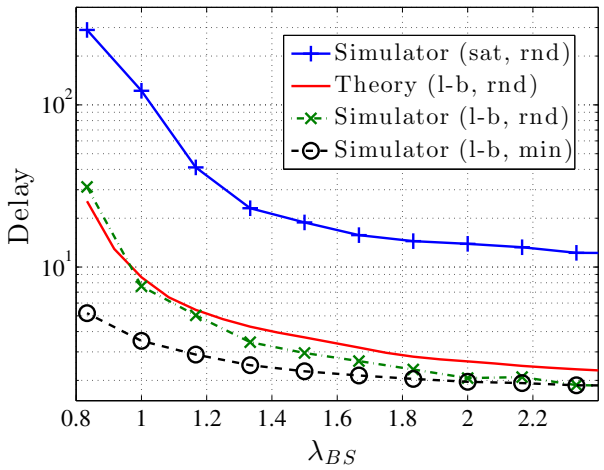


Figure 8.6 – Résultats théoriques et de simulation pour le retard de l'utilisateur moyen. *sat* indique le saturé (Always ON) et *l-b* les cas basés sur la charge de l'interférence. *rnd* indique la sélection aléatoire et *min* le critère des utilisateurs associés minimum de désactivation de BS.

8.5.2 Énergie Vs Délai

Dans cette sous-section, nous sommes intéressés à étudier le compromis entre l'efficacité énergétique et le retard des utilisateurs. En ce qui concerne l'interférence, nous supposons le cas le plus réaliste de la charge basée. Initialement, le réseau est sous-utilisé ($\rho \approx 0,1$), soit E_0 et D_0 la consommation d'énergie et le délai moyen de cet état initial. Puis, progressivement, nous désactivons une partie du réseau, nous définissons comme $\overline{E}/E_0$ le gain d'énergie relative et comme $\overline{D}/D_0$ le retard relatif, pour simplifier, nous appellerons ces mesures énergie gain et retard relatif respectivement.

La figure 8.7 et 8.8 montre le gain d'énergie par rapport au retard relatif pour différents rapports $\frac{E_{on}}{E_{op}}$. Initialement, en observant le cas de $E_{on} = E_{op}$, nous notons que pour une charge faible, la dérivée du gain d'énergie est très élevée, donc il y a la possibilité d'améliorer l'énergie sans coût de retard important. Lorsque la charge du réseau est $\rho \approx 0,5$ la dérivée diminue considérablement, donc le coût du retard est extrêmement élevé par rapport au gain d'énergie Fig. 8.7.

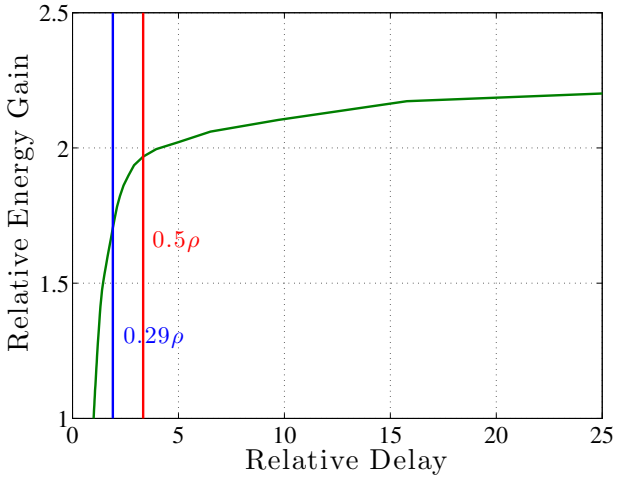


Figure 8.7 – Gain d'énergie relative et retard relatif pour le cas de $E_{on} = E_{op}$. Les deux lignes verticales indiquent le point où la charge du réseau est ρ est égale à 0,29 et 0,5 respectivement

De plus, sur la figure 8.8 il y a deux autres remarques

1. Alors que le rapport entre l'énergie constante et l'énergie opérationnelle augmente, le gain possible en désactivant la BS augmente. Si l'on considère les petites cellules traditionnelles, le fait que le terme d'énergie constante soit beaucoup plus élevé que le terme opérationnel semble irréaliste, mais dans les réseaux futurs (par exemple les drones) cela pourrait être le cas

2. Lorsque le terme d'énergie constante est inférieur à celui d'exploitation, il y a un point tournant dans la courbe de performance. Cela signifie qu'après un point, lorsque nous désactivons plus de BS, l'énergie et les performances de retard s'aggravent. Dans le cas extrême où le coût d'une BS est ON est négligeable par rapport au coût opérationnel $\frac{E_{on}}{E_{op}} \to 0$ il n'y a aucune possibilité d'amélioration de l'énergie, donc, la stratégie optimale consiste simplement à activer toutes les BS Fig. 8.9.

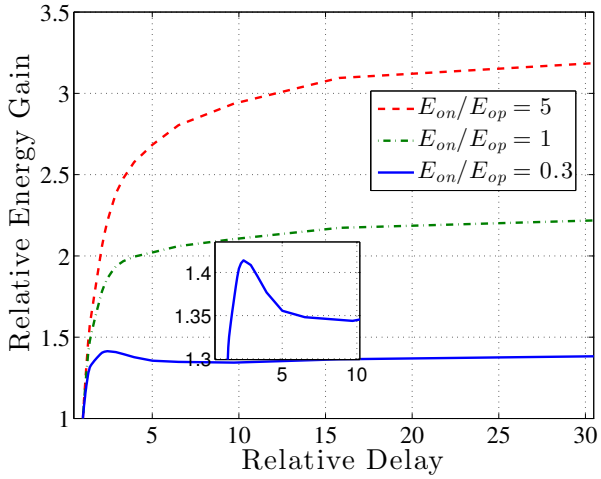


Figure 8.8 – Gain d'énergie relative et retard relatif pour le cas de $\frac{E_{on}}{E_{op}}$ ratios

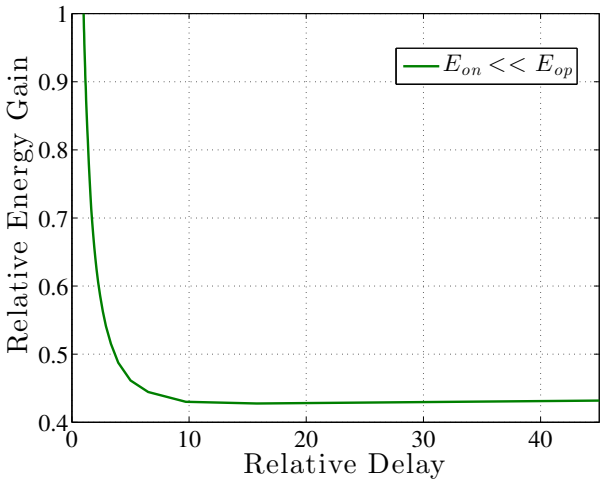
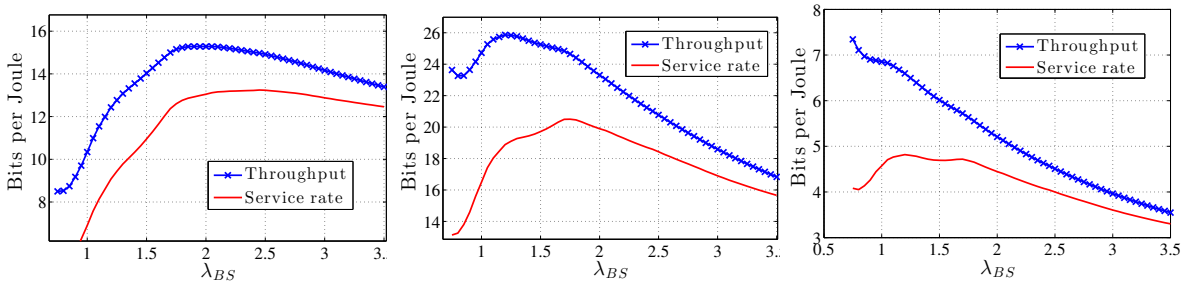


Figure 8.9 – Gain d'énergie relative et retard relatif pour le cas de $\frac{E_{on}}{E_{op}} \to 0$

Une autre mesure intéressante est la consommation d'énergie absolue divisée par le retard par rapport à la Densité BS λ_{BS}. La figure 8.10 représente la métrique mentionnée ci-dessus pour différentes valeurs du rapport $\frac{E_{on}}{E_{op}}$. Si approximativement $\frac{E_{on}}{E_{op}} > 0,1$, le rapport $\frac{\overline{E}}{Delay}$ se met à l'échelle linéairement par rapport à la Densité BS λ_{BS}. Cette propriété pourrait être une règle empirique valable pour la conception de réseau si nous considérons que les deux $\overline{E}$ et $Delay$ sont beaucoup plus complexes par rapport à λ_{BS}.



a: $\frac{E_{on}}{E_{op}} = 0.2$ b: $\frac{E_{on}}{E_{op}} = 1$ c: $\frac{E_{on}}{E_{op}} = 5$

Figure 8.11 – Bits par Joule pour différentes valeurs de ratio $\frac{E_{on}}{E_{op}}$. *Throughout* indique la moyenne statistique du taux, *Taux de service* montre la moyenne harmonique du taux tel que calculé Eq. (4.1)

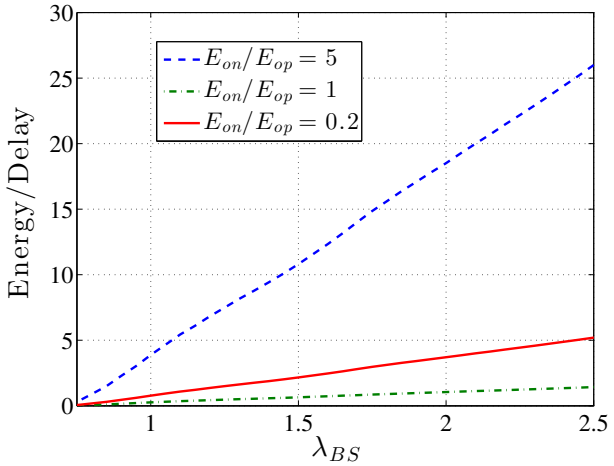


Figure 8.10 – Rapport entre la consommation d'énergie moyenne et le retard moyen de l'utilisateur $\frac{\overline{E}}{Delay}$ par rapport à la densité BS λ_{BS}

8.5.3 Bits par Joule

Une autre mesure de l'efficacité énergétique des réseaux est la quantité de bits de transmission par joule $\overline{R}/\overline{E}$, où $\overline{R}$ représente généralement le débit du système $\sum f_{MCS}(mcs) \cdot r(mcs)$. En ce qui concerne les performances au niveau du flux du réseau, le taux de service $\mu = \left(\sum_{mcs} \frac{f_{MCS}(mcs)}{r(mcs)}\right)^{-1}$ est plus représentatif. Ainsi, respectivement avec les bits par joule, nous définissons les flux par jauge métrique du système. La figure 8.11 représente des bits par joule par rapport à la densité du réseau pour différentes valeurs du ratio $\frac{E_{on}}{E_{op}}$. Quand $E_{on}E_{op}$ il y a une densité de réseau optimale qui maximise les bits par joule et une autre différence qui maximise les flux par joule. En termes de flux par joule, la densité optimale est supérieure à celle en bits par joule.

Dans ce chapitre, nous avons présenté un cadre analytique qui fournit la performance énergétique du réseau et nous avons étudié les compromis entre l'efficacité énergétique du réseau et la QoE de l'utilisateur. Nous observons que l'hypothèse la plus défavorable de sélection aléatoire dont la BS éteint les performances du réseau n'est pas bien pire que la solution plus complexe et sophistiquée d'éteindre la BS avec moins d'utilisateurs associés si le réseau déployé est dense. Nous présentons quelques résultats théoriques sur la réduction de la densité BS dans le cas de la réduction des utilisateurs et nous fournissons une règle empirique, lorsque cette réduction conduira à un gain d'énergie. En outre, si le nombre d'utilisateurs ne sera pas réduit, nous avons vu qu'il y a une capacité d'amélioration de l'énergie sans affecter la QoE d'un utilisateur lorsque l'énergie opérationnelle n'est pas beaucoup plus grande que le terme d'énergie constante. Enfin, nous présentons des mesures intéressantes sur l'efficacité énergétique.

8.6 Chapitre 6: Analyse des performances des réseaux hétérogènes multi-tiers

Les réseaux cellulaires modernes se densifient, sont moins régulièrement planifiés et de plus en plus hétérogènes, en raison des efforts déployés par les opérateurs pour faire face à un manque de données sans précédent. Cette complexité accrue rend cependant l'analyse de la performance difficile. Dans cet article, nous développons un modèle flexible et précis afin d'analyser les performances des grands réseaux cellulaires hétérogènes (HetNets), et de comprendre l'impact des paramètres clés du réseau. Ce modèle se compose de niveaux K de stations de base (BS) situées au hasard, avec des densités différentes, des puissances d'émission et des technologies d'accès radio (RAT). Notre objectif principal est de comprendre l'impact de la dynamique des niveaux de flux sur un tel système, en supposant que les utilisateurs non saturés génèrent des requêtes de téléchargement de façon aléatoire («flux»). Nous le faisons en dérivant analytiquement le retard par écoulement atteint par un tel réseau, ainsi que la probabilité d'encombrement des BS dans différents niveaux (c'est-à-dire, le pourcentage de BS qui seront surchargés). Pour ce faire, nous basons notre analyse sur deux outils principaux: (a) géométrie stochastique, pour comprendre l'impact du hasard topologique et de l'interaction intra- et inter-niveaux dans les cartes de couverture résultantes, en considérant deux modèles d'interférence réalistes; théorie, pour modéliser la concurrence entre les flux simultanés dans la même BS, pour chaque RAT. Nous appliquons notre modèle au cas d'un populaire HetNet à deux niveaux, basé sur le LTE et le WiFi, afin de mieux comprendre les différences de performance des critères d'association d'utilisateurs populaires, tels que le déchargement, l'association Max-SINR et Min-Delay. association. Nos résultats fournissent des informations qualitatives et quantitatives intéressantes sur l'impact de ces politiques d'association et des différentes intensités de trafic.

8.6.1 Introduction

Comme mentionné dans l'introduction de la thèse, HetNets est l'une des solutions les plus prometteuses pour faire face à l'augmentation exponentielle du trafic de données mobiles. Pour alléger le réseau de macrocellules surchargé, les opérateurs déploient en outre de petites cellules pour capturer le trafic dans les points chauds. Un scénario prometteur pour un tel réseau HetNet est la combinaison de cellules macro LTE avec de petites cellules WiFi. De nos jours, il

est possible d'intégrer des points d'accès WiFi dans le réseau central des systèmes cellulaires, et d'effectuer le déchargement du trafic de LTE vers le WiFi.

Les architectures HetNet offrent de nombreux avantages, mais elles conduisent également à des déploiements plus denses, irréguliers et plus hétérogènes, en raison du déploiement souvent imprévu et incrémental de nouvelles BS (à petites cellules) [72], ainsi que des accès radio potentiellement différents. Technologies (RAT). Par conséquent, l'analyse de tels réseaux, par exemple pour la comparaison de protocoles ou la planification de réseau, devient de plus en plus difficile. Tenant compte du fait que les métriques habituellement considérées dans de telles analyses, comme SINR ou capacité, échouent souvent à capturer l'expérience utilisateur réelle, parce qu'elles ne tiennent pas compte de la charge lourde des réseaux cellulaires modernes [9,37], comme nous fait dans les chapitres précédents, nous nous concentrons sur d'autres métriques telles que le délai d'utilisation et la probabilité de congestion.

Ainsi, nous développons dans ce chapitre un modèle flexible et précis pour la performance des futurs HetNets, afin de comprendre l'impact de paramètres réseau importants. Notre modèle se compose de niveaux orthogonaux de K de BSs localisés aléatoirement, avec différentes densités, puissances d'émission et RAT, ainsi que des utilisateurs placés au hasard. Comme dans les chapitres précédents, les utilisateurs sont supposés être non saturés, générant aléatoirement des demandes de nouveaux téléchargements de fichiers / flux de tailles variables, et ils perçoivent les performances en termes de délai moyen pour terminer un tel téléchargement. En d'autres termes, nous nous intéressons à la dynamique des flux ou aux performances au niveau des flux de ce réseau hétérogène. Suite à notre travail jusqu'à présent, nous modélisons les BS en tant que systèmes de files d'attente, qui planifient les flux d'utilisateurs arrivant simultanément selon le planificateur RAT respectif, et les performances liées au réseau sont mesurées en fonction de la charge stationnaire imposée à chaque BS étant congestionnée.

A partir de notre framework présenté dans le chapitre 4 qui analyse les performances au niveau des flux d'un réseau à un seul niveau, nous l'avons étendu dans le cas de plusieurs niveaux. Notre objectif est de fournir un cadre analytique qui analyse la dynamique du niveau d'écoulement dans de grands réseaux hétérogènes multi-niveaux placés au hasard et d'étudier l'impact de différents critères d'association. Par conséquent, dans ce chapitre nous utilisons notre cadre analytique pour étudier l'impact des politiques d'association d'utilisateurs populaires comme *Off-load* (tous les utilisateurs à portée d'un point d'accès WiFi sont associés au réseau WiFi), *Max-SINR* (un utilisateur est associé à la BS offre le meilleur SINR, parmi tous les niveaux), et *Min-Delay* (un utilisateur est associé au niveau qui offre la meilleure combinaison de débit et de charge afin de minimiser la moyenne delay [46]). Nos résultats fournissent des aperçus qualitatifs et quantitatifs intéressants.

Le reste du chapitre est organisé comme ci-dessous. Dans la section 6.2, nous présentons notre modèle de performance au niveau BS avec notre modèle de couche PHY. Dans la section 6.3 nous modélisons mathématiquement les règles d'association, et nous présentons les distributions MCS correspondantes. La section 6.4 examine certains scénarios d'intérêt et applique nos résultats d'analyse pour obtenir des informations. La section 6.5 présente les étapes futures de notre travail.

La plupart de nos hypothèses selon la génération de flux, le modèle de canal, la topologie ou l'association intra-tiers sont les mêmes que celles présentées dans le chapitre 4. Pour la compacité nous les présenterons aussi brièvement que possible avec nos nouvelles hypothèses pour l'environnement à plusieurs niveaux.

Pour le modèle de performance au niveau BS, nous avons supposé que chaque BS subissait

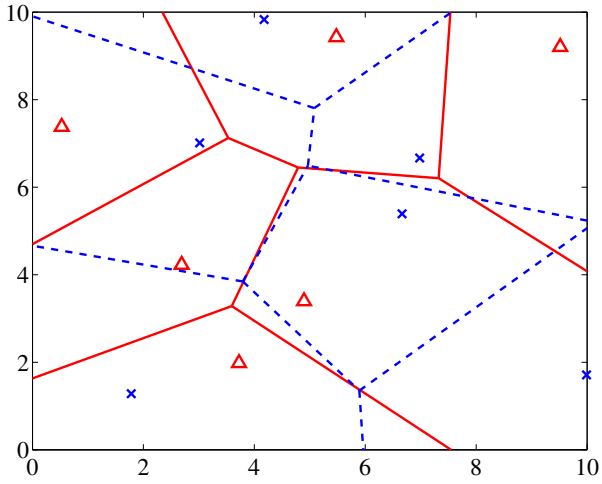


Figure 8.12 – Voronoi Tessellation exemple, 2-tiers avec BS density $\frac{6}{100m^2}$ each

une charge de trafic *dynamique* et nous avons étudié les performances à *flux-niveau* nous avons également fait des suppositions concernant une seule BS choisie aléatoirement.

De plus, nous avons modélisé chaque planificateur BS avec le bon modèle de file d'attente. Comme nous l'expliquons en détail dans le chapitre 3 pour les deux RAT de notre intérêt, les modèles de file d'attente appropriés sont:

1) Pour l' ordonnanceur LTE, il peut être modélisé comme une file d'attente de partage de processeurs M/G/1 multi-classes équitables. 2) En ce qui concerne le planificateur WiFi, nous supposons deux approches différentes i) *Approximation 1:* ressource juste planificateur, cette approximation est valide lorsque la charge BS est faible et peut être utilisée comme limite inférieure du délai, pour des valeurs de charge plus élevées . ii) *Approximation 2:* Pour les charges générales, nous pouvons modéliser le planificateur Wifi comme système de débit équitable et utiliser l'approximation d'Avrachenkov et tout. Pour plus de détails, vous pouvez revenir sur le chapitre 3, section 8.3.2.

Pour la modélisation de la couche PHY, nous avons également énoncé nos hypothèses concernant la topologie du réseau et le modèle de couche physique et nous sommes arrivés à la conclusion que la puissance reçue est monotone en ce qui concerne la distance. Notre critère est simplifié au critère de distance la plus proche, donc, les zones de couverture des BS pourraient être représentées par Voronoi Regions (Tessellations), Fig. 8.12 montre deux réseaux orthogonaux et leurs régions Voronoï, les lignes pleines correspondent à des mosaïques par rapport aux lignes $\bigtriangleup$ du réseau et des tirets sur le réseau $\times$.

Comme nous l'avons vu au chapitre 4, la distribution pmf de chaque MCS $f_{MCS}(mcs)$ dérivée par la probabilité de couverture p_c. La probabilité de couverture indique la probabilité que le SINR d'un utilisateur plus aléatoire dépasse un seuil donné (dans notre cas, c'est le seuil SINR pour chaque MCS). Nous définissons cette probabilité de couverture dans deux cas différents: i) Toujours ON Interference, où nous avons supposé que les BS interférentes transmettaient toujours ii) l'interférence de base de charge, si les BS interféraient seulement pendant la durée utile à un utilisateur.

Maintenant, en utilisant la distribution MCS $f_{MCS}(mcs)$ et la probabilité de couverture p_c

de chaque niveau, nous voulons étudier différents critères d'association et dériver la nouvelle distribution MCS et le pourcentage d'utilisateurs associés à chacun d'eux .

Dans le cas d'un réseau à plusieurs niveaux, la densité d'utilisateur λ_u et pmf de MCS $f_{MCS}(mcs)$ de chaque niveau dépendent également de la politique d'association inter-niveaux. Nous avons présenté ici comment calculer ces deux paramètres pour les schémas d'association de base considérés. Les règles d'association dans cette section indiquent le niveau auquel l'utilisateur sera associé et non la licence de sécurité. Étant donné le niveau, l'utilisateur est associé au BS le plus proche. Les lemmes que nous avons présentés peuvent facilement être étendus à plus de deux niveaux.

8.6.2 Résultats

Nous avons appliqué nos résultats analytiques pour obtenir des vues en considérant certains scénarios. Plus précisément, il est aujourd'hui possible d'intégrer des réseaux WiFi dans les réseaux de base des systèmes cellulaires, et de décharger le trafic vers les points d'accès WiFi. Dans les prochaines versions de 3GPP, une intégration plus étroite des technologies WiFi et LTE est attendue. Pour cette raison, nous choisissons un scénario RAT hétérogène composé de niveaux orthogonaux LTE et WiFi, comme étude de cas. Nous considérerons les paramètres " fixed " suivants pour les deux réseaux: (i) pathloss $\alpha = 4$, (ii) bruit thermique $\sigma^2 = -100$ dBm (iii) $BW_{LTE} = BW_{WiFi} = 20$ MHz, (iv) une antenne par eNodeB et un flux spatial par point d'accès WiFi. Enfin, nous devons mentionner que si le bruit thermique est beaucoup plus faible que l'interférence (ce qui est le cas dans notre système), la valeur de P_{tx} n'affecte pas les résultats, comme le montre [20] . Le reste des paramètres agira comme des variables, et nous discuterons de leur plage de valeurs par scénario.

Pour le réseau WiFi, un certain nombre de configurations différentes et les normes 802.11 pourraient être considérées. Le protocole WiFi traditionnel est réglé sur un canal d'environ 20 MHz. Ce canal est choisi parmi un certain nombre de canaux se chevauchant partiellement, généralement avec le critère de SINR maximum, et ce nombre de canaux diffère selon les pays. En outre, les nouvelles versions de WiFi (n / ac) ont la capacité de liaison de canal afin de fonctionner avec 40 à 160 MHz. Des largeurs de bande plus importantes pourraient également être envisagées via l'agrégation de porteuses dans LTE. Toutes ces capacités de canaux supplémentaires sont orthogonales à notre modèle et hors de notre portée, nous supposons donc pour la simplicité et l'équité que les deux réseaux fonctionnent avec 20 MHz.

Enfin, comme expliqué précédemment, l'implémentation WiFi actuelle fonctionne plus proche d'un ordonnanceur équitable. Cependant, comme mentionné dans le chapitre 3, le planificateur WiFi pourrait être modifié pour éviter le problème " anomalie WiFi " et fonctionner comme ressource équitable [55]. Nous allons donc considérer le WiFi avec les deux types d'ordonnanceurs, afin de mieux comprendre leur impact, individuellement et dans une configuration à deux niveaux.

Après avoir validé nos résultats théoriques sur la probabilité de couverture et la prédiction de l'utilisation et du retard du réseau dans le chapitre 4, nous procédons directement à un autre scénario d'intérêt afin d'obtenir des informations sur la probabilité de congestion d'une BS (la probabilité que la charge d'une BS est $\rho > 1$) et les statistiques de délai d'écoulement dans les grandes topologies aléatoires. Nous étudions tout d'abord l'impact des éléments suivants sur les performances au niveau des flux: (a) la relation MCS-SINR (qui diffère entre WiFi et LTE), (b) le planificateur (débit équitable et ressource équitable), et (c) le type d'interférence (toujours

activé ou basé sur la charge).

Nous le faisons initialement pour les systèmes à un seul niveau, avant de passer à des systèmes à plusieurs niveaux. Cela facilitera également notre discussion subséquente sur les systèmes à deux niveaux, où plusieurs facteurs affectent les performances simultanément.

De plus, nous sommes passés à des HetNets à plusieurs niveaux, ce qui est l'objectif principal de ce document. Nous sommes intéressés à comprendre l'impact de la coexistence de différents RAT dans des fréquences orthogonales. En particulier, notre objectif est de capturer l'impact de différents types de critères d'association entre différents niveaux, sur les gains de performance en introduisant un deuxième niveau. Comme mentionné précédemment, les critères d'association de niveau d'intérêt sont:

Off-load: C'est la politique la plus simple (et la plus agressive), où l'utilisateur, s'il est capable d'établir une connexion avec le réseau WiFi, le fait sans critère supplémentaire.

Max-SINR: Ici, l'utilisateur choisit de s'associer avec le niveau qui fournit le meilleur SINR, essayant ainsi d'améliorer

Min-Delay: Dans ce scénario, l'utilisateur choisit de s'associer à des critères liés à la charge afin de minimiser le délai moyen du système. Dans notre cas, le critère d'association, entre les niveaux, est l'algorithme proposé à [46].

Tout d'abord, nous examinons la politique de déchargement simple. Nous supposons que LTE est le réseau primaire et WiFi le secondaire avec la même densité. Fig. 8.13 présente deux cas différents de ce HetNet à deux niveaux. La différence entre ces deux cas concerne le programmateur WiFi: l'un était le WiFi AP fonctionnant comme un ordonnanceur «idéal», juste pour les ressources, et l'autre comme un débit équitable. De plus, pour des raisons de comparaison, nous incluons comme trame de référence un réseau LTE à un seul niveau avec double densité BS (i.e. le nombre total de BS LTE est égal à la somme de LTE BS et WiFi AP dans les deux autres scénarios) . Fait intéressant, pour le cas saturé, presque tous les scénarios ont le même effet. Ce qui est particulièrement surprenant, c'est que le cas de déchargement est presque le même que celui du LTE à un seul niveau: alors que la densité totale des BS est la même dans les deux cas, le scénario de déchargement utilise le double du spectre. scénario LTE de niveau. Afin d'esquisser l'explication, nous mentionnons: 1) l'association de déchargement n'affecte pas la distribution MCS de chaque niveau, 2) sur l'interférence toujours ON la distribution MCS de chaque niveau ne dépend pas de la densité de la BS (le gain de la la puissance reçue probabiliste supérieure est égale à la perte d'interférence supérieure). 3) Toujours en mode ON, le Wi-Fi capture un peu moins de 50% du trafic. Compte tenu de ce qui précède, le gain du deuxième niveau est seulement que la cardinalité des utilisateurs à la BS décroît, tout comme le simple niveau avec une densité BS plus élevée. L'image change totalement dans le cas de la charge, car le spectre supplémentaire signifie moins d'utilisateurs par bande et donc moins d'interférences.

Néanmoins, si nous nous intéressons aux cas d'interférence basés sur la charge, nous constatons que: (i) le planificateur WiFi affecte fortement les performances globales; (ii) le réseau à deux niveaux surpasse le LTE uniquement pour les deux ordonnanceurs, ce qui est plus en ligne avec ce que nous aurions attendu. Cela souligne en outre l'importance de l'analyse basée sur la charge, qui dans ce cas a non seulement un impact quantitatif, mais aussi un impact qualitatif clair.

Pour le reste de cette section, nous considérons uniquement un réseau WiFi best-case (ie, resource fair), afin de concentrer notre attention sur les politiques d'association, et de comprendre les limites des améliorations de performances en introduisant un niveau WiFi. Pour être plus réaliste, nous supposons maintenant un réseau secondaire plus dense que le réseau primaire. Plus

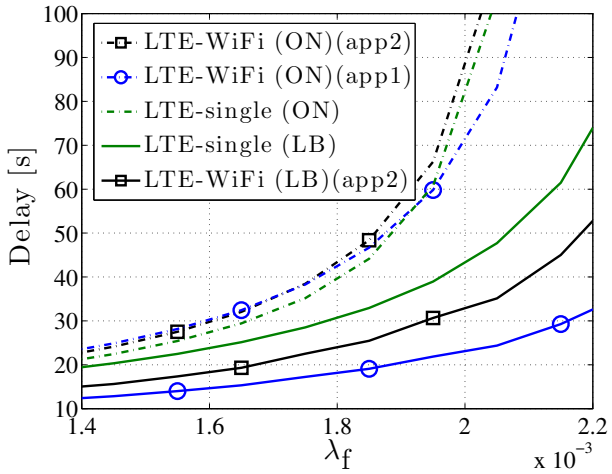


Figure 8.13 – Scenario de déchargement pour différents cas de réseau à deux niveaux, par rapport à la densité de fluxλ_f

précis, un réseau WiFi secondaire avec $\lambda_{WiFi} = 5$, et un réseau LTE primaire avec $\lambda_{LTE} = 1$. La probabilité d'encombrement et le retard par écoulement pour différents débits d'entrée de trafic λ_f sont représentés dans les Fig. 8.14 et 8.15, respectivement.

En examinant les critères de Déchargement et Max-SINR pour le cas saturé, la probabilité d'encombrement des deux cas est presque égale. Cependant, Max-SINR fonctionne mieux, en ce qui concerne le délai moyen. Cependant, compte tenu des cas de brouillage basés sur la charge, la politique de déchargement est beaucoup plus robuste en ce qui concerne la probabilité d'encombrement et surpasse également Max-SINR en ce qui concerne le retard.

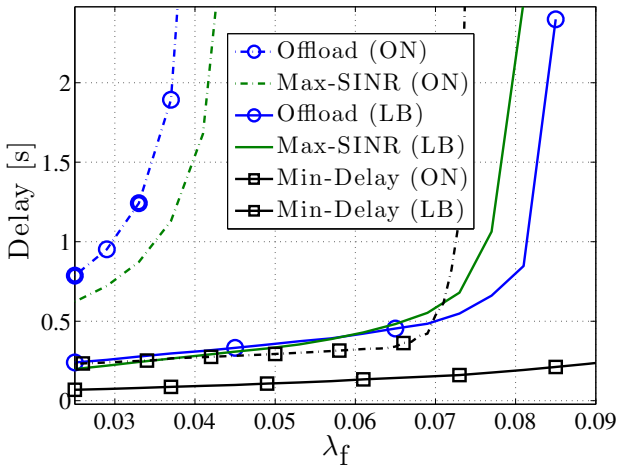


Figure 8.15 – Comparaison du délais de différents schémas d'association, par rapport à la densité de flux λ_f

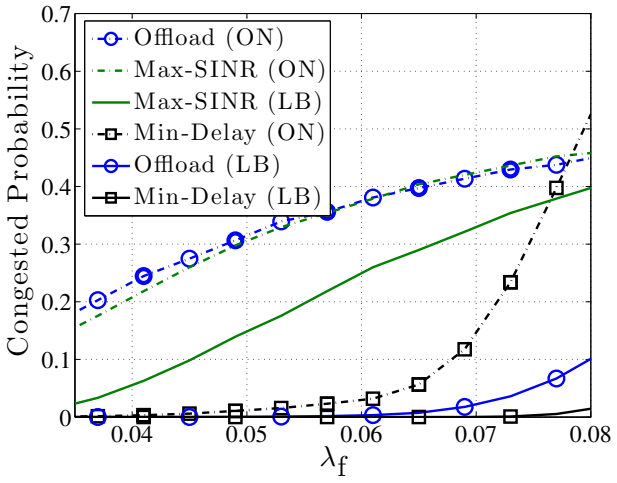


Figure 8.14 – Comparaison de probabilité congestionnée de différents schémas d'association, par rapport à la densité de flux λ_f

Cette discordance entre les cas saturés et les cas basés sur la charge provient du fait que l'analyse saturée est capable de capturer seulement un côté du gain résultant de l'augmentation de la densité du réseau. D'une part, le cas saturé capture correctement le fait qu'une BS «arbitraire» en moyenne doit servir moins d'utilisateurs, dans un réseau plus dense (traitant ainsi un plus petit ρ en raison d'une diminution du numérateur, c'est-à-dire, le trafic d'entrée). D'autre part, il ne parvient pas à capturer que les BS environnantes seront aussi moins chargées, et donc provoque moins d'interférences, ce qui à son tour, conduit à des performances encore meilleures pour les utilisateurs (moins) servis (en raison d'une augmentation de μ et une diminution supplémentaire de ρ). En conséquence, le modèle saturé sous-estime les schémas d'association qui tendent à utiliser davantage le réseau plus dense (WiFi).

Dernier point, mais non des moindres, il est clair à partir de la figure 8.15 que la politique d'association Min-Delay surpasse significativement les deux autres, jusqu'à un ordre de grandeur ou plus, pour les charges élevées. Contrairement aux deux autres politiques qui ne considèrent que la performance de la couche PHY (MAX-SINR) ou naïvement essayer de réduire la charge du réseau primaire (Offload), Min-Delay prend directement en compte la charge réelle, qui joue le rôle clé sur la performance par flux. Il est également intéressant de noter que, en particulier pour les faibles charges, la politique Min-Delay est également assez stable en termes de probabilité d'encombrement (Fig. 8.14). Bien que les politiques d'association considérées soient une version abstraite des politiques détaillées réelles, nous pensons que ces résultats plaident en faveur de mécanismes d'association sophistiqués basés sur la charge dans les futurs HetNets, afin de mieux équilibrer les charges entre les niveaux et garantir la meilleure expérience utilisateur.

Pour résumer, dans ce chapitre nous avons présenté un cadre de travail analytique pour analyser la performance au niveau des flux d'un réseau HetNet à plusieurs niveaux et comparer quelques critères communs d'association utilisés. Trois conclusions principales ressortent de ce chapitre: i) La politique d'ordonnancement pourrait fortement affecter les performances du système au niveau des flux, même si les caractéristiques PHY sont les mêmes que celles observées lors de la comparaison des deux cas du planificateur WiFi.). ii) Les deux approches d'interférence

différentes, toujours activées et basées sur la charge, modifient totalement les performances du système (à un ou plusieurs niveaux), nous devons donc faire très attention à cette hypothèse lorsque nous modélisons un système. iii) Le gain de la politique d'association liée à la charge (min-Delay) est étonnamment élevé comparé aux plus traditionnels (Off-load ou Max-SINR).

8.7 Conclusions et Perspectives

Cette thèse contribue au problème général de la modélisation et de l'analyse des performances des réseaux sans fil placés au hasard. Nous suivons une approche pas si bien étudiée mais très intéressante qui se concentre sur l'analyse de la performance au niveau du flux du réseau. Nous croyons fermement que le réseau devient de plus en plus chargé et orienté sur les paquets. Ces approches fourniront de nombreux avantages et perspectives dans la conception du réseau.

Notre objectif était de développer un cadre qui exporte des informations supplémentaires sur le réseau, telles que le retard des utilisateurs et la probabilité BS encombrée que les approches traditionnelles ne sont pas capables de capturer. Pour ce faire, nous combinons des idées issues de la théorie de la file d'attente pour modéliser la performance dynamique d'un planificateur BS et d'une géométrie stochastique afin de modéliser la topologie du réseau. Nous avons utilisé ces deux outils mathématiques avec notre abstraction de couche PHY afin de créer un cadre analytique solide pour une prédiction précise de la performance au niveau du flux d'un grand réseau placé aléatoirement.

L'analyse mentionnée ci-dessus considère à la fois le cas des BS voisines interférentes ON, ainsi que celui des interférences dépendantes de la charge. Il s'avère que l'écart de performance entre les cas susmentionnés pourrait être assez élevé, affectant non seulement les connaissances quantitatives, mais aussi souvent les conclusions qualitatives, et devrait donc être soigneusement pris en compte lors de la conception du réseau. Un autre paramètre qui affecte les performances du système est la sélection du planificateur; nous avons vu que même si les caractéristiques PHY restent les mêmes, les performances de l'utilisateur (délai) dépendent fortement du planificateur BS, même si les performances de l'opérateur (probabilité d'encombrement) ne changent pas. Voici d'autres commotions générales intéressantes: 1) La performance au niveau du débit du réseau dépend du taux de service qui est habituellement calculé en fonction de la moyenne harmonique des taux et non de la moyenne statistique; Cela signifie que les utilisateurs de bord affectent beaucoup plus sur les performances du réseau. 2) La relation entre les délais des utilisateurs et la charge du réseau n'est pas linéaire, le délai explose lorsque la charge dépasse une limite. Cette limite dépend de la politique d'ordonnancement, mais nous pouvons la définir à 0,9.

De plus, nous avons comparé notre travail avec les frameworks SoA qui utilisent des approches de valeur moyenne pour estimer la performance du réseau et nous avons vu que cette approche surpasse significativement la nôtre en particulier dans le régime de charge élevée charge).

En outre, nous avons analysé théoriquement le coût énergétique d'un tel réseau et nous avons étudié le compromis entre l'efficacité énergétique et la performance du niveau de flux du réseau. Le paramètre réseau principal de notre étude était la densité de la BS, et nous montrons qu'en général, désactiver BS ne conduit pas nécessairement à des économies d'énergie comme on le croit souvent. Nous avons présenté quelques résultats théoriques sur la réduction de la densité BS dans le cas de la réduction de la densité des utilisateurs et nous avons fourni une règle empirique, lorsque cette réduction conduit à un gain d'énergie. D'autre part, nous montrons que

si le nombre d'utilisateurs n'est pas réduit, il existe une possibilité d'amélioration de l'énergie sans affecter la QoE d'un utilisateur lorsque l'énergie opérationnelle n'est pas beaucoup plus grande que la consommation d'énergie constante.

De plus, nous avons élargi notre cadre pour le cas de réseaux hétérogènes considérés comme des topologies à plusieurs niveaux et différentes technologies d'accès radio. Nous modélisons des critères d'association communs et évaluons leur impact sur les performances centrées sur l'utilisateur et sur le réseau. Cette étude montre que si nous considérons la charge dans notre association de niveau, le gain de performance au niveau du flux pourrait être extrêmement élevé.

Cette étude, compte tenu du large éventail de paramètres et de degrés de liberté, ne pourrait jamais être complète, mais nous fournissons des aperçus représentatifs et démontrons l'utilité de notre cadre analytique proposé.

Perspectives

Il y a encore beaucoup de problèmes ouverts dans notre cadre que les travaux futurs pourraient probablement accomplir. Le plus important de ceux-ci est l'étude théorique de la convergence entre la distribution MCS et le taux de service dans le cas de brouillage basé sur la charge. Cette étude devrait se concentrer sur la détermination de la date à laquelle cette convergence est réalisable et s'il est possible de se retrouver avec une solution de forme fermée du taux de service. Nous savons que ce n'est pas une tâche facile du tout, mais nous croyons que ce résultat pourrait aider beaucoup de chercheurs à étudier de façon analytique la performance du réseau dynamique dans le scénario plus réaliste de l'interférence basée sur la charge.

Pour d'autres travaux futurs, il sera intéressant d'appliquer notre cadre, avec différents critères d'association, dans des scénarios d'agrégation de porteuses. Nous devrions examiner les scénarios où les utilisateurs planifient leur trafic sur des niveaux déférents en fonction de leur taille de flux et recherchent des améliorations possibles, par exemple, un utilisateur enverra tout son contenu de type «texte» via sa connexion cellulaire et son contenu volumineux. comme des photos grâce à sa connexion WiFi.

De plus, nous croyons que notre cadre pourrait être modifié pour analyser les scénarios de coexistence LTE et WiFi dans les mêmes bandes et étudier les mécanismes pour que cette coexistence soit la plus juste possible, comme des sous-trames presque vides ou un LTE sens porteur ou optimal. Seuils MCS et niveaux de puissance de transmission pour les deux RAT.

En outre, nous devrions étudier les performances du réseau dans le cas où HetNets fonctionnerait dans les mêmes bandes de fréquence et que les petites cellules seraient en mesure d'effectuer l'extension de la plage cellulaire afin de décharger une partie du trafic de la macrocellule. D'une part, la petite cellule déchargera certains utilisateurs de l'autre, provoquera plus d'interférences.

Enfin, une étude intéressante sera d'élargir notre cadre pour le cas des utilisateurs prioritaires. Cela peut se produire en supposant qu'une partie du réseau servira uniquement aux utilisateurs primaires ou en utilisant certains des résultats déjà développés de la théorie de la file d'attente qui fournissent des expressions analytiques pour les files d'attente avec différentes catégories d'utilisateurs, où nous pouvons utiliser résultats dans les systèmes sans fil ainsi.

Bibliography

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2014–2019," *Whitepaper*, 2015.

[2] F. C. Commission. Advanced wireless services (aws-3), auction 97. [Online]. Available: http://wireless.fcc.gov/auctions/

[3] Qualcomm-Research, "Extending lte advanced to unlicensed spectrum," Qualcomm technologies, Inc, Tech. Rep., 2013.

[4] ——, "Lte in unlicensed spectrum," Qualcomm technologies, Inc, Tech. Rep., 2014.

[5] ——, "Makin the best use of unlicensed spectrum for 1000x," Qualcomm technologies, Inc, Tech. Rep., 2015.

[6] Ericsson, "Carrier wi-fi: the next generation," Ericsson Review, Tech. Rep., 2013.

[7] Alcatel, Ericsson, Qualcomm, Samsung, and Verizon, "Lte-u technical report, coexistance study for lte-u sdl," LTE-U Forum, Tech. Rep., 2015.

[8] Huawei, "U-lte: Unlicensed spectrum utilization of lte," Huawei Technology Co., Ltd, Tech. Rep., 2014.

[9] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in hetnets: old myths and open problems," *Wireless Communications, IEEE*, 2014.

[10] Nokia, "5g use cases and requirements," *White Paper*, 2015.

[11] Ericsson, "5g radio access," *White Paper*, 2015.

[12] *Traffic Engineering: What? Why? How?* Arizona Highway Department, Traffic Engineering Division, 1968.

[13] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Communications*, 2011.

[14] M. Deruyck, W. Vereecken, E. Tanghe, W. Joseph, M. Pickavet, L. Martens, and P. Demeester, "Power consumption in wireless access network," in *Wireless Conference (EW)*, 2010.

[15] *LTE Specifications, http://www.3gpp.org/DynaReport/36-series.htm.*

[16] *802.11 Specifications, http://standards.ieee.org/about/get/802/802.11.html.*

[17] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action.* Cambridge University Press.

[18] F. Baccelli and S. Zuyev, "Stochastic geometry models of mobile communication networks," *Boca Raton, FL: CRC Press*, 1996.

[19] T. X. Brown, "Cellular performance bounds via shotgun cellular systems," *IEEE Journal on Selected Areas in Communications*, 2000.

[20] J. Andrews, F. Baccelli, and R. Ganti, "A tractable approach to coverage and rate in cellular networks," *Communications, IEEE Transactions on*, 2011.

[21] L. Decreusefond and T. T. V. P. Martins, "An analytical model for evaluating outage and handover probability of cellular wireless networks," *The 15th International Symposium on Wireless Personal Multimedia Communications*, 2012.

[22] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, "Modeling and analysis of k-tier downlink heterogeneous cellular networks," *Selected Areas in Communications, IEEE Journal on*, 2012.

[23] X. Lin, J. Andrews, and A. Ghosh, "Modeling, analysis and design for carrier aggregation in heterogeneous cellular networks," *Communications, IEEE Transactions on*, 2013.

[24] S. Singh, H. Dhillon, and J. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *Wireless Communications, IEEE Transactions on*, 2013.

[25] H. Dhillon, M. Kountouris, and J. Andrews, "Downlink coverage probability in mimo hetnets," in *ASILOMAR*, Nov 2012.

[26] T. Novlan, R. Ganti, A. Ghosh, and J. Andrews, "Analytical evaluation of fractional frequency reuse for heterogeneous cellular networks," *Communications, IEEE Transactions on*, 2012.

[27] Y. S. Soh, T. Q. S. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, 2013.

[28] G. Zhang, Q. S. Quek, A. Huang, M. Kountouris, and H. Shan, "Backhaul-aware base station association in two-tier heterogeneous cellular networks," *SPAWC*, 2015.

[29] Y. Li, F. Baccelli, J. G. Andrews, T. Novlan, and J. Zhang, "Modeling and analyzing the coexistence of wi-fi and lte in the unlicensed spectrum," *IEEE Transactions on Wireless Communications*, 2016.

[30] H. Dhillon, R. Ganti, and J. Andrews, "Load-aware modeling and analysis of heterogeneous cellular networks," *Wireless Communications, IEEE Transactions on*, 2013.

[31] I. Siomina and D. Yuan, "Analysis of cell load coupling for lte network planning and optimization," *IEEE Transactions on Wireless Communications*, 2012.

[32] H. S. Dhillon, Y. Li, P. Nuggehalli, Z. Pi, and J. G. Andrews, "Fundamentals of heterogeneous cellular networks with energy harvesting," *IEEE Transactions on Wireless Communications*, 2014.

[33] S. M. Yu and S. L. Kim, "Downlink capacity and base station density in cellular netwrorks," in *WiOpt*, 2013.

[34] T. Bonald and J. Roberts, "Scheduling network traffic," in *ACM SIGMETRICS*, 2007.

[35] S. Aalto, U. Ayesta, S. Borst, V. Misra, and R. Núñez Queija, "Beyond processor sharing," in *ACM SIGMETRICS*, 2007.

[36] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *Networking, IEEE/ACM Transactions on Networking*, 2005.

[37] T. Bonald and A. Proutiere, "Wireless downlink data channels: User performance and cell dimensioning," in *ACM MOBICOM*, 2003.

[38] T. Bonald, B. Sem, H. Nidhi, J. Matthieu, and A. Proutiere, "Flow-level performance and capacity of wireless networks with user mobility," *Queueing Systems: Theory and Applications*, 2009.

[39] N. Sapountzis, S. Sarantidis, T. Spyropoulos, N. Nikaein, and U. Salim, "Reducing the energy consumption of small cell networks subject to qoe constraints," in *GLOBECOM IEEE*, 2014.

[40] 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 Communications*, 2011.

[41] Y. Zhong, M. Haenggi, T. Q. S. Quek, and W. Zhang, "On the stability of static poisson networks under random access," *IEEE Transactions on Wireless Communications*, 2016.

[42] O. Galinina, S. Andreev, M. Gerasimenko, Y. Koucheryavy, N. Himayat, S. P. Yeh, and S. Talwar, "Capturing spatial randomness of heterogeneous cellular/wlan deployments with dynamic traffic," *IEEE Journal on Selected Areas in Communications*, 2014.

[43] M. Karray and M. Jovanovic, "A queueing theoretic approach to the dimensioning of wireless cellular networks serving variable bit-rate calls," *IEEE Transactions on Vehicular Technology*, 2013.

[44] B. Błaszczyszyn, M. Jovanovic, and M. K. Karray, "Performance laws of large heterogeneous cellular networks," in *WiOpt*, 2015.

[45] ——, "How user throughput depends on the traffic demand in large cellular networks," in *WIOPT-SPASWIN*, 2014.

[46] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α -optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking,*, 2012.

[47] "White paper: Ieee 802.11ac migration guide," FLUKE networks, Tech. Rep., 2015.

[48] *OpenAir Interface, www.openairinterface.org.*

[49] S. Sesia, I. Toufik, and M. Baker, *LTE - the UMTS long term evolution : from theory to practice.* Wiley, 2009.

[50] AirMagnet and inc, "802.11n Primer," *Whitepaper*, 2008.

[51] G. Bianchi, "Performance analysis of the ieee 802.11 distributed coordination function," *Selected Areas in Communications, IEEE Journal on*, 2000.

[52] Y. Lin and V. Wong, "Wsn01-1: Frame aggregation and optimal frame size adaptation for ieee 802.11n wlans," in *GLOBECOM IEEE*, 2006.

[53] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in lte cellular networks: Key design issues and a survey," *Communications Surveys Tutorials, IEEE*, 2013.

[54] G. Fayolle, I. Mitrani, and R. Iasnogorodski, "Sharing a processor among many job classes," *J. ACM*, 1980.

[55] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda, "Performance anomaly of 802.11b," in *INFOCOM IEEE*, 2003.

[56] A. Karr, *Probability.* Springer, 1993.

[57] K. Avtachenkov, U. Ayesta, P. Brown, and R. Núñez Queija, "Discriminatory processor sharing revisited," *INFOCOM*, 2005.

[58] F. Baccelli, B. Blaszczyszyn, and F. Tournois, "Spatial averages of downlink coverage characteristics in cdma networks," in *INFOCOM IEEE*, 2002.

[59] T. Bonald and A. Proutière, "On performance bounds for the integration of elastic and adaptive streaming flows," in *ACM SIGMETRICS*, 2004.

[60] G. Arvanitakis and F. Kaltenberger, "Phy layer modeling of lte and wifi rats," Eurecom, Tech. Rep. RR-16-317, 2016.

[61] A. F. Molisch, *Wireless Communications.* Wiley, 2010.

[62] G. Arvanitakis, "Distribution of the number of poisson points in poisson voronoi tessellation," Eurecom, Tech. Rep. RR-15-304, 2014.

[63] G. Arvanitakis and F. Kaltenberger, "Stochastic analysis of two-tier hetnets employing lte and wifi," in *EuCNC*, 2016.

[64] "Strategies for mobile network capacity expansion," *Real Wireless, White Paper*, 2010.

[65] "Looking ahead to 5g," *Nokia Solutions and Networks, White Paper*, Dec. 2013.

[66] X. Ge, S. Tu, G. Mao, C. X. Wang, and T. Han, "5g ultra-dense cellular networks," *IEEE Wireless Communications*, 2016.

[67] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Communications Magazine*, 2011.

[68] J. Wu, Y. Zhang, M. Zukerman, and E. K. N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Communications Surveys Tutorials*, 2015.

[69] E. Dahlman, S. Parkvall, and J. Skold, *4G, LTE-Advanced Pro and The Road to 5G.* Academic Press, 2016.

[70] G. Arvanitakis, T. Spyropoulos, and F. Kaltenberger, "An analytical model for flow-level performance of large, randomly placed small cell networks," *GLOBECOM IEEE*, 2016.

[71] G. Arvanitakis, F. Kaltenberger, and T. Spyropoulos, "An analytical model for flow-level performance in heterogeneous wireless networks," *IEEE Transactions on Wireless Communications.*

[72] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. Thomas, J. Andrews, P. Xia, H. Jo, H. Dhillon, and T. Novlan, "Heterogeneous cellular networks: From theory to practice," *Communications Magazine, IEEE*, 2012.

[73] P. Fotiadis, M. Polignano, L. Chavarria, I. Viering, C. Sartori, A. Lobinger, and K. Pedersen, "Multi-layer traffic steering: Rrc idle absolute priorities amp; potential enhancements," in *VTC IEEE*, 2013.

[74] N. Sapountzis, S. Sarantidis, T. Spyropoulos, N. Nikaein, and U. Salim, "Reducing the energy consumption of small cell networks subject to qoe constraints," in *GLOBECOM IEEE*, Dec 2014.

[75] M. Tanemura, "Statistical distributions of poisson voronoi cells in two and three dimensions," *Forma*, 2003.

[76] ——, "Statistical distributions of shape of poisson voronoi cells," in *Voronoi Conference on Analytic Number Theory and Spatial Tessellations*, 2005.