2015-ENST-0019

EDITE - ED 130

Doctorat ParisTech

THÈSE

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Communication et Électronique »

présentée et soutenue publiquement par

Fidan MEHMETI

le 24 avril 2015

Analyse de la performance et optimisation de l'accès sans fil dans les réseaux cognitifs et hétérogènes

Directeur de thèse : **Thrasyvoulos SPYROPOULOS**

Jury

M. David GESBERT, Professeur, EURECOM	Président du jury
M. Yuming JIANG, Professeur, NTNU	Rapporteur
M. Paolo GIACCONE, Maître de Conférences, Politecnico di Torino	Rapporteur
M. Marcelo DIAS DE AMORIM, Directeur de Recherches, LIP6 UPMC	Examinateur
M. Giovanni NEGLIA, Chargé de Recherches, INRIA	Examinateur

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

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

EURECOM
Sophia Antipolis

DISSERTATION

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

Specialization

Communication and Electronics

École doctorale Informatique, Télécommunications et Électronique (Paris)

Presented by

Fidan MEHMETI

Performance Analysis and Optimization of Wireless Access in Cognitive and Heterogeneous Networks

Defense scheduled on April 24^{th} 2015

before a committee composed of:

Prof. David Gesbert	President of the Jury
Prof. Yuming Jiang	Reporter
Prof. Paolo Giaccone	Reporter
Prof. Marcelo Dias De Amorim	Examiner
Prof. Giovanni Neglia	Examiner
Prof. Thrasyvoulos Spyropoulos	Thesis Supervisor

THÈSE

présentée pour obtenir le grade de
Docteur de
Telecom ParisTech

Spécialité

Communication et Electronique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Fidan MEHMETI

Analyse de la performance et optimisation de l'accès sans fil dans les réseaux cognitifs et hétérogènes

Soutenance de thèse prévue le 24 avril 2015

devant le jury composé de:

Prof. David Gesbert	Président du jury
Prof. Yuming Jiang	Rapporteur
Prof. Paolo Giaccone	Rapporteur
Prof. Marcelo Dias De Amorim	Examinateur
Prof. Giovanni Neglia	Examinateur
Prof. Thrasyvoulos Spyropoulos	Directeur de thèse

Abstract

In the last years there has been an increasing spectrum demand for wireless applications, as a consequence of the rapid penetration of laptops, smartphones and tablets in the technology market, as well as the applications that they provide, which are mainly very bandwidth-hungry. This has become a major concern for mobile network operators, who are forced to often operate very close to (or even beyond) their capacity limits. Due to the static spectrum allocation policies, the issue of spectrum scarcity became a major problem in today's wireless industry. On the other hand, measurements of the utilization of licensed wireless spectrum have revealed that the available spectrum is rather under-utilized, exhibiting high variability across space, frequency, and time [1]. So, in addition to rapidly increasing demand, the current lack of flexibility in dynamically assigning spectrum to match requests over time and space further limits the service levels offered.

Recently, different solutions have been proposed to alleviate these problems. A potential solution is the utilization of *dynamic spectrum access (DSA)*, with *cognitive radio (CR)* as its key enabling technology [2]. For that purpose, Cognitive Radio Networks (CRN) have been proposed to opportunistically discover and exploit (temporarily) unused licensed spectrum bands, in which the secondary users' (SU) activity is subordinated to primary users (PU). This leads to the necessity of SU to adjust its transmission parameters accordingly, so that there are no impairments on the PU QoS. In order to accomplish this goal, cognitive radios have to be equipped with some additional features, such as: spectrum sensing, spectrum decision, spectrum sharing, and spectrum mobility. Based on the subordinated nature of the cognitive radio activity, dynamic spectrum access leads to a model where SU access is *intermittent*, with channel availability patterns being random processes, affecting performance in non-trivial ways, depending on a lot of parameters (the percentage of time an SU is active, the distribution of time intervals during which a PU is active, etc.).

Another important way to deal with the data crunch, which is beneficial for both the cellular operators and mobile users, is to use Heterogeneous Networks (HetNets) comprising of small cells (femtocells, picocells) and/or WiFi networks, as well as aggressive *offloading* from the main (cellular) network. Heterogeneous Networks and offloading also lead to intermittent access (to one or more wireless access technologies), which is also subject of randomness of the availability of the specific interface, and can be modeled in a similar way as the availability in the cognitive case. In this thesis, we have modeled this type of access with alternating renewal processes.

Networks must provide a certain level of QoS to its applications. In the case of delay sensitive applications, like VoIP, streaming live audio and video, teleconferencing etc., the delay induced to a file transmitted by a secondary user is of huge importance. The same holds for a mobile user, who is waiting to find a WiFi AP, and to transmit/receive data through that interface. As expected, the performance of an intermittent user is determined mainly by the activity of the interfering licensed users in the “neighborhood” (percentage of time being able to communicate, duration of availability periods, duration of traffic bursts, etc.). To quantitatively determine to what extent goes this dependency, we need to perform some analysis. As a first step for the analysis comes the need for having realistic *models*. These models and the analysis therein can be used later for the design of efficient protocols for intermittent users, among other things.

Hence, in this thesis we have proposed models that enable us to assess the performance of cognitive radio networks and mobile data offloading. This will enable us to understand to what extent the performance of the secondary users will depend on a number of parameters in a cognitive radio network, as well as to see how can the performance in a mobile offloading system be improved.

After providing in Chapter 1 a short introduction to cognitive radio networks and mobile data offloading, as well as the motivation to our work, we propose in Chapter 2 a model for calculating the average packet delay in a cognitive radio network. We model the activity of a PU on a given channel with an alternating renewal process (ON-OFF stochastic processes), in which the SU can transmit only during the periods of time when the PU is idle (OFF periods). Based on this model and using queueing theory, we derive the expression for average packet delay under generic distribution for ON and OFF periods, and show that the variability of the PU activity on a given channel plays a much more important role than the duty cycle itself.

While in Chapter 2 the main focus is on the delay, as the key figure of merit, in Chapter 3 our main metric of interest is throughput. When a single radio is used for both transmission and sensing, an interesting tradeoff arises: when one or more channels among the currently available ones are lost (e.g. primary user returns), should the node start scanning immediately, or instead it should continue transmitting over the remaining available channels? Using renewal-reward theory, we show that if the goal is to maximize the average (long-term) throughput, the answer to this question depends on the statistics of the availability periods. Specifically, for relatively homogeneous channels, we show that it is optimal to start scanning immediately, while for heterogeneous channels, it is often better to defer scanning, even if multiple channels are lost.

Depending on the nature of the interaction between the SU and PU, there are two frequently encountered types of spectrum access: *underlay* and *interweave*. In Chapters 2 and 3 we analyze the interweave mode of access. In Chapter 4, we propose models for *underlay* access mode, as well as slightly less generic (but highly accurate models) for interweave access than in Chapter 2, that can enable proper comparison between the two modes. With these models we are not only able to calculate the average file delay in a CRN, but also the average data rate. We compare the performance of both modes and answer which one is better, in terms of delay and throughput, for a given set of parameters. On top of that, we also propose hybrid policies that are a combination of the two modes, improving the performance even more.

While in Part I we are concerned with the performance of cognitive radio networks, in Part II of this thesis we analyze analytically the two types of mobile data offloading: *on-the-spot*, and *delayed* offloading. The analysis is again based on modeling the availability of the interface of interest with alternating renewal processes. First, in Chapter 5 we propose a model for on-the-spot mobile offloading, in which we model the activity of the mobile user with a two-dimensional Markov chain (with WiFi and cellular access), whose outcome is the average file delay in an on-the-spot offloading system. Essentially, this model is the same as the underlay model of Chapter 4. The other parameter we are able to calculate is the *offloading efficiency*, which represents the percentage of data transmitted through the WiFi interface. We further extend our model to capture the case with a heterogeneous network, comprising of multiple interfaces, like: no network, WiFi, UMTS, EDGE, 4G, etc. For this more complex case, we provide a closed-form solution as well.

The other type of mobile data offloading, namely *delayed offloading*, is the topic of Chapter 6. The coverage with AP is again modeled with alternating renewal processes. However, under this setup the mobile user will not transmit/receive data when there is no WiFi up to a certain deadline. We propose a model to calculate the average file delay in such a system based on 2D Markov chains. Yet similar, there is a difference with the model of Chapter 5 that comes from the dependency of the transition rates on the state of the system. This model provides a closed form expression for the delay, as well as for the percentage of data sent over the WiFi interface. Furthermore, we propose some simple energy and cost models, which we use to solve some optimization problems, and to optimize the performance in terms of delay and cost.

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

Acknowledgements

First and foremost, I would like to thank my supervisor at Eurecom, Prof. Thrasyvoulos Spyropoulos. The work on this thesis was made possible through his continuous guidance and support. He has been a great teacher and excellent advisor patiently guiding me through research during the last three years. Working with him was a great experience and an invaluable lesson for my future professional career.

I would also like to thank the colleagues from our research group, Pavlos and Delia, with whom I had the pleasure to collaborate. They were always willing to discuss various research problems and offer me their valuable advice.

The life here in Sophia-Antipolis, far away from my family and friends was not always easy for me. However, there were some good friends here and home, that made the life easier for me and helped me to overcome some difficult situations. For that, I would specially like to thank my good friend and colleague from Eurecom, Vuk. I had so many enjoyable discussions with him during lunch time, and many trips during weekends and holidays having a great fun. Another friend I would like to give special thanks is my friend from Kosovo, Rinor. Although we could see each other only few times in the previous three years, we were in touch very often online, and he was always showing interest on how my work was going. Special thanks would go to Erdet for helping me with some drawings, and to my cousin Afrim for so many conversations we had.

I would also like to thank my precious, younger and smarter sister Vlera for her love and support. Finally, and most importantly, I dedicate this thesis to the best parents in the world, my father Shenasi and my mother Emine, for their infinite love, support, comprehension and so many other wonderful things they have taught me and have done for me so far. They were always there supporting me when I needed them, during the most difficult situations in my life. There are no words with which I can express my gratitude towards them. Without them, I could hardly arrive at this point, in completing my PhD studies. I would also like to thank them and my sister for visiting me here at *Côte d'Azur*, and for spending so many magnificent moments together. I hope that one day I will be able to make them up for what they have done for me.

Contents

Abstract . . . i
Acknowledgements . . . v
Contents . . . vii
List of Figures . . . xi
List of Tables . . . xv
Acronyms . . . xvii

1 Introduction 1
1.1 Cognitive Radio Networks . . . 1
1.2 Mobile Data Offloading . . . 4
1.3 Motivation and Contributions of the Thesis . . . 5
1.3.1 Motivation . . . 5
1.3.2 Contributions and Outline . . . 6

I Modeling, Analysis and Optimization of Cognitive Radio Networks 11

2 Who Interrupted Me? Analyzing the Effect of PU Activity on Cognitive User Performance 13
2.1 Introduction . . . 13
2.2 Performance modeling . . . 14
2.2.1 Arrival and Service Processes for SU Traffic . . . 16
2.3 Performance analysis . . . 26
2.4 Related work . . . 28
2.5 Conclusion . . . 28

3 To Scan or Not To Scan: The Effect of Channel Heterogeneity on Optimal Scanning Policies 31
3.1 Introduction . . . 31
3.2 Related Work . . . 33
3.3 Analysis . . . 34
3.3.1 Problem Setup . . . 34
3.3.2 Analytical Model . . . 35
3.3.3 Exponential IID OFF Periods . . . 37
3.3.3.1 Finding an optimal threshold . . . 39
3.3.4 Heterogeneous exponentially distributed channels . . . 40
3.3.5 The online (adaptive) algorithm . . . 42

3.4 Simulation results . . . 43
3.4.1 Homogeneous channels . . . 43
3.4.2 Initial cost . . . 43
3.4.3 Heterogeneous channels . . . 44
3.4.4 The adaptive algorithm . . . 46
3.4.5 Generic OFF periods . . . 46
3.5 Conclusion . . . 47

4 Stay or Switch? Analysis and Comparison of Interweave and Underlay Spectrum Access in Cognitive Radio Networks 49
4.1 Introduction . . . 49
4.2 Performance modeling of spectrum access . . . 52
4.2.1 Problem setup for underlay access . . . 52
4.2.2 Problem setup for interweave access . . . 52
4.3 Delay analysis of underlay and interweave access . . . 54
4.3.1 Delay analysis for interweave access . . . 55
4.3.2 Delay analysis for underlay access . . . 65
4.3.3 Analytical comparison of delays in underlay and interweave mode . . . 66
4.3.4 The delay minimization policy . . . 67
4.4 Throughput analysis of underlay and interweave access . . . 71
4.4.1 Analytical comparison of throughput . . . 71
4.4.2 Throughput maximization policy . . . 73
4.5 Simulation results . . . 75
4.5.1 Validation of the delay models . . . 75
4.5.2 Validation of the throughput models . . . 78
4.5.3 Delay minimization policies . . . 78
4.5.4 Throughput maximization policies . . . 80
4.6 Related Work . . . 82
4.7 Conclusion . . . 83

II Modeling, Analysis and Optimization of Mobile Data Offloading 85

5 Performance Analysis of On-the-Spot Mobile Data Offloading 87
5.1 Introduction . . . 87
5.2 Analyzing Offloading in Homogeneous Cellular Networks . . . 89
5.2.1 Problem Setup . . . 89
5.2.2 Delay Analysis . . . 91
5.2.3 Low utilization approximation . . . 94
5.2.4 High utilization approximation . . . 94
5.2.5 Moderate utilization approximation . . . 95
5.2.6 Generic file size distribution approximation . . . 95
5.2.7 Offloading efficiency . . . 96
5.3 Analyzing Offloading in Heterogeneous Cellular Networks . . . 97
5.3.1 The model . . . 97
5.3.2 Analysis . . . 98

5.4 Simulation results . . . 102
5.4.1 Basic model validation . . . 102
5.4.1.1 Validation of the main delay result . . . 103
5.4.1.2 Validation against non-exponential ON-OFF periods . . . 104
5.4.1.3 Validation of non-exp flow sizes . . . 104
5.4.1.4 Validation of approximations . . . 104
5.4.2 Limitations of the 2-level model . . . 105
5.4.3 M-level model validation . . . 107
5.4.4 Non-exponential assumptions . . . 108
5.4.5 Offloading Gains . . . 110
5.5 Related Work . . . 112
5.6 Conclusion . . . 113

6 Is it Worth to be Patient? Analysis and Optimization of Delayed Mobile Data Offloading 115
6.1 Introduction . . . 115
6.2 Analysis of Delayed Offloading . . . 117
6.2.1 Performance of WiFi queue . . . 118
6.2.2 Low utilization approximation . . . 124
6.2.3 High utilization approximation . . . 125
6.3 Performance evaluation . . . 126
6.3.1 Validation of main delay result . . . 126
6.3.2 Validation of approximations . . . 128
6.3.3 Variable WiFi rates and non-exponential parameters . . . 129
6.3.4 Delayed offloading gains . . . 130
6.4 Optimizing Delayed Offloading . . . 130
6.4.1 Optimization problems . . . 131
6.4.2 Optimization evaluation . . . 134
6.5 Related Work . . . 135
6.6 Conclusion . . . 136

7 Conclusions and Future Work 137
7.1 Future work . . . 138

8 Résumé 141
8.1 Motivation et contributions de la thèse . . . 142
8.2 Contribution et bref aperçu . . . 143

List of Figures

2.1 The activity pattern of the primary user. 15
2.2 Renewal cycle. 21
2.3 System time for exp-OFF. 26
2.4 System time for BP-OFF. 26
2.5 WLAN primary user. 27
2.6 Cellular primary user. 27

3.1 The channel model. 34
3.2 The transmission and scanning phases. 35
3.3 Homogeneous exp. channels. 44
3.4 Homogeneous exp. channels. 44
3.5 Initial scanning cost. 44
3.6 Initial scanning cost. 44
3.7 Heterogeneous exp. channels. 45
3.8 Heterogeneous exp. channels. 45
3.9 Heterogenerous exp. channels with low var. 45
3.10 Throughput for different policies. 45
3.11 Homogeneous uniform. 46
3.12 Heterogeneous uniform. 46
3.13 Homogeneous Pareto. 47
3.14 Heterogeneous Pareto. 47

4.1 The illustration of the underlay mode with: a) idle PU (high periods), b) active PU (low periods). 50
4.2 The illustration of the interweave mode with: a) idle PU (high periods), b) active PU (low periods). 50
4.3 The distribution of the scanning time for low PU duty cycles. 55
4.4 The distribution of the scanning time for two group of channels. 55
4.5 The 2D Markov chain for exponentially distributed scanning time. 56
4.6 The 2D Markov chain for Erlang-distributed scanning time. 59
4.7 The 2D Markov chain for hyperexponentially distributed scanning time. 62
4.8 The 2D Markov chain for the underlay model. 65
4.9 The renewals for the underlay CRN case. 72
4.10 The renewals for the interweave CRN case. 72
4.11 The throughput policy. 73
4.12 The delay for underlay spectrum access. 76

4.13 The delay for interweave spectrum access in the cellular scenario. 76
4.14 The delay for interweave spectrum access in the WiFi scenario. 76
4.15 The delay for generic underlay spectrum access. 76
4.16 The delay for generic interweave spectrum access. 77
4.17 The throughput for underlay spectrum access. 77
4.18 The throughput for interweave spectrum access. 78
4.19 The static delay policy for different λ and exp. scanning times. 78
4.20 The static delay policy for different λ and hyperexp. scan. times. 79
4.21 The generic static delay policy for different λ. 79
4.22 The dynamic delay policy for Pareto OFF periods. 80
4.23 The throughput static policy for two different low period data rates. 80
4.24 Max. Scanning times for different ratios of data rates. 81
4.25 The throughput dynamic policy for Pareto OFF periods. 81

5.1 The WiFi network availability model. 90
5.2 The 2D Markov chain for the simple on-the-spot mobile data offloading scenario of Def. 3 and 4. 92
5.3 The reduced Markov chain for $\rho \to 0$. 95
5.4 The multi-network availability model. 98
5.5 The 2D Markov chain for multilevel on-the-spot mobile data offloading model. . 99
5.6 Pedestrian user. 103
5.7 Vehicular user. 103
5.8 BP vehicular periods. 103
5.9 Generic flow sizes. 103
5.10 The low utilization approx. 105
5.11 The high utilization approx. 105
5.12 The approx. for AR=0.5. 105
5.13 Variable WiFi rates. 105
5.14 Variable rates for $AR = 0.5$. 107
5.15 Variable rate with $AR = 0.75$. 107
5.16 Variable rate case with lower cellular rate. 107
5.17 The approximation with the 2 level model. 107
5.18 The transmission delay. 109
5.19 The transmission delay. 109
5.20 Bounded Pareto periods. 109
5.21 Generic flow sizes. 109
5.22 Different traffic rates. 111
5.23 Offloading efficiencies. 111
5.24 The design problem with increasing LTE coverage. 111
5.25 The design problem with increasing WiFi coverage. 111

6.1 The WiFi network availability model. 118
6.2 The 2D Markov chain for the WiFi queue in delayed offloading. 119
6.3 The reduced Markov chain for $\rho \to 0$. 124
6.4 The average delay for pedestrian users' scenarios. 127
6.5 The average delay for vehicular users' scenarios. 127

6.6 The delay for BP ON-OFF periods vs. theory. 128
6.7 Low utilization delay approximation for $AR = 0.75$. 128
6.8 Low utilization p_r approx. for $AR = 0.75$. 128
6.9 High utilization delay approximation for $AR = 0.5$. 128
6.10 Variable WiFi rates with the same average as theory. 129
6.11 The delay for deterministic deadlines vs. theory. 129
6.12 Deterministic packets. 129
6.13 The delay for BP packet sizes vs. theory. 129
6.14 Offloading gains for delayed vs. on-the-spot offloading. 131
6.15 The delay function for the optimization problem. 131
6.16 The delay vs. cost curve for high cellular rate. 134
6.17 The delay vs. cost curve for low cellular rate. 134

8.1 Le modèle d'activité d'un utilisateur principal (PU). 144
8.2 Le délai de système pour exp-OFF. 146
8.3 Le délai de système pour BP-OFF. 147
8.4 Canaux homogènes exponentiels. 149
8.5 Canaux homogènes exponentiels. 150
8.6 Le débit pour les différentes politiques (règles). 152
8.7 La chaîne de Markov 2D pour le temps de scannage avec distribution exponentielle. 152
8.8 La chaîne de Markov 2D pour le temps de scannage Erlang distribué. 153
8.9 La chaîne de Markov 2D pour le temps de scannage hyper-exponentielle. 155
8.10 La chaîne de Markov 2D pour le modèl underlay. 156
8.11 Les renouvellements pour le cas underlay. 157
8.12 Les renouvellements pour le cas interweave. 157
8.13 Le délai pour l'accès underlay au spectre. 158
8.14 Le débit pour l'accès interweave au spectre. 159
8.15 La politique statique de délai pour λ différentes et temps exponentiel de scannage. 160
8.16 La politique dynamique de délai pour des périodes OFF de Pareto. 161
8.17 La politique dynamique de débit pour des périodes OFF de Pareto. 161
8.18 Le modèle de la disponibilité du réseau WiFi. 162
8.19 La chaîne Markov 2D pour le modèle de déchargement on-the-spot des données mobiles avec plusieurs niveaux. 163
8.20 L'utilisateur piétonne. 164
8.21 L'utilisateur en véhicule. 165
8.22 Le délai de transmission. 166

List of Tables

2.1 Variables and Shorthand Notation. 17
2.2 The ratio of delays for two different channels. 28

4.1 Variables and Shorthand Notation. 55
4.2 The analytical comparison of underlay and interweave modes. 67
4.3 Summary of the delay policies. 70

5.1 Variables and Shorthand Notation. 91
5.2 The parameters for different access technologies [3, 4]. 108

6.1 Variables and Shorthand Notation. 118
6.2 Probability of reneging for pedestrian and vehicular scenarios. 127
6.3 Optimal deterministic deadline times vs theory. 135

8.1 Les variables et la notation abrégée. 145
8.2 Le rapport des délais pour deux canaux différents. 148
8.3 Les variables et la notation abrégée. 154
8.4 La comparaison analytique des modes underlay et interweave. 155
8.5 Résumé des politiques dynamiques. 156
8.6 Les paramètres pour différentes technologies d'accès [3, 4]. 164

Acronyms

2D MC Two Dimensional Markov Chain

3G Third Generation

3GPP Third Generation Partnership Project

4G Fourth Generation

AP Access Point

ANDSF Access Network Discovery and Selection Function

BP Bounded Pareto

BTS Base Transceiver Station

CTMC Continuous Time Markov Chain

CR Cognitive Radio

CRN Cognitive Radio Network

CDF Cumulative Distribution Function

CCDF Complementary Cumulative Distribution Function

DFR Decreasing Failure Rate

DSA Dynamic Spectrum Access

EDGE Enhanced Data Rates for GSM Evolution

FCFS First Come First Served

GPRS General Packet Radio Service

GSM Global System for Mobile Communications

HSPA High Speed Packet Access

HSDPA High-Speed Downlink Packet Access

HetNets Heterogeneous Networks

IID Independent and Identically Distributed

IFR Increasing Failure Rate

IP Internet Protocol

IFOM IP Flow Mobility

LTE Long Term Evolution

LCFS Last Come First Served

KKT Karush-Kuhn-Tucker

MC Markov Chain

MAC Medium Access Control

MSC Mobile Switching Center

PU Primary User

pdf Probability Density Function

PS Processor Sharing

QoS Quality of Service

SU Secondary User

SIPTO Selected IP Traffic Offload

SNR Signal to Noise Ratio

TCP Transmission Control Protocol

UMTS Universal Mobile Telecommunications System

VoIP Voice over IP

WiFi Wireless Fidelity

WLAN Wireless Local Area Networks

Chapter 1

Introduction

1.1 Cognitive Radio Networks

In the last years there has been an increased spectrum demand for wireless applications. Due to the static spectrum allocation policies, the issue of spectrum scarcity became a major problem in today's wireless industry. On the other hand, a large portion of the assigned spectrum is underutilized over time, space and frequency, and measurements reveal that the level of under utilization can go up to 85% even in the dense populated metropolitan areas [1].

Dynamic spectrum access techniques have recently been proposed to alleviate these problems. *Cognitive radio* [2] is the key enabling technology for dynamic spectrum access. Due to the significant advantages they offer in terms of efficient spectrum utilization, *cognitive networks* have been on the focus of research in wireless communications in the last years. In a cognitive network, there exist licensed users which are assigned the spectrum from the regulation authority with certain charges, such as cellular operators, or without charges (users in the ISM band). These are called *primary users* (PU). There exist also users that utilize the spectrum opportunistically whenever they find it available. These users are known as *cognitive* or *secondary users* (SU), and they are subordinated to primary users' activities. Hence, they have to adapt their transmission parameters, such as: power, type of modulation, coding, etc., in order not to cause impairments to PUs.

In order to exploit dynamically the spectrum, cognitive users must be equipped with additional functionality. All the capabilities of cognitive networks can be realized through spectrum management function that includes: *spectrum sensing*, *spectrum decision*, *spectrum sharing*, and *spectrum mobility.*

One of the most important features the cognitive users need to possess is their ability to sense a wide range of licensed or unlicensed bands and detect the presence (absence) of primary users on them. This is known as *spectrum sensing.* Based on the outcomes of that process, the SU decides if it can communicate on a given channel, and when it should re-adapt or stop completely the transmission or reception. It is very important, since any mistakes during the sensing procedure can either degrade the performance of the licensed user, or can prevent the secondary user from utilizing a given portion of the spectrum although there are no PUs active on it. The former mistake is known as *missdetection*, and it occurs during the sensing phase when a SU fails in detecting a PU signal. This can happen, e.g. when the SU is relatively far from the PU, and the PU signal arrives with a very low power at SU, or e.g. it can be due to the presence of fading in the PU signal at the moment when the SU is sensing. The later mistake can be a consequence of the very low threshold posed by the sensing unit to detect the presence of PU signals, making them very sensitive, in which case even the background noise with slightly higher power can be detected as a PU signal. This is known as the *false alarm*.

There are several techniques of implementing spectrum sensing. The most well known, and thoroughly used are: *energy detection*, *matched-filter detection*, and *feature detection*. In energy detection spectrum sensing, the sensor has a threshold in which all the received signals with higher energy than the threshold value are perceived as PU signals. This form of sensing is the simplest one. If the sensing is based on a filter that provides the highest SNR ratio (matched-filter), then it is called *matched-filter* detection. Cognitive users with sensing devices which exploit one or more of the properties of the received signal waveform are said to implement *featured-detection* spectrum sensing.

Depending on the network architecture, the decision if a channel is being used by a PU or is idle can be made in different ways. Namely, if we are considering a centralized network architecture, in which there is a central entity that decides for each mobile SU node when and how to access a channel, the decision from the master unit is made only after receiving the sensing outcomes for corresponding channels from individual SUs in the coverage region. There are several ways how the sensing results can be combined from the central unit to decide if a channel is busy or idle [1]. If out of all sensing results for a given channel, it suffices only one SU node to have sent the information that the given channel is busy for the central unit to decide that the considered channel is busy, then that approach is called "OR" rule (related to the logical operator OR).

The other type of decision making for a given channel in a centralized network is the "AND" rule. According to this rule (following the logical operator AND), all the SU nodes need to confirm that the given channel is busy, for it to be considered as such by the central unit, and to command the nodes to behave correspondingly towards it. So, in this case, it suffices one node to claim for a channel to be idle, and it can be declared as such. Obviously, both aforementioned approaches can produce a lot of erroneous decisions. A safer way would be if the central entity decides upon the majority of "votes" from individual nodes. It means that the channel is declared as busy, if most of the nodes' sensors have detected a PU signal on that channel. If one wants to further decrease the probability of missdetections, the central unit can use the rule "m-out-of-n". Under this option, if there are n SUs that sense a given channel, then that channel to be considered as busy, there must be at least m nodes ($m < n$) sensing it busy.

Nevertheless, the situation is slightly different in a decentralized wireless network (ad-hoc network). Namely, in such a network each SU decides individually if a given channel is busy or idle. However, if there is no exchange of information between the nodes of their sensing results, and each node decides based on its own measurements, then we deal with *non-cooperative* spectrum sensing. On the other hand, if nodes exchange their information between them, and if each node decides on the availability of a given channel based also on the data from its neighbors, then that form of sensing is known as *cooperative* spectrum sensing.

The other important feature of a CR is the capability to decide which is the most appropriate channel among the available ones. Of course, this will depend on the type of the applications the SU is interested in. This is known as *spectrum decision*. Usually, the spectrum decision process consists of two phases [5, 6]: *the channel characterization phase* and *the decision phase*.

The first phase is very important since the different spectrum holes[1] lie in the different parts of the spectrum, have different bandwidths, etc. Hence, it very important to characterize each available channel by: availability, interference, path loss, link errors, link delays, etc. After all the spectrum holes have been characterized, the SU will choose the appropriate channel that best suits her.

After the PU reclaims back its original channel, the SU currently residing on the same channel needs to either completely stop the transmission process, or to modify its transmission parameters so that no harmful interference is caused to the licensed user. In case the SU has to cease immediately the transmission, it needs to either evacuate itself from that channel, and start looking for another available one, or it should remain silent on the same channel and wait until it becomes idle again. This capability of cognitive users is known as *spectrum mobility.* The question if the SU should stay and wait on the same channel or should start searching for a new channel is not that easy to answer. It depends on many factors, such as the nature of primary users' activities on other channels, the time needed to switch the radios from one to another frequency, sensing time, etc. The total time needed to find a new available channel comprises of two parts. The first one is the *scanning time*, which represents the total time spent on sensing all the channels until an available one is found. The second component is the *switching delay.* This is the time spent while switching from a busy sensed channel to another channel waiting to be sensed.

Since the wireless channel in principle is of shared nature, then it requires coordination between the cognitive users to access it. This kind of coordination is prone to be solved by a traditional MAC protocol. However, the coexistence paradigm in a cognitive radio network between primary and secondary users urges for an extra feature on the side of CRs. It is the *spectrum sharing* [1] capability that CRs also need to possess.

Spectrum sharing techniques in a CRN can be classified based on different aspects, such as: architecture (*centralized* and *distributed*), spectrum allocation (*cooperative* and *non-cooperative*), spectrum access techniques, etc. [5]. The most interesting classification is based on the spectrum access techniques. There are three well known approaches that can be categorized under the spectrum access techniques.

Spectrum access techniques can be classified as: *underlay*, *interweave*, and *overlay.* In the *underlay mode*, the SU reduces its transmission power when a PU is utilizing a given channel such that the maximum interference level a PU can tolerate is met. When the PU ceases transmission, the SU can increase its power to the maximum level. On the other hand, in the *interweave* mode, the cognitive user cannot transmit at all when there is an active PU on that channel. However, when an SU finds an idle channel it can transmit with the highest power under the constraint that it does not cause interference to neighboring channels. Several factors (scanning time, nature of PU activity, the traffic intensity, etc.) impact the overall performance, and the fact which one of these two modes provides better results for the SU.

As opposed to the two previous techniques, in the *overlay* mode the cognitive user serves as a relay to a PU. For this, the SU as a reward gets the right to access a portion of the PU spectrum. Due to its increased complexity, the overlay access technique is less interesting.

Since the secondary user (SU) activity is subordinated to primary users, there will be interruptions during cognitive user transmissions. When on a given channel there are no active PU,

[1] By spectrum hole we usually mean the part of the spectrum consisting of a single or multiple contiguous channels that can be utilized under certain constraints from cognitive users.

an SU can communicate normally as a licensed user, if it has data to receive/send. The interruptions take place when a PU reclaims back its channel, for which it has the exclusive rights of utilization. At the same moment, the SU will have to change its transmission parameters (power, modulation, coding, direction of antenna, etc.), such that there is no deterioration of PU transmissions.

Depending on the type of spectrum access technique used by the SU, the nature of interruptions can be twofold. In interweave spectrum access, the SU has to completely cease its transmission, when a PU has something to send/receive. During the whole busy period of the PU, the SU is not allowed to transmit any data. During these time intervals, a SU can either (i) stay on the same channel, and wait for the PU to finish its busy period and become idle again; or (ii) initiate the scanning process and look for another available channel by sensing the presence of PUs on the different spectrum bands, with no possibility of data transmission. On the other hand, in the case of underlay spectrum access, when the PU reclaims back its channel, the SU already transmitting there normally will not completely cease its transmission. Instead, the SU will have to re-adapt its transmission power (or even coding, modulation, or other parameters), such that the PU will maintain its QoS. Usually, the quantitative measure that constrains the transmission power of a SU is the maximum allowed interference at the PU. Otherwise, during PU idle periods, the SU can transmit like a “normal” user.

The above mentioned interruptions make media access by SUs *intermittent*, especially in densely utilized parts of the spectrum. Intermittent transmission might not only occur when using a licensed channel as a secondary user, but also when using HetNets (cellular, small cells and WiFi). In fact, HetNets lead to intermittent access to one or more wireless access technologies (small cells or WiFi). For example, in a scenario with complete cellular network coverage and only sporadic availability with WiFi at hotspots in a certain geographic region, we say that the mobile user has only *intermittent* access to the WiFi interface. This kind of intermittent access to the WiFi can be very attractive for both the mobile user and the operator. The problems related to this paradigm are known as *mobile data offloading*, and the goal is to relieve the main (cellular) network from overloading.

1.2 Mobile Data Offloading

An enormous growth in the mobile data traffic has been reported in the last years. This increase in the traffic demand is due to a significant penetration of laptops, smartphones, and tablets in the technology market, as well as Web 2.0 and various streaming applications which have high requirements in terms of bandwidth. It is expected that the mobile traffic will increase even more rapidly in the following period [7]. According to the same prediction, mobile video traffic in the next two years will comprise 66% of the total mobile traffic, as opposed to 51% in 2012.

This increase in traffic demand is overloading the cellular networks (especially in metro areas) forcing them to operate close or even beyond their capacity limits, causing thus a significant degradation to 3G services. As possible solutions to alleviate this capacity crunch problem are upgrading to higher capacity access technologies, such as LTE or LTE-Advanced, as well as the deployment of additional network infrastructure [8]. However, reports already suggest that such solutions are bound to face the same problems [9]. Furthermore, these solutions may not be cost-effective from the operators’ perspective: they imply an increased cost (for power, location rents, deployment and maintenance), without a similar increase in revenues [10] due to flat rate plans and the fact that few users consume a lot of traffic (3% of users consume 40% of the traffic [10]).

Instead, a better option is to offload some of the traffic through Femtocells (SIPTO, LIPA [11]), or to use WiFi instead. According to measurements, in 2012 33% of the mobile data traffic were offloaded [7]. Projections say that these figures will increase to 46% by 2017 [7].

While there are still ongoing discussions on which of the two approaches (Femtocell or WiFi) is better, as there are arguments for the one or the other side, we give priority to offload by using WiFi access points as the most efficient way from both operator's and customer's point of view. However, we also study the case of heterogeneous networks, when there are multiple wireless access technologies possible, or even when there is no WiFi. WiFi offloading has become a very popular solution due to some of the often cited advantages compared to Femtocells, such as: lower cost, higher data rates, lower ownership costs [8], etc. Also, wireless operators have already deployed or bought a large number of WiFi access points (AP) [8], as is the case with AT&T. As a result, WiFi offloading has attracted a lot of attention recently.

There exist two types of offloading: *on-the-spot* and *delayed* offloading [12]. In *on-the-spot offloading*, whenever there is WiFi available, all traffic is sent over the WiFi network; otherwise, all traffic is sent over the cellular interface. Currently, the smartphones can only switch between interfaces. Using both interfaces in parallel, as well as per flow offloading (IFOP) are currently also being considered in 3GPP [11].

More recently, *delayed* offloading has been proposed. The idea behind it is to introduce a certain threshold (deadline) up to which some traffic can be delayed if there is no WiFi availability. If up to that moment no access point is detected, the data are transmitted through the cellular network. Otherwise, the traffic is sent via the WiFi interface. Obviously, postponing the transmission of some data up to the time when there will be a WiFi AP in the vicinity of the user can provide further benefits to both the operator and the user. The main advantage for the cellular operator is the decongestion of its base stations. Since the deployment and maintenance of the WiFi APs is much cheaper than of the BTS, the operator can offer lower prices for the traffic exchanged through the WiFi interface. Also, there are some energy benefits for the mobile user when transmitting over the WiFi interface.

1.3 Motivation and Contributions of the Thesis

1.3.1 Motivation

Wireless networks need to provide a certain level of QoS to its users and applications. Quantitatively, this QoS is expressed through various parameters, such as the mean packet delay, average throughput, the jitter, bandwidth, etc. Among the most important figures of interest are the *delay* and *throughput*. Depending on the nature of application provided, the one or the other can be more important. For example, in the case of applications, like VoIP, streaming live audio and video, teleconferencing, online games etc., the delay induced to a transmitted or received file is very important. On the other hand, a number of applications of interest, such as: file download, peer-to-peer file exchange, cloud-based services, software-based services, etc., are often throughput-sensitive and for them a very important parameter is the average throughput.

As we have seen in the previous paragraph, it is of huge importance to know the key figures of merit an application is provided from a network. This is even more emphasized in intermittent networks where the mobile users do not have permanently the possibility to communicate, and can either transmit/receive data only occasionally, or cannot make use of the full potential they have during some time periods. Hence, the performance evaluation of intermittent wireless networks is one of the most important tasks in designing mobile networks.

In the following, we present the main reasons why it is important to perform analysis for both cognitive radio networks and mobile data offloading.

Cognitive Radio Networks. As expected, the performance of a cognitive user is determined mainly by the activity of interfering licensed users in the "neighborhood" (percentage of time being idle, duration of idle periods, duration of traffic bursts, etc.). In a CRN, when it comes to delay-sensitive applications, it is very important to quantitatively determine the average delay a CR packet, file, or flow experiences in such a network, and to see based on the outcome whether the required QoS for a CR application is met. The same holds for throughput-sensitive applications, for which it is very important to determine the average data rate.

Another important question that emerges is to what extent the SU should reduce its transmission power in the underlay spectrum access. For a normal operation, there is a maximum interference level a PU can tolerate [13]. The SU should know this value, and the distance from the PU. The maximum allowed transmission power is then calculated as the sum of the interference budget, the free space loss from SU to PU, and other types of losses that are specific to wireless communications, such as fading and shadowing [14].

Besides aforementioned questions, another important issue might come up. Namely, if a CR can use both underlay and interweave spectrum access modes, which one should it use so that a certain parameter (delay or throughput) is optimized? Or, if it is possible to switch from one mode to another, how and under which conditions this hybrid implementation can improve the overall performance?

Mobile Data Offloading. For a user with the possibility of switching between interfaces using either on-the-spot or delayed mobile data offloading it is of crucial importance to assess correctly its QoS parameters. A user would like to know how much she would have to wait more while using the delayed offloading. What are the gains, in terms of delay reduction if she uses occasionally a WiFi interface with a higher data rate? How much less she will be paying if she would be using an offloading system? What are the savings in terms of the energy consumption of the battery? How tolerant to delays should the user be, in order to really benefit from offloading?

Similar questions of interest may come up from the cellular operator side as well. Nevertheless, the most important parameter for the operator is the *offloading efficiency*, which is the percentage of data a user transmits through the WiFi interface, i.e. the percentage of data the operator gets rid off from its BTS.

The first step in assessing the performance of wireless networks is to have the appropriate *models*, which are realistic and can capture to a great extent the behavior of the different factors and their nature of interaction in a given system. With the appropriate models and analysis, the obtained results can be further used for the design of efficient spectrum scheduling, scanning, and handoff strategies, among other things.

In the following section, we summarize the contributions and present the outline of the thesis.

1.3.2 Contributions and Outline

The focus of this thesis is on assessing analytically the performance of mobile users with only intermittent access to the resources. This is very important, since by doing that we are able to infer if the mobile users can achieve the QoS imposed by the different types of applications they are supporting. We consider two types of this communication paradigm: *cognitive radio networks* and *mobile data offloading.* In cognitive radio networks, the performance of secondary users depends heavily on the type of primary user activity (i.e. how variable is the duration of communication when it has data to send/receive), the PU duty cycle (i.e. the percentage of time she is occupying the channel), as well as on the SU traffic characteristics (intensity, packet size). Similarly, for a mobile user willing to exploit the data offloading feature through WiFi AP, the performance depends on the coverage with APs, the number of users, the traffic intensity, and the file sizes. Hence, our main goal in this thesis is to propose models that enable us to understand how different factors and to what extent affect the performance of mobile users of interest. This further allows us to find the best technique a mobile user can use in order to achieve the optimal performance, by solving different optimization problems.

An outline of the dissertation is provided below and the contributions achieved in each chapter are summarized.

Chapter 2 - Who Interrupted Me? Analyzing the Effect of PU Activity on Cognitive User Performance.

As a first step to better understand the performance of cognitive users in a CRN is to have the adequate models. Our focus in this chapter is to analyze the average delay incurred to a SU packet. We model the PU activity with an alternating renewal process, and use queueing theory to derive the expected packet delay. An advantage of our model is that it holds for generic packet sizes. Our model finally leads to a closed-form result, and shows a complex interplay between the network and traffic parameters.

In this chapter, we also show that the variability of the primary user activity can be much more important than the duty cycle itself, and that depending on the coefficient of variation of the PU busy periods, it is implied that even channels with a very low duty cycle and high variability can yield to higher delays than channels with very high duty cycle, but with very low variability. This is a very interesting outcome of our model, and can be used when designing scheduling algorithms. As an implication of our model, a simple scheduling scheme can be proposed. Namely, for each flow we calculate the average delay it would experience on all the channels based on our formula, and then we can pick the channel that provides the smallest delay.

The works related to this chapter are:

- *F. Mehmeti, T. Spyropoulos, "Who Interrupted Me? Analyzing the Effect of PU Activity on Cognitive User Performance", in Proc. of IEEE International Conference on Communications (IEEE ICC 2013), Budapest, Hungary, 2013.*
- *F. Mehmeti, T. Spyropoulos, "Analysis of Cognitive User Performance Under Generic Primary User Activity", Tech. Report, RR-12-274, Eurecom, 2012.*

Chapter 3 - To Scan or Not To Scan: The Effect of Channel Heterogeneity on Optimal Scanning Policies.

While in Chapter 2 we are interested on the SU performance in terms of delay, in Chapter 3 our main interest lies on the average throughput analysis and its optimization. We consider the scanning problem in scenarios where multiple channels are pooled together. More importantly, instead of optimizing the scanning sequence when scanning is triggered, we investigate the complementary optimization problem of *when* to trigger this scanning function in order to optimize the (long-term) throughput that can be maintained by the SU. We introduce the notion of *threshold value* as the number of channels we are allowed to lose, before the initiation of the scanning procedure. This threshold value provides the best results in terms of the average data rate. It is proven that no such threshold value exists for the case of homogeneous (i.i.d.) channels if there is no initial cost to be paid at the beginning of the scanning process. However, this value exists for heterogeneous independent channels and its value depends on the variability of the OFF periods of the channels in use.

We also propose an adaptive algorithm that determines the instants when to stop transmission depending on which channel is lost. Namely, if a channel which is better (in terms of the larger average duration of the OFF periods) than the average of the channel pool is lost, we show that it is better to stop the transmission process and to initiate the scanning procedure, and vice versa. We show by extensive simulations that this algorithm provides the highest throughput.

The work related to this chapter is:

- *F. Mehmeti, T. Spyropoulos, "To Scan or Not To Scan: The Effect of Channel Heterogeneity on Optimal Scanning Policies", in Proc. of IEEE International Conference on Sensing, Communications, and Networking (IEEE SECON 2013), New Orleans, USA, 2013.*

Chapter 4 - Stay or Switch? Analysis and Comparison of Interweave and Underlay Spectrum Access in Cognitive Radio Networks.

In Chapter 2, we propose a generic model to find the average delay (in closed-form) in interweave CRN. Next, in Chapter 4, using queueing theory, based on two dimensional Markov chains, we derive closed-form expressions for the expected delay for underlay and interweave spectrum access as a function of key network parameters (average PU idle time, transmission rates, scanning time statistics), and user traffic statistics (traffic intensity, file size). For the underlay model of Chapter 4, we need to make some simplifying assumptions (the random variables of interest are subject to exponential distributions) in order to get closed-form formulas for the average delay. However, using simulations we show that even if we depart from the model assumptions of Chapter 4, our model can predict the delay with a significant accuracy. In order to be able to directly compare the performance of both spectrum access modes, the models for both of them need to rely on the same assumptions. Hence, in Chapter 4 we also propose a model (based on 2D Markov chains) that leads to a closed-form formula for the average delay. For this case as well, we show that even if we consider generic distributions for some of the variables, we can still get very good fit with the results relying on the model of Chapter 2. We can use these results to directly compare the performance of the two modes, and derive the conditions that would make the one or the other preferable. Finally, we also use these insights to propose a "hybrid" policy, that can switch between the two modes dynamically, in order to further improve the delay performance.

We perform the same steps for the case of average (long term) data rate for both modes using the renewal-reward theory, providing closed-form expressions for individual performances, comparing them analytically, and finally optimizing. Using a wide range of realistic simulation scenarios, we validate our analytical predictions extensively, and explore the conditions under which underlay or interweave policy performs better.

Simulation-wise we show that when the periods of time with primary user activity are subject to decreasing-failure rate distributions, the improvement in the performance of a CRN both in terms of the delay and throughput can go to 50% compared to the best static policy (the better spectrum access technique under the given conditions among the underlay and interweave spectrum access). This is already an important improvement in the key figures of interest.

The works related to this chapter are:

- *F. Mehmeti, T. Spyropoulos, "Stay or Switch? Analysis and Comparison of Interweave and Underlay Spectrum Access in Cognitive Radio Networks", submitted to IEEE Transactions on Mobile Computing, November 2014.*
- *F. Mehmeti, T. Spyropoulos, "Underlay vs. Interweave: Which one is better?", Tech. Report, RR-14-296, Eurecom, 2014.*

Chapter 5 - Performance Analysis of On-the-Spot Mobile Data Offloading.

In Chapter 5, we propose a queueing analytic model for performance analysis of on-the-spot mobile data offloading. The model in this chapter is identical to the underlay model of Chapter 4. We consider a simple scenario where the user can choose between WiFi and a single cellular technology, and derive general formulas, as well as simpler approximations for some interesting utilization regimes (low, medium and high utilization), for the expected delay and offloading efficiency as a function of WiFi availability, traffic intensity, and other key parameters. We further generalize our analysis to the case when multiple cellular technologies (and respective rates), such as: 3G, 4G, etc., are available to a user, e.g. depending on her location, and/or different rates are offered by the same technology (e.g. rate adaptation, indoor/outdoor, etc.). There we use a more complex model, based also on the Markov chain theory, which we solve by using probability generating functions.

We validate extensively our model in scenarios where most parameters of interest are taken from real measured data, and which might diverge from our assumptions, and show that even in such scenarios our model is able to predict the performance of the system with a significant accuracy. This is particularly important given that most measurements of the availability periods show a heavy-tailed behavior. We use our model to provide some preliminary answers to the questions of offloading efficiency and delay improvements through WiFi-based offloading.

The works related to this chapter are:

- *F. Mehmeti, T. Spyropoulos, "Performance Analysis of On-the-Spot Mobile Data Offloading", in Proc. of IEEE Global Telecommunications Conferenece (IEEE GLOBECOM), Atlanta, USA, 2013.*
- *F. Mehmeti, T. Spyropoulos, "Performance Analysis of Mobile Data Offloading in Heterogeneous Networks", submitted to IEEE Transactions on Mobile Computing, September, 2014.*

Chapter 6 - Is it Worth to be Patient? Analysis and Optimization of Delayed Mobile Data Offloading.

The main contributions of Chapter 6 can be summarized as follows. We propose a queueing analytic model for the problem of delayed offloading, based on two-dimensional Markov chains, and derive expressions for the average delay, and other performance metrics as function of deadlines and other key system parameters. We also give closed-form approximations for different regimes of interest. We provide some insights how the cellular queue process can be modeled, and what kind of approximations can be used there to provide a more reliable system model in terms of better describing what actually happens in a real system. The Markov chain model that we use in Chapter 6 is slightly different from that of Chapter 5. While in the Markov chain of Chapter 5 all the transition rates are independent of the state in which the system is, in one part of the Markov chain of our model in Chapter 6 the transition rate depends on the state of the system.

We validate our results extensively, using also scenarios and parameters observed in real measurement traces that depart from the assumptions made in our model. We formulate and solve basic cost-performance optimization problems, and derive the achievable tradeoff regions as a function of network parameters (WiFi availability, user load, etc.) in hand. Finally, we show that our results hold for other service type disciplines, such as Processor-Sharing (PS), Last Come First Served (LCFS), etc.

The works related to this chapter are:

- *F. Mehmeti, T. Spyropoulos, "Is it Worth to be Patient? Analysis and Optimization of Delayed Mobile Data Offloading", in Proc. of IEEE International Conference on Computer Communications (IEEE Infocom 2014), Toronto, Canada, 2014.*
- *F. Mehmeti, T. Spyropoulos, "Performance Modeling, Analysis and Optimization of Delayed Mobile Data Offloading under Different Service Disciplines", to be submitted to IEEE/ACM Transactions on Networking, March, 2015.*
- *F. Mehmeti, T. Spyropoulos, "Optimization of Delayed Mobile Data Offloading", Tech. Report, RR-13-286, Eurecom, 2013.*

Part I

Modeling, Analysis and Optimization of Cognitive Radio Networks

Chapter 2

Who Interrupted Me? Analyzing the Effect of PU Activity on Cognitive User Performance

Cognitive Networks have been proposed to opportunistically discover and exploit (temporarily) unused licensed spectrum bands. With the exception of TV white spaces, secondary users (SUs) can access the medium only *intermittently*, due to deferring to primary user (PU) transmissions and scanning for new channels. This raises the following questions: (i) what sort of delays can an SU expect on a channel given the PU utilization of this channel? (ii) how do specific characteristics of the PU activity patterns (e.g. burstiness) further affect performance? These questions are of key importance for the design of efficient algorithms for scheduling, spectrum handoff, etc. In this chapter, we propose a queueing analytical model to answer them. We model the PU activity pattern as an ON-OFF alternating renewal process with generic ON and OFF durations, and derive a closed form expression for packet delays by solving a variant of the M/G/1 queue. Contrary to the common belief that low utilization channels are good channels, we show that the expected SU delay on a channel, and thus the best channel to use, is a subtle interplay between the ON and OFF duration distributions of the primary users, and the SU traffic load. We validate our analysis against simulations for different PU activity profiles.

2.1 Introduction

Measurements of the utilization of licensed wireless spectrum have (somewhat counter-intuitively) revealed that the available spectrum is rather under-utilized, exhibiting high variability across space, frequency, and time [1]. Yet, the current lack of flexibility in dynamically assigning spectrum to match demand over time and space further limits the service levels offered, in addition to rapidly increasing demand [15].

Cognitive radios and networks have been proposed to address this problem. Cognitive users (also referred to as “Secondary Users (SU)”) can sense a range of licensed or unlicensed bands and, if found idle, opportunistically use one or more of them to meet the user/application demands. However, most of these channels are only *temporarily* available, when the licensed user (also referred to as “Primary User (PU)”) is not transmitting or receiving on them (one exception are TV white spaces [16], which can be known in advance and available for very long periods of time).

Such interruptions make media access by SUs *intermittent*, especially in densely utilized parts of the spectrum. When the SU cannot transmit anymore on the current channel(s), it will (in the simplest case) either have to wait for the channel to become available again or switch its radio to scanning mode (assuming a single radio) to discover other available channels. This can delay ongoing or new SU transmissions. The probability and duration of such delays depends on the PU's activity characteristics (percentage of time being idle, duration of idle periods, duration of traffic bursts, etc.). These characteristics can be highly variant, since the (PU) channels might belong to different primary wireless systems, be used to carry different types of traffic (e.g. voice, file transfer, web browsing, video streaming, etc.), and be governed by different protocols managing the access to this channel. All of these features impact the PU activity observed on a chosen channel by a cognitive user in non-trivial ways.

These observations lead to some important questions: (i) What is the level of performance (e.g. delay) that a secondary user should expect on a channel, given the PU activity characteristics on this channel? (ii) Does the *average* "amount" of PU activity (e.g. being active for 40% of the time) suffice to predict SU performance, or can differences between PU activity profiles (e.g. distribution of active periods) further affect SU performance and to what extent? These questions have implications for the design of efficient spectrum scheduling, scanning, and handoff strategies, among other things.

To this end, in this chapter we propose a queueing analytic model for the performance of secondary user transmissions for *general* PU activity patterns. Contrary to the common belief that a channel with lower average PU activity is a better channel, our results reveal that the SU delay is in fact a subtle interplay between PU activity statistics (1^{st}, 2^{nd} and 3^{rd} moments of the random durations of PU *active* and *idle* periods), and the secondary user traffic load. As one example, for applications with low traffic intensity (e.g. machine-to-machine communications [17]), choosing the channel with the lowest PU utilization might lead to higher delays (in fact, arbitrarily worse *in theory*).

Our model and analysis are presented in detail in Section 2.2. Then, in Section 2.3 we validate our theory using simulations of different PU activity profiles including some realistic models, recently proposed [18, 19]. Related work is given in Section 2.4. We conclude our work in Section 2.5.

2.2 Performance modeling

Primary User Model: Consider a single channel used by one or more primary users. We assume that the state of this channel can be either active ("ON"), i.e. a primary packet is transmitted (it is indifferent to the SU whether this is data, signaling, uplink or downlink traffic) or idle ("OFF"), as depicted in Fig. 2.1. The exact duration of ON and OFF periods depends first on user behavior (e.g. arrival process of PU traffic flows) [18]. Furthermore, it depends on the type of traffic (e.g. short VoIP packets vs. long file transfer packets), system details (e.g. number of independent users multiplexed on the channel) and intricate protocol interactions (e.g. MAC layer carrier sense, TCP mechanisms etc.). Such details cannot be known or inferred by the secondary user.

In order to allow for the maximum amount of generality, while maintaining analytical tractability, we thus model the ON-OFF activity pattern of PUs as an *alternating renewal*

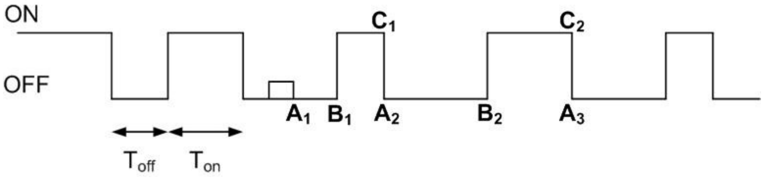

Figure 2.1: The activity pattern of the primary user.

process [20]: $(T_{ON}^{(n)}, T_{OFF}^{(n)}), n \geq 1$, where n denotes the number of $ON - OFF$ cycles elapsed until time t. The duration of any ON period, $T_{ON}^{(n)}$ (OFF period $T_{OFF}^{(n)}$, respectively), is a random variable distributed according to some probability distribution F_{ON} (F_{OFF}), and independently of other ON or OFF periods[1]. This model for PU activity is significantly more flexible in capturing different types of PUs than simple memoryless or "half-memoryless" models used in related work [21, 22]. As we will see in Section 2.3, we can accommodate recently proposed models of PU activity that attempt to mimic real traffic data sets.

Secondary User Model: In this chapter, we will assume that SUs can only access a channel at times where there is no PU activity (i.e. during OFF periods). While our model could be extended in other directions, we defer this to future work. [2]

The success of an SU packet transmission depends on the duration of that packet and the (remaining) duration of the OFF period. We assume that if a PU starts transmitting before the SU packet is completely sent, the SU transmission is considered lost (collision), and has to be retransmitted in the next idle (OFF) period. We also assume that the spectrum sensing ability of cognitive users is perfect. This means that there are no missdetections or false alarms. Such events however are orthogonal to our model and analysis, and could be captured in the PU ON-OFF process, if needed.

In practice, a cognitive (SU) user can have access to multiple channels. A number of architectures and protocols have been proposed [1] to discover and access such channels. To avoid including fine architectural details that could make our analysis intractable and reduce the generality of our results, we choose to maintain the abstraction of a single stochastic ON-OFF process. When the state of the process is ON, it means that the SU cannot transmit any new or queued traffic. This ON period might correspond, among other things, to: (i) the SU waiting on a busy channel until it becomes available again, in situations where spectrum mobility is expensive or unlikely to yield better results; (ii) a scanning period during which the SU tries to gain or regain one (or more) idle channel(s). Similarly, an OFF period (during which the SU can transmit) may correspond to using different spectrum holes "back-to-back", e.g. when backup channels are maintained [23].

[1] As usual, we denote probability distribution functions with capital letters, e.g. F(t), and the respective density function with lowercase, e.g. f(t).

[2] Our model does not apply directly to spectrum sharing based on "interference temperature", where SUs are allowed a budget of interference while concurrently transmitting with PUs. Yet, the FCC has recently commented that the interference temperature approach might not be a workable concept, increasing interference in the frequency bands where it would be used [1].

2.2.1 Arrival and Service Processes for SU Traffic

Our goal is to use a queueing model to derive the expected delay SU packets will experience, depending on the type of PU channel activity they "compete" with. We will assume that packets are generated at the cognitive user according to a Poisson process with rate λ. This will allow us to focus on the effect of PU activity on the service process, and get exact expressions for queueing delays. We defer the treatment of generic arrivals to future work. Without loss of generality, we also assume that packet sizes are fixed and equal to Δ (size normalized for transmission rate). A generic packet size distribution can be easily integrated in our derivations.

The service time that an SU packet experiences depends on the packet size, but also on the state of the PU (ON or OFF) when the packet arrives. While this points to an M/G/1 system [24], this is not the appropriate model. In fact, there are two service time distributions, *S'* and *S"*, depending on the (SU) queue length at the time of arrival:

Service S': Consider SU packets arriving to find the queue empty (no SU packets in queue or being transmitted). Then,

(i) if the PU is ON, the packet will have to wait until the next OFF cycle;

(ii) if the PU is OFF, it can immediately start transmission.

The success of this (first) transmission attempt depends on the duration (case (i)) or the remaining duration (case (ii)) of this OFF cycle. In case of failure (PU restarts too soon), the SU packet will have to wait and retry in the next OFF period(s).

Service S": Consider an SU packet that finds the SU system busy. This packet will have to queue until all packets in front of it are successfully transmitted. However, it is guaranteed that its "service" will start in an OFF period, since the packet in front of it *just* finished transmitting (successfully). [3]

The next two Lemmas derive the expected service times $E[S']$ and $E[S'']$, as a function of the PU activity profile (ON-OFF statistics). Theorem 1 then combines the two service time distributions to derive the total system delay for arriving SU packets, including both queueing and service time. We assume throughout unlimited queue size and FCFS service order.

Before we proceed, we summarize in Table 2.1 a number of variables and useful (shorthand) notations that we will use in our results and derivations.

Lemma 1. *The mean service time of non-queued packets $E[S']$ is given by*

$$\begin{aligned} E[S'] \quad &= \Delta + \frac{P}{1-p_1p_2}\left(\Delta\cdot p_2 + \left(e^{\lambda\Delta}-1\right)\left(\Delta + E[T_1] - \frac{1}{\lambda}\right)\right) \\ + \quad & \frac{P}{1-p_1p_2}\left(E[T_{ON}] + \left(E[T_2]-\frac{1}{\lambda}\right)(1-p_2)\right). \end{aligned} \tag{2.1}$$

Proof. The delay of an S' packet (finding no other SU packets queueing or in transmission) depends on its arrival time, relative to the PU state during and after that time. The key to deriving this delay is to notice the following: If we considered a single, isolated packet, we could use the *inspection paradox* to derive the expected delay [20]; e.g. renewal theory tells us that the

[3] We remind the reader that the term "system time" is typically used for the total delay (queueing + service) of a packet, while "service" begins when the packet arrives at the front of the SU queue and lasts until the packet is successfully transmitted.

Table 2.1: Variables and Shorthand Notation.

Variable	Definition/Description
T_{ON}	Duration of ON periods
T_{OFF}	Duration of OFF periods
$T_{OFF}^{(f)}$	$T_{OFF} \| T_{OFF} < \Delta$: OFF period with duration smaller than Δ
$T_{OFF}^{(e)}$	Excess OFF period
$f_{OFF}^{(e)}(x)$	$\frac{1-F_{OFF}(x)}{E[T_{OFF}]}$
p_1	$\int_0^\infty e^{-\lambda t_{off}} f_{OFF}(t_{off}) dt_{off}$
p_2	$\int_0^\infty e^{-\lambda t_{on}} f_{ON}(t_{on}) dt_{on}$
P	$\int_0^\infty e^{-\lambda t_{off,e}} f_{OFF}^{(e)}(t_{off,e}) dt_{off,e}$
p	Prob.of transmission success in an OFF period: $p = P\left[T_{OFF} > \Delta\right]$
N	Number of (extra) ON-OFF cycles until successful transmission
T_1	Duration of (extra) ON-OFF cycles (for packet arriving in ON period)
T_2	Duration of (extra) ON-OFF cycles (for packet arriving in OFF period)

stationary probability of arriving during an ON period is $\frac{E[T_{ON}]}{E[T_{ON}]+E[T_{OFF}]}$. However, this is only the limiting case, when the SU traffic arrival rate λ goes to 0 (i.e. SU traffic is very sporadic). In fact, the time until the arrival of the next S' packet starts counting *from the point the last queued packet got transmitted, which can only occur during an OFF period.* The situation is depicted in Fig. 2.1. The higher λ the higher is the probability of the packet arriving in the same OFF period.

To account for this effect, we assume a given realization ("sample path") of the ON-OFF process, represented by a vector t_s of ON and OFF durations: $t_s = \{t_1^{OFF,e}, t_1^{ON}, t_2^{OFF}, t_2^{ON}, \dots\}$, where index 1 corresponds to the (OFF) cycle when the last packet of the previous busy period got successfully transmitted. Note that for the first OFF cycle we consider the remaining ("excess") time $t_1^{OFF,e}$ right *after* the end of the last packet transmission (that ended a "busy" cycle). The delay S' on this sample path can then be expressed as follows (we will later take the expectation over all sample paths):

$$S' = I_{OFF}^1 S_{OFF}^{'} + I_{ON}^1 S_{ON}^{'} + I_{OFF}^2 S_{OFF}^{'} + I_{ON}^2 S_{ON}^{'} + \dots, \tag{2.2}$$

where I_{OFF}^i and I_{ON}^i are indicator random variables, which have value 1, only if the S' packet arrival happens in that OFF (ON) period. Clearly, only one such term can be non-zero for a given sample path. We separate this sum into two terms

$$Y_{OFF} = \sum_{i=1}^{\infty} I_{OFF}^{(i)} S_{OFF}' \text{ and } Y_{ON} = \sum_{i=1}^{\infty} I_{ON}^{(i)} S_{ON}'. \tag{2.3}$$

For this sample path, delay S' will depend on the time until the next SU packet arrival, which is exponential with rate λ. The expectation of terms in Y_{OFF} is then

$$E[I_{OFF}^{(i)} S_{OFF}^{'}] = \int_{A_i}^{B_i} \Delta \lambda e^{-\lambda \cdot x} dx + \int_{B_i - \Delta}^{B_i} \left(B_i - x + T_1\right) \lambda e^{-\lambda \cdot x} dx, \tag{2.4}$$

where $A_1 = 0, A_2 = t_1^{OFF,e} + t_1^{ON}, A_3 = t_1^{OFF,e} + t_1^{ON} + t_2^{OFF} + t_2^{ON}, \dots$ and $B_1 = t_1^{OFF,e}, B_2 = t_1^{OFF,e} + t_1^{ON} + t_2^{OFF}, \dots$, as depicted in Fig. 2.1.

The first integral is the case when the OFF period the packet arrives in is long enough for the packet to be transmitted immediately (i.e. delay $S' = \Delta$). The second integral is the case when the (remaining) OFF period is smaller than the packet size: then, transmission fails, a delay equal to that remaining time is "paid" (note that this delay is between 0 and Δ, otherwise transmission would be successful), and additional ON-OFF periods must be experienced before successful transmission. The number of such periods is a random variable, denoted by N. The total duration of these periods is thus $T_1 = \sum_{i=1}^{N} T_{ON} + \sum_{i=1}^{N-1} T_{OFF}^{(f)}$. N is a stopping time, so the expectation $E[T_1]$ can be found using Wald's equation [20]:

$$E[T_1] = \frac{1}{p}E[T_{ON}] + \left(\frac{1}{p} - 1\right) E[T_{OFF} \mid T_{OFF} < \Delta], \tag{2.5}$$

For the conditional expectation that appears in Eq.(2.5) we have $E\left[T_{OFF} \mid T_{OFF} < \Delta\right] = \int_0^{\Delta} \frac{x \cdot f_{OFF}(x)}{F_{OFF}(\Delta)} \mathrm{d}x$. Since the duration of OFF cycles is independent and distributed as $F_{OFF}(x)$, N is geometrically distributed with probability $p = P[T_{OFF} > \Delta] = 1 - F_{OFF}(\Delta)$. Calculating the integrals in Eq.(2.4) yields

$$E[I_{OFF}^{(i)} S_{OFF}^{'}] = \Delta e^{-\lambda A_i} + \left(e^{\lambda\Delta} - 1\right) e^{-\lambda B_i} \left(\Delta + E[T_1] - \frac{1}{\lambda}\right). \tag{2.6}$$

Summing over all the OFF terms of Y_{OFF} in Eq.(2.3),

$$E\left[Y_{OFF} \mid t_s\right] = \Delta \sum_i e^{-\lambda A_i} + \left(e^{\lambda\Delta} - 1\right) \left(\Delta + E[T_1] - \frac{1}{\lambda}\right) \sum_i e^{-\lambda B_i}.$$

This is the expectation of Y_{OFF}, conditional on the ON-OFF sample path t_s. Finally, we take the expectation over all possible sample paths

$$E\left[Y_{OFF}\right] = \int_0^{\infty} E\left[Y_{OFF} \mid t_s\right] f_{t_s}\left(x_1^{OFF,e}, x_1^{ON}, \ldots\right) dx_1^{OFF,e} dx_1^{ON} \ldots \tag{2.7}$$

Since ON and OFF periods are IID, we can split this integral into a product of expectations, and after some calculus we get

$$E\left[Y_{OFF}\right] = \Delta \left(1 + P p_2 \sum_i (p_1 p_2)^i\right) + \left(e^{\lambda\Delta} - 1\right) \left(\Delta + E[T_1] - \frac{1}{\lambda}\right) P \sum_i (p_1 p_2)^i,$$

where $P = \int_0^{\infty} e^{-\lambda t_{off,e}} f_{T_{off,e}}(t_{off,e}) dt_{off,e}$. Calculating the geometric sums we have

$$E\left[Y_{OFF}\right] = \Delta + \frac{P}{1 - p_1 p_2} \left(\Delta p_2 + \left(e^{\lambda\Delta} - 1\right) \left(\Delta + E[T_1] - \frac{1}{\lambda}\right)\right). \tag{2.8}$$

Using similar steps, we can calculate the term Y_{ON} of Eq.(2.3), related to packets S' arriving during an ON period

$$E[I_{ON}^{(i)} S_{ON}^{'}] = \int_{B_i}^{C_i} \left(C_i - x + T_2\right) \lambda e^{-\lambda \cdot x} dx, \tag{2.9}$$

where $C_i \in \left\{t_1^{ON} + t_1^{OFF,e}, t_1^{ON} + t_1^{OFF,e} + t_2^{OFF} + t_2^{ON}, \ldots\right\}$. C_i are also shown in Fig. 2.1. T_2 is the additional delay caused by unsuccessful packet transmissions, after the first excess ON period $T_2 = \sum_{i=1}^{N-1}\left(T_{ON} + T_{OFF}^{(f)}\right) + \Delta$. Hence,

$$E[T_2] = \left(\frac{1}{p} - 1\right)\left(E[T_{ON}] + E[T_{OFF} \mid T_{OFF} < \Delta]\right) + \Delta. \tag{2.10}$$

After solving the integral in (2.9), we get

$$E[I_{ON}^{(i)} S_{ON}^{'}] = (C_i - B_i)\, e^{-\lambda B_i} + \left(E[T_2] - \frac{1}{\lambda}\right)\left(e^{-\lambda B_i} - e^{-\lambda C_i}\right). \tag{2.11}$$

Summing over all terms in Eq.(2.3)

$$E\left[Y_{ON} \mid t_s\right] = \sum_i (C_i - B_i)\, e^{-\lambda B_i} + \left(E[T_2] - \frac{1}{\lambda}\right)\sum_i \left(e^{-\lambda B_i} - e^{-\lambda C_i}\right).$$

This is the expectation of Y_{ON}, conditioned on the ON-OFF sample path t_s. Finally, we take the expectation over all possible sample paths

$$E\left[Y_{ON}\right] = \int_0^{\infty} E\left[Y_{ON} \mid t_s\right] f_{t_s}\left(x_1^{OFF,e}, x_1^{ON}, \ldots\right) dx_1^{OFF,e} dx_1^{ON} \ldots \tag{2.12}$$

As before, this integral can be split into a product of expectations, which after some calculus yields

$$E[Y_{ON}] = P \cdot E[T_{ON}] \sum_i (p_1 p_2)^i + \left(E[T_2] - \frac{1}{\lambda}\right)\left(P \sum_i (p_1 p_2)^i - P p_2 \sum_i (p_1 p_2)^i\right) \tag{2.13}$$

Calculating the geometric sums in Eq.(2.13) we have

$$E\left[Y_{ON}\right] = \frac{P}{1 - p_1 p_2}\left(E[T_{ON}] + \left(E[T_2] - \frac{1}{\lambda}\right)(1 - p_2)\right). \tag{2.14}$$

Finally, by summing Eq.(2.14) with Eq.(2.8) we have Eq.(2.1).

□

As mentioned previously, if the arriving packet finds other packets in the system, it has to be queued and wait until its turn. A type 2 packet can start its service only during an OFF period. The following Lemma gives the average service time of a type 2 packet.

Lemma 2. *The mean service time of queued packets $E[S'']$ is given by*

$$E[S^{''}] = \Delta + \int_0^{\Delta} x f_{OFF}^{(e)}(x) dx + \frac{1}{p}\left(E[T_{ON}] + E\left[T_{OFF} \mid T_{OFF} < \Delta\right]\right)\int_0^{\Delta} f_{OFF}^{(e)}(x) dx. \tag{2.15}$$

Proof. The service time of a packet that finds the system busy is

$$S^{''} = \Delta + I\left(T_{OFF}^{(e)(f)} + \sum_{i=1}^{N} T_{ON} + \sum_{i=1}^{N-1} T_{OFF}^{(f)}\right). \tag{2.16}$$

$T_{OFF}^{(e)(f)}$ is the random variable that denotes the first OFF period in which a type 2 packet enters service (excess OFF period) and is smaller than the packet size. I is an indicator random variable that describes the inability of the packet to be transmitted in the first attempt. It is defined as

$$I = \begin{cases} 1 & \text{if } T_{OFF}^{(e)} < \Delta \\ 0 & \text{otherwise} \end{cases}$$

Its expectation is $E[I] = P[T_{OFF}^{(e)} < \Delta]$. Taking the expectations in Eq.(2.16)

$$E[S^{''}] = \Delta + E\left[IT_{OFF}^{(e)(f)}\right] + E\left[I\right]E\left[\sum_{i=1}^{N} T_{ON}\right] + E\left[I\right]E\left[\sum_{i=1}^{N-1} T_{OFF}^{(f)}\right], \tag{2.17}$$

and rearranging we obtain

$$E[S^{''}] = \Delta + E\left[IT_{OFF}^{(e)(f)}\right] + P\left(T_{OFF}^{(e)} < \Delta\right)\frac{1}{p}E[T_{ON}] + P\left(T_{OFF}^{(e)} < \Delta\right)\left(\frac{1}{p} - 1\right)E\left[T_{OFF} \mid T_{OFF} < \Delta\right]. \tag{2.18}$$

In the previous equation, I and $T_{OFF}^{(e)(f)}$ are not independent. Hence,

$$E\left[IT_{OFF}^{(e)(f)}\right] = \int_0^{\Delta} x f_{OFF}^{(e)}(x)dx. \tag{2.19}$$

For the probability of failure in the first attempt we have

$$P\left(T_{OFF}^{(e)} < \Delta\right) = \int_0^{\Delta} f_{OFF}^{(e)}(x)dx. \tag{2.20}$$

Replacing Eq.(2.20) and Eq.(2.19) into Eq.(2.18), we obtain Eq.(2.15).

□

The system described here has two different service times, depending on whether the arriving customer finds or not other customers in the system. As a result, we cannot simply use the Pollaczek-Khinchin (P-K) formula to derive the queueing delay for our system [24]. Nevertheless, we can still follow the "tagged-user" approach to find this delay.

Theorem 1. *Let an SU access a channel with generic PU activity, such that it experiences two service time distributions S' and S'' with known first and second moments. Then, the total system delay for SU packets is equal to*

$$E[T] = \frac{E[S^{'}]}{1 + \lambda\left(E[S'] - E[S'']\right)} + \frac{\lambda E[S''^2]}{2(1 - \lambda E[S''])} + \frac{\lambda\left(E[S'^2] - E[S''^2]\right)}{2 + 2\lambda\left(E[S'] - E[S'']\right)}. \tag{2.21}$$

Proof. The system delay of an SU packet consists of its queueing delay $E[T_Q]$ and its service time $E[S]$. We first consider the service time. An arriving packet will find the system busy with probability ρ, which is the utilization of the system, and idle with probability $1 - \rho$. So, the service time can be given as $E\left[S\right] = \left(1 - \rho\right)E[S^{'}] + \rho E[S^{''}]$. Applying Little's Law on the service part of the system, we get that $\rho = \lambda E\left[S\right]$. Substituting this above equation and solving for $E[S]$, we have

$$E\left[S\right] = \frac{E[S^{'}]}{1 + \lambda\left(E\left[S'\right] - E\left[S''\right]\right)}. \tag{2.22}$$

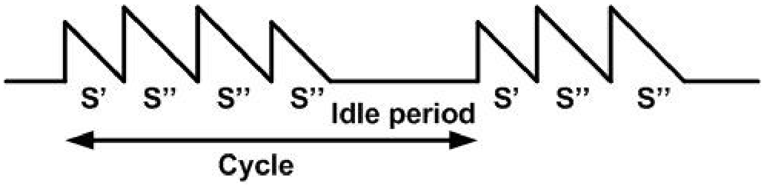

Figure 2.2: Renewal cycle.

We now consider the queueing delay (incurred with probability ρ). A packet arriving in the queue finds a packet in service, and it will have to wait for the remaining (i.e. excess) service time S_e of that packet. Assume further that it finds additional N_Q packets in front of it in the queue. Then, the expected queueing time for that packet is

$$E\left[T_Q\right] = E\left[N_Q\right] E\left[S''\right] + E\left[S_e\right].$$

Using Little's law $E\left[N_Q\right] = \lambda E\left[T_Q\right]$ and rearranging we get

$$E\left[T_Q\right] = \frac{E\left[S_e\right]}{1 - \lambda E\left[S''\right]}. \tag{2.23}$$

The mean excess time $E\left[S_e\right]$ differs from that of an $M/G/1$. We will find it using renewal-reward theory [20]. A renewal starts whenever a packet arrives and finds the system empty. The busy period (B) ends when all the packets that were generated in the meantime get transmitted and the system becomes idle again. The period until a next busy cycle begins is the idle period (I). So, in our case a cycle consists of a busy and an idle period. From renewal-reward theory we know that the mean excess time is equal to the ratio between the mean reward during a cycle and the mean duration of the cycle $E\left[S_e\right] = \frac{E[R]}{E[X]}$. The reward is defined as the remaining service during an arrival, similarly to the M/G/1 case (this is illustrated in Fig. 2.2).

From Fig. 2.2 we can infer that the excess time is

$$E\left[S_e\right] = \frac{E[\frac{1}{2}S'^2] + ME[\frac{1}{2}S''^2]}{E[B] + E[I]}. \tag{2.24}$$

In Eq.(2.24), M is the average number of arrivals finding the server busy during a renewal cycle, and is

$$M = \lambda E\left[B\right]. \tag{2.25}$$

In a long run, the utilization of the system is $\rho = \frac{E[B]}{E[B]+E[I]}$. The average idle period is $E[I] = \frac{1}{\lambda}$. Then, for the average busy period we have

$$E\left[B\right] = \frac{E\left[S'\right]}{1 - \lambda E\left[S''\right]}. \tag{2.26}$$

Replacing Eq.(2.26) into Eq.(2.25) and Eq.(2.24), as well as Eq.(2.25) into Eq.(2.24), we have

$$E[S_e] = \frac{\lambda E[S''^2]}{2} + \frac{\lambda\left(E[S'^2] - E[S''^2]\right)}{2 + 2\lambda\left(E[S'] - E[S'']\right)}\left(1 - \lambda E[S'']\right). \tag{2.27}$$

Now, replacing Eq.(2.27) into Eq.(2.23) gives us

$$E[T_Q] = \frac{\lambda E[S''^2]}{2(1-\lambda E[S''])} + \frac{\lambda\left(E[S'^2] - E[S''^2]\right)}{2 + 2\lambda\left(E[S'] - E[S'']\right)}. \tag{2.28}$$

Finally, by replacing Eq.(2.22) and Eq.(2.28) into $E[T] = E[S] + E[T_Q]$, we obtain Eq.(2.21). □

We also need the 2^{nd} moments for S' and S'' in Eq.(2.21). We will derive them in the following Lemmas.

Lemma 3. *The second moment of non-queued packets $E[S'^2]$ is given by*

$$\begin{aligned} E[S'^2] \quad &= \Delta^2 + \frac{P}{1-p_1p_2}\left(\Delta^2 p_2 + 2\Delta e^{\lambda\Delta}\left(E[T_1] - \frac{1}{\lambda}\right) + \left(e^{\lambda\Delta} - 1\right)\left(\left(E[T_1] - \frac{1}{\lambda}\right)^2 + \frac{1}{\lambda^2} + \Delta^2\right)\right) \\ + \quad &\frac{P}{1-p_1p_2}\left(E[T_{ON}^2] + 2E[T_{ON}]\left(E[T_2] - \frac{1}{\lambda}\right) + \left(\left(E[T_2] - \frac{1}{\lambda}\right)^2 + \frac{1}{\lambda^2}\right)(1-p_2)\right) \end{aligned} \tag{2.29}$$

Proof. The proof for the second moment of the non-queued packets follows similar steps as the proof in Lemma 2.1. First, we express the second moment as

$$S'^2 = I^1_{OFF} S'^2_{OFF} + I^1_{ON} S'^2_{ON} + I^2_{OFF} S'^2_{OFF} + I^2_{ON} S'^2_{ON} + \ldots, \tag{2.30}$$

where the indicator random variables I^i_{ON} and I^i_{OFF} are identical as in the proof of Lemma 2.1. We separate the sum in Eq.(2.30) into two terms

$$Y^2_{OFF} = \sum_{i=1}^{\infty} I^{(i)}_{OFF} S'^2_{OFF} \text{ and } Y^2_{ON} = \sum_{i=1}^{\infty} I^{(i)}_{ON} S'^2_{ON}. \tag{2.31}$$

The expectation of terms in Y^2_{OFF} is

$$E[I^{(i)}_{OFF} S'_{OFF}] = \int_{A_i}^{B_i} \Delta^2 \lambda e^{-\lambda\cdot x} dx + \int_{B_i-\Delta}^{B_i} (B_i - x + T_1)^2 \lambda e^{-\lambda\cdot x} dx, \tag{2.32}$$

where $A_1 = 0, A_2 = t_1^{OFF,e} + t_1^{ON}, A_3 = t_1^{OFF,e} + t_1^{ON} + t_2^{OFF} + t_2^{ON}, \ldots$ and $B_1 = t_1^{OFF,e}, B_2 = t_1^{OFF,e} + t_1^{ON} + t_2^{OFF}, \ldots$, as depicted in Fig. 2.1. Calculating the integrals in Eq.(2.32) we get

$$E[I^{(i)}_{OFF} S'_{OFF}] = \Delta^2 e^{-\lambda A_i} + 2\Delta e^{-\lambda(B_i-\Delta)}\left(E[T_1] - \frac{1}{\lambda}\right) + e^{-\lambda B_i}\left(e^{\lambda\Delta} - 1\right)\left(\left(E[T_1] - \frac{1}{\lambda}\right)^2 + \frac{1}{\lambda^2} + \Delta^2\right) \tag{2.33}$$

Summing over all terms of Y^2_{OFF} in Eq.(2.31) we have

$$\begin{aligned} E[Y^2_{OFF} \mid t_s] \quad &= \Delta^2 \sum_i e^{-\lambda A_i} + 2\Delta \sum_i e^{-\lambda(B_i-\Delta)}\left(E[T_1] - \frac{1}{\lambda}\right) \\ + \quad &\sum_i e^{-\lambda B_i}\left(e^{\lambda\Delta} - 1\right)\left(\left(E[T_1] - \frac{1}{\lambda}\right)^2 + \frac{1}{\lambda^2} + \Delta^2\right). \end{aligned} \tag{2.34}$$

Eq.(2.34) is the expectation of Y_{OFF}^2 conditioned on the given realization t_s of the sample space. We find Y_{OFF}^2 by taking the expectation of the all possible sample paths

$$E\left[Y_{OFF}^2\right] = \int_0^\infty E\left[Y_{OFF}^2 \mid t_s\right] f_{t_s}\left(x_1^{OFF,e}, x_1^{ON}, \ldots\right) dx_1^{OFF,e} dx_1^{ON} \ldots \tag{2.35}$$

As before, this integral is then split into a product of expectations and after some calculus we get

$$E[Y_{OFF}^2] = \Delta^2 + \frac{P}{1-p_1p_2}\left(\Delta^2 p_2 + 2\Delta e^{\lambda\Delta}\left(E[T_1] - \frac{1}{\lambda}\right) + \left(e^{\lambda\Delta} - 1\right)\left(\left(E[T_1] - \frac{1}{\lambda}\right)^2 + \frac{1}{\lambda^2} + \Delta^2\right)\right) \tag{2.36}$$

The term Y_{ON}^2 that is related to the packets arriving during an ON period, is derived as follows. The expectation of terms of Y_{ON}^2 in Eq.(2.31) are

$$E[I_{ON}^{(i)} S_{ON}^{\prime 2}] = \int_{B_i}^{C_i} \left(C_i - x + T_2\right)^2 \lambda e^{-\lambda \cdot x} dx, \tag{2.37}$$

After solving the integral in Eq.(2.37) we obtain

$$\begin{aligned} E[I_{ON}^{(i)} S_{ON}^{\prime 2}] &= \left(C_i - B_i\right)^2 e^{-\lambda B_i} + 2\left(E[T_2] - \frac{1}{\lambda}\right)\left(C_i - B_i\right) e^{-\lambda B_i} \\ + & \left(\left(E[T_2] - \frac{1}{\lambda}\right)^2 + \frac{1}{\lambda^2}\right)\left(e^{-\lambda B_i} - e^{-\lambda C_i}\right), \end{aligned} \tag{2.38}$$

where $C_i \in \left\{t_1^{ON} + t_1^{OFF,e}, t_1^{ON} + t_1^{OFF,e} + t_2^{OFF} + t_2^{ON}, \ldots\right\}$. C_i are also shown in Fig. 2.1. Summing over the all Y_{ON}^2 terms in Eq.(2.31) yields

$$\begin{aligned} E[Y_{ON}^2 \mid t_s] &= \sum_i \left(C_i - B_i\right)^2 e^{-\lambda B_i} + 2\left(E[T_2] - \frac{1}{\lambda}\right)\sum_i \left(C_i - B_i\right) e^{-\lambda B_i} \\ + & \left(\left(E[T_2] - \frac{1}{\lambda}\right)^2 + \frac{1}{\lambda^2}\right)\sum_i \left(e^{-\lambda B_i} - e^{-\lambda C_i}\right). \end{aligned} \tag{2.39}$$

The expectation over all the sample paths is

$$E\left[Y_{ON}^2\right] = \int_0^\infty E\left[Y_{ON}^2 \mid t_s\right] f_{t_s}\left(x_1^{OFF,e}, x_1^{ON}, \ldots\right) dx_1^{OFF,e} dx_1^{ON} \ldots \tag{2.40}$$

As before, after performing some simple calculus steps we get

$$E[Y_{ON}^2] = \frac{P}{1-p_1p_2}\left(E[T_{ON}^2] + 2E[T_{ON}]\left(E[T_2] - \frac{1}{\lambda}\right) + \left(\left(E[T_2] - \frac{1}{\lambda}\right)^2 + \frac{1}{\lambda^2}\right)(1-p_2)\right) \tag{2.41}$$

Finally by summing Eq.(2.41) with Eq.(2.36) we obtain Eq.(2.29). □

Lemma 4. *The second moment of queued packets $E[S^{''2}]$ is given by*

$$E[S^{''2}] = \Omega_1 + \Omega_2, \tag{2.42}$$

where

$$\Omega_1 = \Delta^2 + 2\Delta \int_0^\Delta x f_{OFF}^{(e)}(x)dx + 2\Delta \int_0^\Delta f_{OFF}^{(e)}(x)dx \left(\frac{1}{p}E[T_{ON}] + \left(\frac{1}{p}-1\right) E\left[T_{OFF} \mid T_{OFF} < \Delta\right]\right),$$

and

$$\begin{aligned}\Omega_2 &= \int_0^\Delta x^2 f_{OFF}^{(e)}(x)dx + 2\int_0^\Delta x f_{OFF}^{(e)}(x)dx \left(\frac{1}{p}E[T_{ON}] + \left(\frac{1}{p}-1\right) E[T_{OFF} \mid T_{OFF} < \Delta]\right) \\ &+ \int_0^\Delta f_{OFF}^{(e)}(x)dx \left(\frac{2(1-p)}{p^2}E[T_{ON}]^2 + \frac{1}{p}E[T_{ON}^2] + \frac{2}{p}E[T_{ON}]\left(\frac{1}{p}-1\right)E[T_{OFF} \mid T_{OFF} < \Delta]\right) \\ &+ \int_0^\Delta f_{OFF}^{(e)}(x)dx \left(\frac{(1-p)(2-p)}{p^2}E[T_{OFF} \mid T_{OFF} < \Delta]^2 + \left(\frac{1}{p}-1\right) Var\left(T_{OFF} \mid T_{OFF} < \Delta\right)\right)\end{aligned}$$

Proof. The random variable $S^{''}$, as seen before, is

$$S^{''} = \Delta + I\left(T_{OFF}^{(e)(f)} + \sum_{i=1}^{N} T_{ON} + \sum_{i=1}^{N-1} T_{OFF}^{(f)}\right). \tag{2.43}$$

Squaring both sides of previous equation we have

$$S^{''2} = \Delta^2 + 2\Delta I T_{OFF}^{(e)(f)} + 2\Delta I \sum_{i=1}^{N} T_{ON} + 2\Delta I \sum_{i=1}^{N-1} T_{OFF}^{(f)} + I\left(T_{OFF}^{(e)(f)} + \sum_{i=1}^{N} T_{ON} + \sum_{i=1}^{N-1} T_{OFF}^{(f)}\right)^2. \tag{2.44}$$

Taking the expectation of both sides of Eq.(2.44) we have

$$E\left[S^{''2}\right] = \Omega_1 + \Omega_2. \tag{2.45}$$

In Eq.(2.45)

$$\Omega_1 = E\left[\Delta^2 + 2\Delta I T_{OFF}^{(e)(f)} + 2\Delta I \sum_{i=1}^{N} T_{ON} + 2\Delta I \sum_{i=1}^{N-1} T_{OFF}^{(f)}\right],$$

which is equivalent to

$$\Omega_1 = \Delta^2 + 2\Delta E[I T_{OFF}^{(e)(f)}] + 2\Delta E\left[I \sum_{i=1}^{N} T_{ON}\right] + 2\Delta E\left[I \sum_{i=1}^{N-1} T_{OFF}^{(f)}\right].$$

Beside the random variables I and $T_{OFF}^{(e)(f)}$ which are not independent, all the other random variables are independent. Hence,

$$\Omega_1 = \Delta^2 + 2\Delta E\left[I T_{OFF}^{(e)(f)}\right] + 2\Delta E[I] E\left[\sum_{i=1}^{N} T_{ON}\right] + 2\Delta E[I]\left[\sum_{i=1}^{N-1} T_{OFF}^{(f)}\right]. \tag{2.46}$$

In the last equation $E\left[IT_{OFF}^{(e)(f)}\right] = \int_0^\Delta x f_{OFF}^{(e)}(x)dx$. The expectation of the random variable I is $E[I] = P\left(T_{OFF}^{(e)} < \Delta\right) = \int_0^\Delta f_{OFF}^{(e)}(x)dx$. By using the two previous expressions and the Wald's equality in Eq.(2.46) we get

$$\begin{aligned}\Omega_1 &= \Delta^2 + 2\Delta \int_0^\Delta x f_{OFF}^{(e)}(x)dx + 2\Delta \int_0^\Delta f_{OFF}^{(e)}(x)dx \frac{1}{p} E[T_{ON}] \\ + &\quad 2\Delta \int_0^\Delta f_{OFF}^{(e)}(x)dx \cdot \left(\frac{1}{p} - 1\right) E\left[T_{OFF} \mid T_{OFF} < \Delta\right]. \end{aligned} \tag{2.47}$$

The other part of Eq.(2.45) is given by

$$\Omega_2 = E\left[I\left(T_{OFF}^{(e)(f)} + \sum_{i=1}^{N} T_{ON} + \sum_{i=1}^{N-1} T_{OFF}^{(f)}\right)^2\right]. \tag{2.48}$$

After taking the square of the right-side of Eq.(2.48) we have

$$\begin{aligned}\Omega_2 &= E\left[IT_{OFF}^{(e)(f)2}\right] + E[I]E\left[\left(\sum_{i=1}^{N} T_{ON}\right)^2\right] + E[I]E\left[\left(\sum_{i=1}^{N-1} T_{OFF}^{(f)}\right)^2\right] \\ + &\quad 2\left[IT_{OFF}^{(e)(f)}\right] E\left[\sum_{i=1}^{N-1} T_{OFF}^{(f)}\right] + 2E[I]E\left[\sum_{i=1}^{N} T_{ON}\right] E\left[\sum_{i=1}^{N-1} T_{OFF}^{(f)}\right]. \end{aligned} \tag{2.49}$$

In Eq.(2.49)

$$E\left[IT_{OFF}^{(e)(f)2}\right] = \int_0^\Delta x^2 f_{OFF}^{(e)}(x)dx.$$

Using Wald's equation in Eq.(2.49), similarly as in Eq.(2.46) we have

$$\begin{aligned}\Omega_2 &= \int_0^\Delta x^2 f_{OFF}^{(e)}(x)dx + 2\int_0^\Delta x f_{OFF}^{(e)}(x)dx \left(\frac{1}{p}E[T_{ON}] + \left(\frac{1}{p} - 1\right) E[T_{OFF} \mid T_{OFF} < \Delta]\right) \\ + &\quad \int_0^\Delta f_{OFF}^{(e)}(x)dx \left(\frac{2(1-p)}{p^2} E[T_{ON}]^2 + \frac{1}{p}E[T_{ON}^2] + \frac{2}{p}E[T_{ON}]\left(\frac{1}{p} - 1\right) E[T_{OFF} \mid T_{OFF} < \Delta]\right) \\ + &\quad \int_0^\Delta f_{OFF}^{(e)}(x)dx \left(\frac{(1-p)(2-p)}{p^2} E[T_{OFF} \mid T_{OFF} < \Delta]^2 + \left(\frac{1}{p} - 1\right) Var\left(T_{OFF} \mid T_{OFF} < \Delta\right)\right)\end{aligned}$$

Finally, summing the last equation with Eq.(2.47), we get Eq.(2.42). □

Our analytical results suggest that the exact SU performance has an intricate dependence on PU characteristics that goes beyond channel utilization. At the same time, the key additional statistics needed are the second (and in congested cases) the third moments of PU active and idle periods (the success probabilities could be approximated using second moments through Chebyshev's inequality). This implies that by collecting such statistics for different channels, an SU can use our result to evaluate each channel's predicted performance, and choose according to application needs. In the next section, we validate our analysis, and also explore the effect of active and idle period variability on performance.

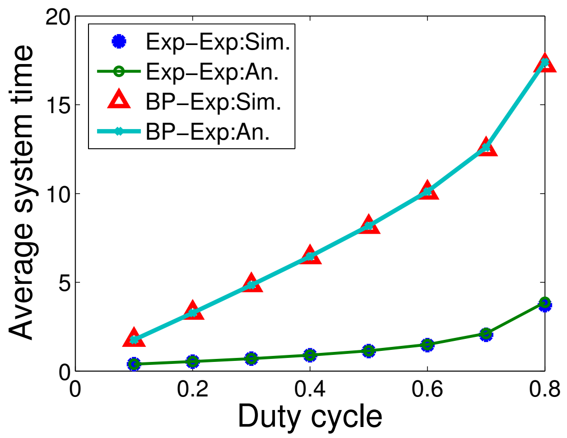

Figure 2.3: System time for exp-OFF.

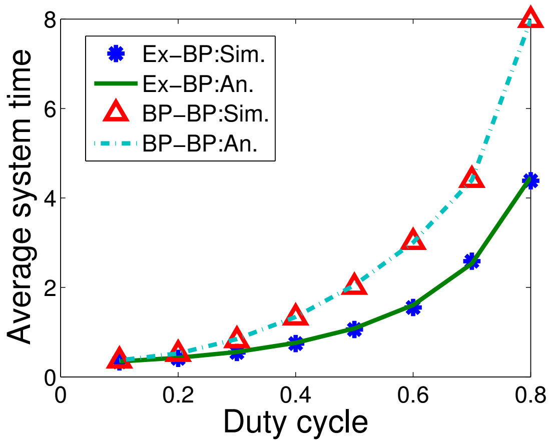

Figure 2.4: System time for BP-OFF.

2.3 Performance analysis

The traditional metric for characterizing the PU activity is the duty cycle. It is defined as $\frac{E[T_{ON}]}{E[T_{ON}]+E[T_{OFF}]}$. Unless otherwise stated, in all the scenarios below the packet size is taken to be 0.25, although other values of Δ lead to the same conclusions. To validate our theory against simulations we take combinations of exponentially (Exp) and Bounded Pareto (BP) distributed ON-OFF periods (as an example of "heavy-tailed" distributions). For the Bounded Pareto distribution, we take the lower bound $L = 0.215$, upper bound $H = 400$, and the shape parameter $\alpha = 1.2$.

Fig. 2.3 shows the average packet delay in a cognitive network for two different primary user activity scenarios for exponentially distributed OFF periods. The arrival rate is $\lambda = 0.1$. For the exp-exp distributions a low primary user activity (duty cycle of 0.2) gives a utilization of 0.05. When the duty cycle is 0.8, the utilization is 0.29. For the BP ON periods, lower relative primary user activity of 0.2 corresponds to a utilization of 0.09, and higher relative licensed user activity of 0.8 to a server utilization of 0.41. Different values of duty cycle give different levels of utilization, since the mean service time depends on the values of $E[T_{ON}]$ and $E[T_{OFF}]$. i.e. of duty cycle. The first thing to observe is a good match between theory and simulations. Furthermore, we can also observe that higher duty cycle implies higher delays, which is expected because the higher the duty cycle is, the primary user is more active, and there is less time for the cognitive user to operate. We can also see that for the same average ON and OFF durations (the same duty cycle), the delays are higher when the primary user has busy periods with higher variability. This is the first interesting conclusion that comes out of our model. Despite the two channels looking similar, from the point of view of average PU activity, variability can further affect delays. This is reminiscent of the inspection paradox [20], albeit the dynamics of Equations (2.1) and (2.15) are in fact more complex.

Fig. 2.4 shows the packet delays for Bounded Pareto distributed OFF periods. The arrival rate is low (0.01). This arrival rate corresponds to sparse traffic. As we can observe from Fig. 2.4, there is a good match between theory and simulations for the generic distributed OFF periods, also.

So far, we have considered some standard distributions for the general ON and OFF periods,

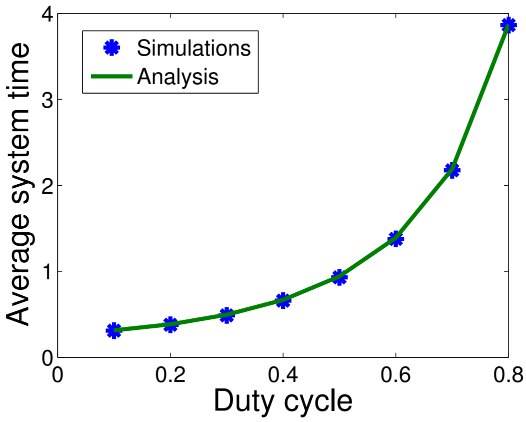

Figure 2.5: WLAN primary user.

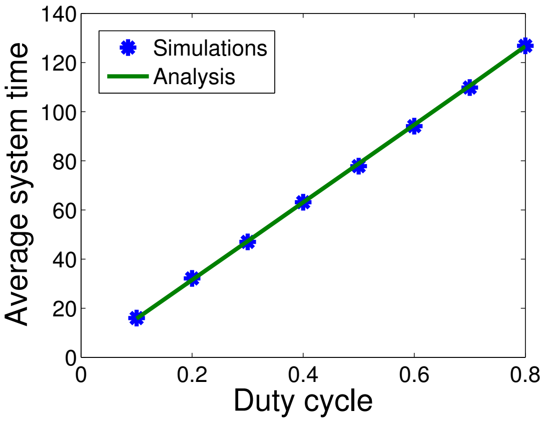

Figure 2.6: Cellular primary user.

to see the effect of high or low variance. We are also interested to see how our model can predict performance under more “realistic” PU patterns. To this end, we consider two recently proposed models for PU activity, one for cellular channels [18] and one for WiFi channels [19]. We have tried to implement the proposed models according to the respective descriptions, although some details are not specified there.

Fig. 2.5 shows the packet delay incurred by a WLAN network. The distributions for this simulation are taken from [19]. The ON periods are deterministic, while OFF periods have bimodal distributions. The arrival rate is 0.01. As we can see our theory provides a very good match with simulations.

Fig. 2.6 shows the packet delay in a cellular network. The model description from [18] is being used. The ON periods underlay a multimodal distribution, while the OFF periods are exponentially distributed. The packet size is 0.01. The arrival rate is $\lambda = 0.1$. For a duty cycle of 0.8, this arrival rate corresponds to an (maximum) utilization of 0.74. More details about the two models can be found in [18] and [19].

We have established so far that (i) our analytical model correctly predicts performance in all generic PU channels considered, and under various levels of congestion (we note here that we have performed a large number of other scenarios, with similar conclusions), and (ii) that even if two channels have similar *average* PU activity, variability (e.g. PU traffic burstiness) can further degrade performance. We now go a step further and consider a channel A with high average PU activity (duty cycle of 0.6) and exponentially distributed ON (activity) durations with variance equal to 1. We put it “against” a channel B with much lower PU activity (duty cycle 0.3), but log normally distributed ON periods (which have a heavier tail than exponential). Keeping the mean of the ON period unchanged, and increasing the variance of the log normal distribution gives us an interesting insight into the effect of both PU average activity and variability on cognitive user performance.

Table 2.2 displays the ratio of $\frac{\text{SU delay on channel B}}{\text{SU delay on channel A}}$.

We can observe from Table 2.2 that for similar variance, channel B which is less “busy” is better. However, by increasing the variance, the ratio keeps growing and exceeds 3 for a variance higher than 50. In fact, in theory, this difference can become arbitrarily high (i.e. for real heavy-tailed distributions, like Pareto with parameter < 2). We can thus conclude that, contrary to

Table 2.2: The ratio of delays for two different channels.

Variance for CH B	1	10	20	30	40	50	60
Ratio of Delays	0.5	1.2	1.8	2.3	2.6	3	3.2

common belief that tends to consider an underutilized part of the spectrum as “good spectrum” for cognitive users, the impact of variability of the primary user activity is much more important than the duty cycle itself. This could be key for example in delay-sensitive, but low throughput applications like M2M.

2.4 Related work

Some interesting works also model the PU user with a stochastic ON-OFF process, but assume a 2-state Markov Chain for it [25, 26]. Service times are derived from their 2-state Markov chain, and system times using an $M/G^Y/1$ system with bulk departures. While the exponential assumption is more convenient for analysis, it turns out to be inaccurate for both cellular [27], and WiFi systems [19]. As our analysis suggests, it can also lead to (arbitrarily) inaccurate predictions. In a recent work [28], a simple approximate model is used for delay prediction of an SU packet, assuming generic OFF periods, but this model also suffers from large inaccuracies, when the OFF periods are not exponential.

In order to depart from the strong exponential assumption, some recent works [21, 29, 30] have capitalized on the measurement-based study of [26], in which the Poisson approximation seems to be decent for call arrivals, but call duration is generically distributed. These works model SUs together with PUs, as an M/G/1 system with priorities and preemption. M/G/1 systems with priorities have been long analyzed (see e.g. [31]). Nevertheless, there are some important caveats in the above models. *First*, the system is preemptive-resume, that is, SU packets when preempted by a PU transmission can resume transmission from the point they stopped. In practice, SU packets will “collide” when a PU is detected, and have to *restart* in the next available cycle (possibly colliding again). Hence, this model is only approximately accurate for small packets (or long OFF periods). *Second*, while PU call arrivals might be approximated by a Poisson distribution, this *does not* mean that OFF periods are exponentially distributed, nor that the SU can directly infer the ON and OFF duration distributions. These depend on a number of system specifics and protocol interactions across the stack.

Consequently, we believe our model allows one to capture a much larger number of channels and PU activity patterns, and the SU to be able to measure and evaluate the predicted performance of a channel, without the need to know anything about the type and number of PUs multiplexed on a channel, protocols used, and PU job statistics.

2.5 Conclusion

In this chapter, we have proposed a queueing analytical model for the performance of cognitive users under generic ON-OFF primary channel models, and we have validated it against both synthetic and realistic PU channel models. We have shown that variability of primary user activity is very important, and often more important than the utilization itself. This is the key for the protocol design. The actual delay is a complex interplay between secondary traffic characteristics (intensity and packet sizes) and channel characteristics (1st and 2nd moments of idle and available durations of PU). In future work, we intend to extend our model to multihop networks, and also use our results to design better spectrum management, resource allocation and scheduling algorithms.

Chapter 3

To Scan or Not To Scan: The Effect of Channel Heterogeneity on Optimal Scanning Policies

Cognitive Networks have been proposed to opportunistically discover and exploit (temporarily) unused licensed spectrum bands. For a number of applications, high throughput is the key figure of merit, while the application is still elastic enough to be supported at different rates. To this end, the cognitive node will try to discover and pool together a number of (at the time available) primary channels to provide a given target throughput. When a single radio is used for both transmission and channel scanning, an interesting trade off arises: when one or more channels of the currently available ones are lost (e.g. primary user returns), should the node start scanning immediately or continue transmitting over the remaining channels. Using renewal-reward theory, we show that if the goal is to maximize the average (long-term) throughput, the answer to this question depends on the statistics of the channel availability periods. Specifically, for relatively homogeneous channels, we show that it is optimal to start scanning immediately, while for heterogeneous channels, it is often better to defer scanning, even if multiple channels are lost. Simulations for a range of different channel characteristics validate our analytical findings and suggest that triggering the scanning function at the right times, can improve performance considerably.

3.1 Introduction

In the last years there has been an increased spectrum demand from wireless applications and data. Due to the static spectrum allocation policy, spectrum scarcity is a major problem in today's wireless industry. At the same time, a large portion of the assigned spectrum is underutilized [1]. Dynamic spectrum access techniques have recently been proposed to overcome these problems, and Cognitive radio [32] is the key enabling technology. In a cognitive network, there exist licensed users which are provided the spectrum from the regulation authority, as well as users that utilize the spectrum opportunistically whenever they find it available. The former ones are known as primary users (PUs), while the later ones are known as cognitive (secondary) users (SUs). Various spectrum management functions [5] such as spectrum sensing, spectrum decision, spectrum sharing, and spectrum mobility, are then used to enable this seamless sharing.

A cognitive user can use one or more channels when no other primary users are active in that region. However, as soon as a primary user returns to the channel, the SU must interrupt its transmission at that moment[1]. Spectrum sensing [33] is a key feature of cognitive radios. It enables an SU to detect available frequency bands (not occupied by PUs), as well as the detection of the PU returning on that channel. Sensing techniques such as energy detection, matched filter approach, and waveform sampling [5] have been extensively studied.

Spectrum scanning, on the other hand, decides how to choose the channels to sense first, among the large number of possible frequencies a wideband cognitive radio could have access to. Intuitively, spending time to sense channels which have a high PU activity, and thus a low probability to be found available, can waste resources. As a result, there has been a considerable amount of work regarding the problem of channel scanning. Most of these works are concerned with finding the optimal sequence order to scan the channels, so that a certain parameter is optimized. In [23], authors try to minimize the scanning time, i.e. to get an available channel in the fastest way. In [34], the goal is to minimize the probability of link failure. Some other papers [35,36] propose greedy algorithms for the sequence of channel scanning for both reactive and pro-active spectrum handoff.

Cognitive radios are also capable of utilizing multiple channels in parallel ("pooling"), in order to increase the aggregate capacity [37]. In this context, an interesting problem arises when a single radio is used for both sensing and transmission, due to the need for low cost, energy, and/or complexity for small cognitive devices [37]. This means that a node cannot simultaneously transmit and sense, i.e. when sensing, it must interrupt its transmission. Should the node then initiate the scanning function immediately after is loses one of its channels currently in use? Or should it continue with the channels left, until one (or more) channels are lost? The tradeoff is the following: if it chooses to scan, then the capacity on the remaining available channels is wasted, and nothing useful is sent; if on the other hand it defers scanning, then it will transmit for some time at a lower rate (than it potentially could) *and* it will then have to scan longer to find *more* available channels. This latter intuition suggests that perhaps it is best to trigger scanning immediately, and maximize the amount of time one sends at the maximum rate.

Our goal in this chapter is to analyze this tradeoff as a function of the characteristics of all the channels available to the SU. We will assume an application that requires a maximum amount of throughput, but is elastic (i.e. could also operate at lower rates - e.g. streaming, file downloading, etc.), and we will try to maximize the average (long-term) throughput it is offered by the SU. To achieve this, we apply renewal reward theory [20]. Our contributions can be summarized as follows: (i) In accordance with the above intuition, we prove that when channel availability/idle periods are exponentially distributed with similar mean duration (*homogeneous*), it is indeed optimal to scan immediately when a single channel is lost; (ii) Contrary to the above intuition, when there is an initial scanning cost, or channel availability periods are heterogeneous, we prove that it is not optimal to scan immediately, but an optimal threshold exists; (iii) We provide a method to predict this threshold based on the channel characteristics, and show that this depends on the coefficient of variation between the mean availability periods of different channels; (iv) We provide an online algorithm that can take advantage of knowledge of which channel was just lost, to further improve performance; (v) Finally, using simulations, we provide evidence that our conclusions and algorithms are valid even when the channel availability durations follow a general (non-memoryless) distribution.

[1] We will assume throughout that the "interweave" model of sharing is used [1].

The chapter is organized as follows. In the next section, we discuss some related work. We present our problem setup and provide analytical results about the optimal scanning actions, for a specific class of channel availability distributions (exponential) in Section 3.3. We then validate our theory against simulations in Section 3.4, and also explore the case of general availability periods. We conclude our work in Section 3.5.

3.2 Related Work

There has been a large amount of research in the area of spectrum scanning in cognitive networks. Yet, most of these papers are concerned with determining the scanning sequence of the channels to be sensed, in order to optimize certain system parameters [38–41]. In some of those papers, some greedy algorithms relying on dynamic programming have been proposed. In some other, on the other hand, non-greedy algorithms are proposed that do not need the dynamic programming for the solution to be obtained. In [35], the authors propose an algorithm based on dynamic programming for the optimal scanning sequence. However, this work refers to proactive spectrum mobility schemes, where the spectrum sensing process is initiated before losing the channel. The goal of this algorithm is to minimize the *extended data delivery time*. This is the time needed to transmit a packet due to multiple spectrum handoffs. They also propose a sub-optimal greedy algorithm for the optimization of their goal. The authors claim that there are only six comparisons to be made in order to have the channel sequence that minimizes the delivery time. In [36], a similar algorithm has been proposed for the reactive spectrum mobility schemes (like the ones we consider). However, the proposed algorithm is not optimal and intends to minimize only the time it takes for a packet to be successfully transmitted assuming that there will be spectrum handoffs.

In a more related work [23], the authors propose a scanning sequence in order to minimize the average time to get a new channel, after the current one in use is being lost. They prove that this can be accomplished if channels are sorted in descending order of their probabilities to be found available (related to the duty cycle of that channel). However, they do not consider how long the acquired channel will be available for (dependent on both the first and second moments of the availability period [42]). To this end, in [34], a channel scanning sequence is proposed to minimize the probability of link failure between two cognitive users, i.e. a transmitter and a receiver during a transmission session. This can be done by sorting channels in descending order according to their expected excess idle periods durations.

Furthermore, [23] does not consider when the scanning process should be triggered, but rather begins from the starting time of such a scanning with a goal to acquire an amount of “missing” bandwidth, and investigates the order of scanning (backup) channels so as to minimize the delay to acquire the required bandwidth (possibly pooling together multiple channels). While finding additional channels fast is an important problem, it is also orthogonal to our problem. Different scanning sequences (optimal or suboptimal) could be incorporated into our model, as long as we know the expected amount of time to acquire L channels using that sequence.

Summarizing, unlike many related works, we consider the scanning problem in scenarios when multiple channels are pooled together. More importantly, instead of optimizing the scanning sequence when scanning is triggered, we investigate the complementary optimization problem of *when* to trigger this scanning function in order to optimize the (long-term) throughput that can be maintained by the SU. To our best knowledge, this is the first work in this direction.

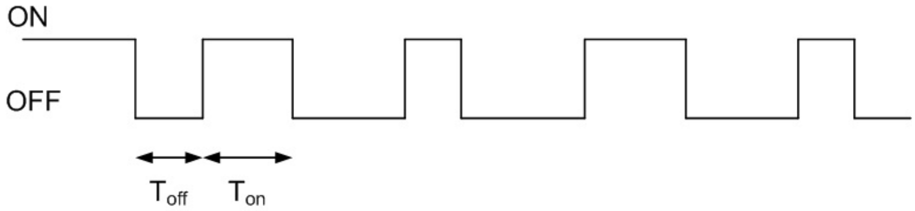

Figure 3.1: The channel model.

3.3 Analysis

3.3.1 Problem Setup

Consider a single channel used by one or more primary users. We assume that the state of this channel can be either active ('ON'), i.e. the primary user is active, or idle ('OFF'). A secondary user can only transmit during an OFF period (interweave model). The exact duration of the ON and OFF periods depends on the user behavior, the type of traffic, system details and protocol interaction. Such details cannot be known by the secondary user.

We will model the ON-OFF activity pattern of PUs as an alternating renewal process [20] $\left(T_{ON}^{(n)}, T_{OFF}^{(n)}\right), n \geq 1$, as shown in Figure 3.1. n denotes the number of ON-OFF cycles elapsed until time t. The duration of any ON (OFF) period $T_{ON}^{(n)}$ $\left(T_{OFF}^{(n)}\right)$, is a random variable distributed according to some probability distribution F_{ON} (F_{OFF}), and independently of other ON or OFF periods.

We assume that the spectrum, available to the SU, is divided into channels with identical bandwidths B. Furthermore, we consider applications that require a (maximum) bandwidth of $C = NB$, but could also operate at lower rates. This could be the case for example, for file downloading or P2P applications [43], where C (or N) is dictated by the cognitive radio technology (not more than N channels can be pooled together), or multimedia streaming applications with multiple rate coding [37]. In addition to these N channels used, we assume there are also other (backup) channels that can be scanned and used (if available), when one or more of the used channels are lost. After losing a certain channel, we add it to the list of backup channels.

In this context, we will consider the following problem: *A threshold value L is chosen, such that after* any *L channels (out of the N) are lost we must start scanning and regain L new channels. What is a good value for L?* Note that the choice of $L = 1$ corresponds to the usual case where scanning starts immediately after any of the available channels is lost.

We make here some final remarks about the practical implications of the above problem setting. We assume that a single radio and antenna is used. Although such a cognitive radio is usually wideband (e.g. some 10s of MHz), allowing for more than 1 channel to be grouped together, even if not adjacent (using a digital filter), it can only transmit or scan at any time, but not both (since scanning usually requires changing the center frequency as well, and introduces switching delays in the order of ms). In contrast, to detect that one or more channels out of the used ones is lost (i.e. a PU has started transmitting), we could just switch the radio periodically to receive mode (*on the same band*), take a short time sample (usually in the order of μs), and do an FFT to identify which band(s) out of the (up to) N used ones has high enough energy (implying activity). The energy threshold and the frequency of such sensing periods pose a

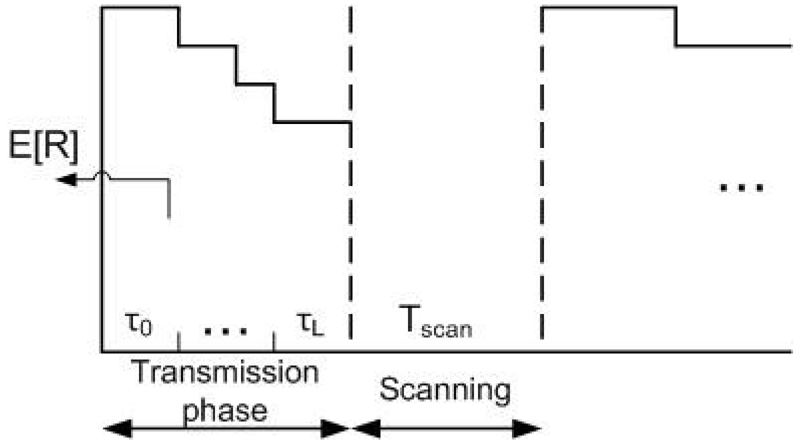

Figure 3.2: The transmission and scanning phases.

tradeoff between the time lost not transmitting and false positive/negative for PU detection, and is beyond the scope of this chapter. Yet, for the sake of the subsequent analysis, we can safely ignore these interleaved sensing periods as they are orders of magnitude shorter, and assume that PUs are detected correctly.

3.3.2 Analytical Model

Figure 3.2 illustrates our model, assuming that at each instant at most one channel can be lost. A cycle consists of the transmission phase and the scanning phase. τ_i represents the time during which $N-i$ channels are being used for transmission ($i < L$). The scanning phase represents the time needed to regain the L missing channels. The area under the stair-case curve represents the total amount of data transmitted during a cycle. Our goal is to maximize the long-term throughput, namely the size of this area up to a time t, when t is large.

To this end, we will use renewal-reward theory [20]. We model the process shown in Figure 3.2 as a renewal process, where a renewal occurs each time the necessary N channels are gathered and we can resume our transmission. A renewal cycle consists of two phases: transmission and scanning phase[2]. We can define a reward for each cycle as the amount of data transmitted during the transmission phase of the cycle.

If we denote by $R(t)$ the reward earned by time t, the average (long-term) throughput is the mean reward rate $\frac{R(t)}{t}$, which from renewal-reward theory we know it to be

$$lim_{t\to\infty}\frac{E\left[R\left(t\right)\right]}{t}=\frac{E\left[R\right]}{E\left[T\right]}. \tag{3.1}$$

In Eq.(3.1), $E\left[R\right]$ is the average reward per cycle, while $E\left[T\right]$ is the average cycle duration. Let's denote the reward rate by X. So, the average reward rate is given by

$$E[X]=\frac{E[R]}{E[T]}. \tag{3.2}$$

[2] We stress here that we do not claim this process to be a renewal process, but rather use the renewal-reward theory as a tool to derive analytical insight regarding the throughput achievable by different policies. This insight will be validated against simulations.

Our objective is to maximize the average reward rate, that is the average data rate per cycle, by choosing the right threshold L, as a function of the OFF (idle) period characteristics of the channels available to an SU node.

To calculate the values of $E[R]$ and $E[T]$, we need to understand the random variables τ_i. At the beginning of a transmission phase, N channels are available that were either already being used, or were found to be available during the scanning phase. As a result, the remaining availability time for each of these channels, say channel j, is an *excess* random variable (excess OFF period) denoted as $T^{(e)}_{OFF,j}$. Then, τ_0 corresponds to the time until *any* of the N channels is lost,

$$\tau_0 = \min\left(T^{(e)}_{OFF,1}, T^{(e)}_{OFF,2}, \dots T^{(e)}_{OFF,N}\right). \tag{3.3}$$

Similarly, τ_1 denotes the amount of time exactly $N-1$ channels are in use (time between the 1^{st} and the 2^{nd} channels are lost),

$$\tau_1 = \min\left(T^{(e)}_{OFF,1}, T^{(e)}_{OFF,2}, \dots T^{(e)}_{OFF,(N-1)}\right). \tag{3.4}$$

Similarly for the rest of τ_i.

We can thus express the average area below the curve (mean reward) as

$$E[R] = NB\sum_{i=0}^{L} E[\tau_i] - B\sum_{i=0}^{L} iE[\tau_i]. \tag{3.5}$$

The average cycle duration is

$$E[T] = \sum_{i=0}^{L} E[\tau_i] + E[T_{scan}(L)], \tag{3.6}$$

where $T_{scan}(L)$ is the scanning time needed to acquire the missing channels.

We can now say that our goal is to find the threshold value L that will maximize

$$E\left[X\right] = \frac{C\sum_{i=0}^{L} E[\tau_i] - B\sum_{i=0}^{L} iE[\tau_i]}{\sum_{i=0}^{L} E[\tau_i] + E[T_{scan}(L)]}. \tag{3.7}$$

This is quite involved in the general case. In the remainder, we will consider analytically the cases of arbitrary ON periods and exponential OFF periods (with the same or different mean durations). The assumption for exponentially distributed OFF periods is made for analytical tractability. Nevertheless, this assumption is often not far from reality, as measurements from [19,44] suggest that the OFF periods can be approximated quite well with exponential distributions. In Section 3.4, we will further consider generic OFF periods as well (with increasing and decreasing failure rates).

We make here a final remark about scanning. We will assume that the scanning periods for each channel are considered to be identical. Additional features could be included in our theory. The duration of a channel scanning period is T_I, and we also assume that during the scanning period the probability that other channels are lost is low; this is reasonable (and also supported by our simulations) since the scanning period is usually lower than the durations of the OFF periods for each channel. For example, assume that the scanning period is $T_I = 1$ ms. For a duty cycle (the ratio of time the primary user is active on that channel) of 0.5, on average we need to scan two channels in order to get one free channel. So, the average scanning time is $T_s = 2$ ms. If the OFF periods are exponentially distributed with mean 1 s, then the probability that an available channel will be lost while scanning is $P[T_{OFF} < T_s] = 1 - e^{-T_s} = 0.002$. While this value can be larger for some channels with shorter OFF periods, this is still small enough to ignore it in our analysis and only reintroduce this in simulations.

Finally, the duration of the scanning phase will also be a function of the number of channels that must be acquired:

$$E[T_{scan}] = l(L)T_I, \tag{3.8}$$

where $l(L)$ is the average number of channels that need to be scanned, in order to regain the L missing ones. It is easy to see that $l(L) \geq L$, since one might need to sense more than one channel to find one that is available. The exact function depends on the sequencing algorithm, and can be simply plugged into the above equations. Unless otherwise stated, we will assume w.l.o.g. throughout that $l(L)$ is linear, that is, if T_s is the total time to acquire one channel, then kT_s is the total time to acquire k channels.

3.3.3 Exponential IID OFF Periods

We will consider first the simpler case of PU activities having independent identically distributed (IID) OFF periods that are exponentially distributed, with mean $E[T_{OFF}] = \frac{1}{\lambda_{off}}$. The average durations of ON and OFF periods can be inferred in different ways [44]. Our first result is the following:

Result 2. *For channels with homogeneous (i.i.d.) exponentially distributed OFF periods it is always optimal to scan immediately after one channel is lost.*

To derive this, we start by the fact that for N independent exponentially distributed random variables X_1, X_2,..., X_N with parameters λ_1, λ_2,..., λ_N, the expectation of the minimum of these random variables is

$$E\left[\min\left(X_1, X_2, \ldots, X_N\right)\right] = \frac{1}{\lambda_1 + \lambda_2 + \ldots + \lambda_N}. \tag{3.9}$$

Since $\lambda_i = \lambda_{off}, \forall i$, it holds that

$$\begin{aligned} E[\tau_0] &= \frac{1}{N\lambda_{off}}, \\ E[\tau_1] &= \frac{1}{(N-1)\,\lambda_{off}}, \end{aligned}$$

and similarly for $E[\tau_i]$.

The mean reward for the duration τ_0 is

$$\frac{NB}{N\lambda_{off}} = \frac{B}{\lambda_{off}}. \tag{3.10}$$

Similarly, the mean reward for the duration τ_1 (i.e. if the node continues transmitting using the remaining $N-1$ channels) is

$$\frac{(N-1)B}{(N-1)\lambda_{off}} = \frac{B}{\lambda_{off}}. \tag{3.11}$$

It is easy to see then that, based on Eq.(3.7), the following inequality must hold so that postponing the scanning process after a channel is lost gives a higher expected rate per cycle, compared to immediately scanning

$$\frac{\frac{B}{\lambda_{off}}}{\frac{1}{N\lambda_{off}}+T_s} < \frac{\frac{B}{\lambda_{off}}+\frac{B}{\lambda_{off}}}{\frac{1}{N\lambda_{off}}+\frac{1}{(N-1)\lambda_{off}}+2T_s}. \tag{3.12}$$

After rearranging, this gives

$$N\lambda_{off} < (N-1)\lambda_{off}, \tag{3.13}$$

which can not be satisfied. This proves that we cannot increase our average throughput by stopping after 2 channels are lost (instead of 1). Since the above argument holds for any N, it is easy to see that we can only lose more by increasing L further, which proves our claim.

So far, we have considered that there is no initial cost for the scanning process. However, it is expected that switching to scanning mode (adjust the devices, determine the first channel to sense, etc.) will incur some initial setup cost, before the actual sensing process can commence for the first channel. In this case, we show that, depending on this initial cost, the optimal value for L (the scanning threshold) might be larger than 1.

Result 3. *For i.i.d. exponentially distributed OFF periods, where an initial scanning cost T_0 exists, scanning immediately ($L = 1$) is not optimal when*

$$T_0 \geq \frac{1}{N(N-1)\lambda_{off}}. \tag{3.14}$$

The analysis is exactly the same as before, except the factor T_0 added to the scanning time for both policies. Hence the inequality that needs to be satisfied is

$$\frac{\frac{B}{\lambda_{off}}}{\frac{1}{N\lambda_{off}}+T_0+T_s} < \frac{\frac{B}{\lambda_{off}}+\frac{B}{\lambda_{off}}}{\frac{1}{N\lambda_{off}}+\frac{1}{(N-1)\lambda_{off}}+T_0+2T_s}. \tag{3.15}$$

Rearranging again gives us the above value for T_0. When the setup (initial) scanning cost is higher than this value, then it is better to not scan immediately, so as to amortize this cost.

As explained earlier, we have assumed that the expected cost (time) to acquire k channels is linear in k. This could be the case for example, if the channels to be sensed during the scanning phase are picked randomly from the list of backup channels. If a better or optimal sequence is provided (e.g. as in [23]) then the time to get a second channel would be higher on average than the time to get the first channel (since, channels with high availability probability are scanned first). This could be easily included in our model by adding an extra cost Δ to the time to get the second channel in the above inequalities. Obviously, for the case of no initial scanning cost, this would not change Result 2. For the case of an initial scanning cost, it is easy to see that the required condition changes to

$$T_0 \geq \frac{1}{N(N-1)\lambda_{off}}+\Delta. \tag{3.16}$$

3.3.3.1 Finding an optimal threshold

We have so far proven conditions for which scanning immediately is optimal or not. In the case of IID exponential OFF periods, we can also find the optimal threshold explicitly. If we assume a threshold L, the average transmission time during a cycle is

$$\sum_{i=0}^{L} E[\tau_i] = \frac{1}{N\lambda_{off}} + \frac{1}{(N-1)\,\lambda_{off}} + \ldots + \frac{1}{(N-i)\,\lambda_{off}}. \tag{3.17}$$

This gives

$$\sum_{i=0}^{L} E[\tau_i] = \frac{1}{\lambda_{off}}\left(\frac{1}{N} + \frac{1}{N-1} + \ldots + \frac{1}{N-L}\right) = \frac{1}{\lambda_{off}}\sum_{i=0}^{L}\frac{1}{N-i}. \tag{3.18}$$

The last equation can be rewritten as

$$\sum_{i=0}^{L} E[\tau_i] = \frac{1}{\lambda_{off}}\left(\sum_{i=1}^{N}\frac{1}{i} - \sum_{i=1}^{N-L-1}\frac{1}{i}\right). \tag{3.19}$$

Using the Euler's approximation [45] $H_n = \sum_{i=1}^{n}\frac{1}{i} = \ln n + \frac{1}{2n} + 0.57721$ we obtain

$$\sum_{i=0}^{L} E[\tau_i] = \frac{1}{\lambda_{off}}\left(H_N - H_{N-L-1}\right). \tag{3.20}$$

After simple calculus operation we can obtain

$$\sum_{i=0}^{L} E[\tau_i] = \frac{1}{\lambda_{off}}\left(\ln\frac{N}{N-L-1} - \frac{1}{2}\frac{L+1}{N\,(N-L-1)}\right). \tag{3.21}$$

The expected amount of transmitted data before L channels are lost is

$$\frac{1}{N\lambda_{off}}C + \frac{1}{(N-1)\,\lambda_{off}}(C-B) + \ldots + \frac{1}{(N-L)\,\lambda_{off}}(C-LB). \tag{3.22}$$

Since $C = NB$, after rearranging the last equation we obtain

$$E[R] = \frac{(L+1)\,B}{\lambda_{off}}. \tag{3.23}$$

Replacing Eq.(3.21), Eq.(3.23) and Eq.(3.8) into Eq.(3.7), we have

$$E[X] = \frac{\frac{(L+1)B}{\lambda_{off}}}{\frac{1}{\lambda_{off}}\left(\ln\frac{N}{N-L-1} - \frac{1}{2}\frac{L+1}{N(N-L-1)}\right) + l(L)T_I}. \tag{3.24}$$

We could thus differentiate $E[X]$ above, with respect to L (and round to the closest integer), in order to find the optimal value, when one exists.

3.3.4 Heterogeneous exponentially distributed channels

We have shown that, unless a large enough initial setup cost for the scanning function exists, when channel OFF periods are exponential and of similar duration it is always optimal to scan immediately when any channel is lost. Here, we will consider the more realistic case of heterogeneous availability periods. We will still consider exponential durations to maintain tractability, and assume that the rate λ_i[3] for the OFF duration of channel i is drawn from some distribution ($G(\lambda)$). The main result of this section is the following:

Result 4. *For heterogeneous exponentially distributed OFF periods, it is not always optimal to scan immediately. Instead, the transmission should proceed with the remaining channels, if they satisfy the following relation:*

$$c_\lambda^2 \geq 1, \tag{3.25}$$

where c_λ is the (sample) coefficient of variation of the OFF periods for the channels remaining available.

There are two interesting things to notice about the above result. First, unlike the homogeneous channel case, scanning immediately is suboptimal and a better threshold than $L = 1$ can be found. Second, this threshold increases when the variability of the pool of channels available to the SU increases. We will now go ahead and derive this result.

Since the OFF periods are exponential, we can derive the average throughput for $L = 1$ (scan immediately) as

$$E[X_1] = \frac{\frac{BN}{\sum_i \lambda_i}}{\frac{1}{\sum_i \lambda_i} + T_s}. \tag{3.26}$$

Assume now that a channel with rate λ_{lost} is lost first. This is a random variable with probability $\frac{\lambda_{lost}}{\sum_i \lambda_i}$. If the node continues transmitting until another channel is lost, then the average throughput achieved is given by

$$E[X_2|\lambda_{lost}] = \frac{\frac{BN}{\sum_i \lambda_i} + \frac{B(N-1)}{\sum_i \lambda_i - \lambda_{lost}}}{\frac{1}{\sum_i \lambda_i} + \frac{1}{\sum_i \lambda_i - \lambda_{lost}} + 2T_s}. \tag{3.27}$$

We would like to uncondition and get the mean value of $E[X_2]$. If we denote $E[X_2|\lambda_{lost}]$ as a function $f(\lambda_{lost})$, then we would like to know $E[f(\lambda_{lost})]$. Rearranging Eq.(3.27), we obtain

$$E[f\left(\lambda_{lost}\right)] = E\left[\frac{a - b\lambda_{lost}}{c - d\lambda_{lost}}\right], \tag{3.28}$$

where

$$a = B(2N-1)\sum_i \lambda_i, \qquad b = BN,$$

$$c = 2\sum_i \lambda_i \left(1 + T_s \sum_i \lambda_i\right), \quad d = 1 + 2T_s \sum_i \lambda_i.$$

[3] From now on, we will denote $\lambda_{off,i}$ simply as λ_i.

The above expectation depends on the distribution of λ_i values and is not easy to calculate in the general case. We can use however Jensen's inequality to convert this to a function of $E[\lambda_{lost}]$ itself. We thus check for the convexity of the function $f(\lambda_{lost})$. It's second derivative is

$$f''(\lambda_{lost}) = \frac{2d(ad-bc)}{(c-d\lambda_{lost})^3}. \tag{3.29}$$

We can easily prove that the term $c - d\lambda_{lost}$ is always larger than 0. For convexity, the term in the numerator must satisfy

$$ad - bc = B\sum_i \lambda_i \left[2(N-1)T_s\sum_i \lambda_i - 1\right] > 0.$$

This implies that the function $f(\lambda_{lost})$ is always convex under the condition $T_s\sum_i \lambda_i \geq \frac{1}{2(N-1)}$. This condition could be satisfied when the number of channels N used is large and/or when there are enough "bad" channels in the pool of available ones, so that the product $T_s\sum_i \lambda_i$ exceeds 1 ("bad" channels would correspond to channels with quick fluctuations between ON and OFF states, unlike e.g. the case of TV white spaces).

Using Jensen's inequality [20], and the convexity of $f(\lambda_{lost})$ ($E[f(x)] \geq f(E[x])$), we can now replace the *necessary* condition for the $L = 1$ to not be optimal

$$E[X_1] \leq E[f(\lambda_{lost})], \tag{3.30}$$

with the *sufficient* condition

$$E[X_1] \leq f(E[\lambda_{lost}]). \tag{3.31}$$

This yields

$$\frac{a - bE[\lambda_{lost}]}{c - dE[\lambda_{lost}]} \geq \frac{BN}{1 + T_s\sum_i \lambda_i}. \tag{3.32}$$

After solving the above inequality we obtain

$$E[\lambda_{lost}] \geq \frac{2bc - a(d+1)}{b(d-1)}, \tag{3.33}$$

and replacing the expressions for a, b, c, and d we get

$$E[\lambda_{lost}] \geq \frac{1 + T_s\sum_i \lambda_i}{NT_s}. \tag{3.34}$$

This already provides a condition related to the statistics of the existing and the lost channels. If it holds that $T_s\sum_i \lambda_i >> 1$, from Eq.(3.34) we get

$$E[\lambda_{lost}] \geq \frac{\sum_i \lambda_i}{N}. \tag{3.35}$$

However, since the probability of losing a channel is proportional to its OFF period rate λ, $E[\lambda_{lost}]$ is always greater than the sample average of the current N channels $\frac{\sum_i \lambda_i}{N}$, and it is always better (on average) to continue transmitting.

The second case of interest is when $T_s \sum_i \lambda_i \geq \frac{1}{2(N-1)}$, but it is not larger than 1. The average rate of the first lost channel is

$$E[\lambda_{lost}] = \sum_i \lambda_i \frac{\lambda_i}{\sum_i \lambda_i} = \frac{\sum_i \lambda_i^2}{\sum_i \lambda_i}. \tag{3.36}$$

Replacing Eq.(3.36) into Eq.(3.34) we have

$$\frac{\sum_i \lambda_i^2}{\sum_i \lambda_i} \geq \frac{1 + T_s \sum_i \lambda_i}{NT_s}. \tag{3.37}$$

After some simple calculus steps we get

$$\frac{\frac{1}{N} \sum_i \lambda_i^2}{\frac{1}{N^2} \left(\sum_i \lambda_i\right)^2} \geq \frac{1 + T_s \sum_i \lambda_i}{T_s \sum_i \lambda_i}. \tag{3.38}$$

On the left hand side of Eq.(3.38) we have the ratio of the second moment of the rates of the OFF periods for the channels in use and the square of their means. This can be written through the coefficient of variation $c_v^2 = \frac{Var(X)}{(E[X])^2}$ as

$$c_\lambda^2 + 1 \geq \frac{1}{T_s \sum_i \lambda_i} + 1 \geq 2. \tag{3.39}$$

This means that the coefficient of variation must fulfill the condition

$$c_\lambda^2 \geq 1. \tag{3.40}$$

It is important to note that the above conditions we derived are sufficient, but not necessary. We may be allowing a lot of slack through the step where we use Jensen's inequality. However, our goal was to show that, contrary to the homogeneous case, scanning less frequently (but for more channels) can improve performance. The fact that we can find reasonable regimes for channel characteristics where this holds, despite the stricter condition, only strengthens our argument. Furthermore, while the above result proves conditions for the existence of a (non-trivial) scanning threshold, it could also be used in a recursive manner to derive the optimal threshold. We can already predict from the above theory that this threshold will increase when the variability of the channels accessible by the SU increases. In the remainder we will call this approach the *offline* algorithm. This algorithm maximizes the expected throughput, given that the threshold must be chosen only once and at the beginning.

3.3.5 The online (adaptive) algorithm

Based on the above insight, we can also take advantage of knowledge of which channel was in fact lost and propose a simple algorithm that we expect to further improve the average throughput. We will call this the *online* algorithm. The purpose of this algorithm is to decide about the threshold "on the fly", depending on which channel we lose. The condition regarding the lost channel can be derived departing from the comparison of $E[X_1] < X_2$ from Eq.(3.26) and Eq.(3.27). Solving this inequality, as before, we obtain

$$\lambda_{lost} > \frac{1 + T_s \sum_{i=1}^{N} \lambda_i}{NT_s}. \tag{3.41}$$

If the channel we lose has duration rates for the OFF periods that are larger than the right hand side of Eq.(3.41), then it is better to keep transmitting after losing that channel. Similarly, if for the second lost channel the above inequality still holds, transmission is not stopped. If after losing the ith channel, the above inequality does not hold any more, then we start to scan until we regain the missing i channels.

A special case would be if there exists the relation $T_s \sum_i \lambda_i >> 1$. Then, Eq.(3.41) reduces to

$$\lambda_{lost} > \frac{\sum_i \lambda_i}{N}. \tag{3.42}$$

Then, if the lost channel is worse (has higher λ) than the average of the channels in use, then it is better to resume transmission. This is rather intuitive, since by getting rid of channels with low availability we are left only with the ones with high availability, and if we decide to scan instead, the new channel is expected to only make the mean rate of the channel pool worse, *on average*. On the other hand, if we lose a good channel then we should interrupt, since the rest of the channels are probably worse than average and scanning could improve this situation.

There are a number of important design decisions involved in implementing this algorithm in practice, but this is beyond the scope of this thesis.

3.4 Simulation results

3.4.1 Homogeneous channels

We have already seen in Section 3.3 that for i.i.d. exponentially distributed OFF periods, it is better to start scanning immediately. So, there is no threshold value that provides higher data rate. In that case, it is better to scan as soon as we lose the first channel. Fig. 3.3 shows the average throughput for a radio that uses $N = 5$ channels to transmit its data. The ON periods for the primary users activities in all the channels are identical and chosen from a uniform distribution in the range between 10 and 20 s^{-1}. The OFF periods are also identical for all the channels with the average duration of $\frac{1}{\lambda_{off}} = 1\,s$. The number of backup channels is large enough. Unless otherwise stated, the channel bandwidth is $1\,MHz$, while the sensing period for each channel is $1ms$ for all the scenarios. We scan channels from the backup list sequentially one at a time. The x axis gives the threshold value used in each case. From Fig. 3.3 we can observe that the best result is achieved if we start scanning right away after losing the first channel, and by increasing the threshold value the average data rate decreases. Throughout this section we will user MATLAB as simulation tool.

Fig. 3.4 illustrates the dependence of the average data rate per cycle on the threshold value for i.i.d. exponentially distributed OFF periods, where $\lambda_{off} = 1\,s^{-1}$. We show the effect for a larger number of channels in use ($N = 10$). The other parameters are identical to the scenario of Fig. 3.3. Here also, the plot proves our claim that there is no threshold value that provides better results in terms of the data rate for exponentially i.i.d. channels.

3.4.2 Initial cost

As we have shown in Section 3.3 when an initial (setup) scanning cost exists, the optimal threshold might move from $L = 1$ to higher values. Fig. 3.5 illustrates the case when the initial cost value is 0.06 s, that is slightly higher than the theoretical minimum value obtained from

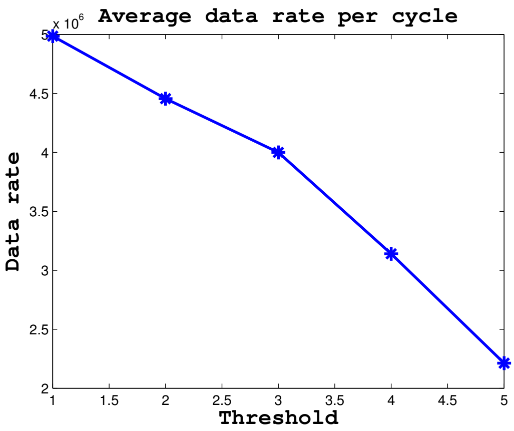

Figure 3.3: Homogeneous exp. channels.

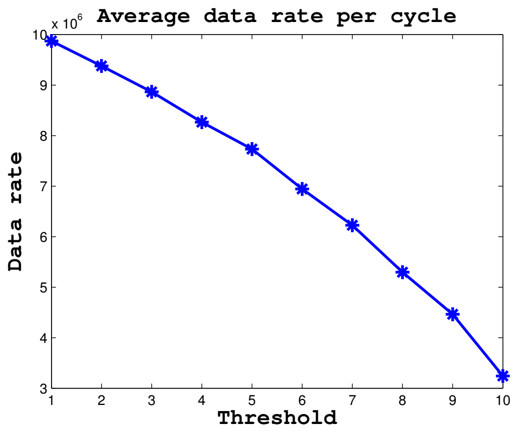

Figure 3.4: Homogeneous exp. channels.

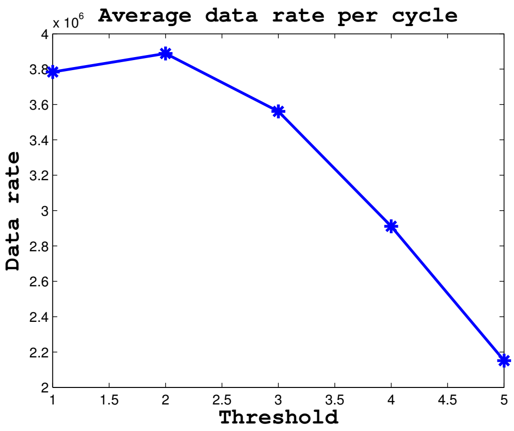

Figure 3.5: Initial scanning cost.

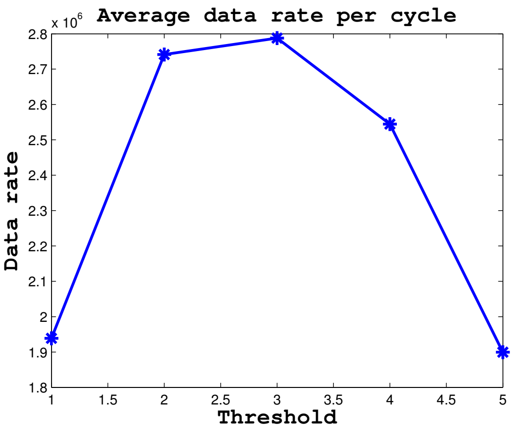

Figure 3.6: Initial scanning cost.

our theory Eq.(3.14). The other parameters are exactly the same as those from the scenario which corresponds to Fig. 3.3. From Fig. 3.5 we can see that the threshold value that provides the best result is $L = 2$.

By intuition, if the initial scanning cost is higher, then the ideal threshold value moves towards higher values. If the initial cost is 0.26 s, then the dependence of the average data rate per cycle on the threshold value is depicted in Fig. 3.6. All the simulation parameters are the same as in Fig. 3.5. From this plot we can observe that the best result is achieved for a threshold value of $L = 3$. As expected, by increasing the initial cost, the ideal threshold value has been increased, too.

3.4.3 Heterogeneous channels

Now, we will consider the case with channels where the OFF periods are not drawn from identically exponential distributions. Fig. 3.7 shows the average throughput for the scenario with 15 channels. The OFF periods of these channels are independent exponentially distributed with different λ's. The values of the channel parameters λ_i are drawn from a uniform distribution in

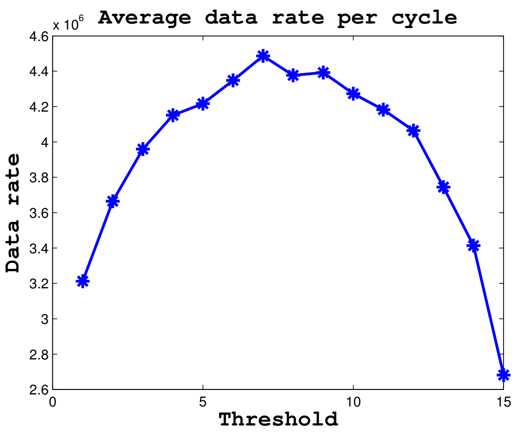

Figure 3.7: Heterogeneous exp. channels.

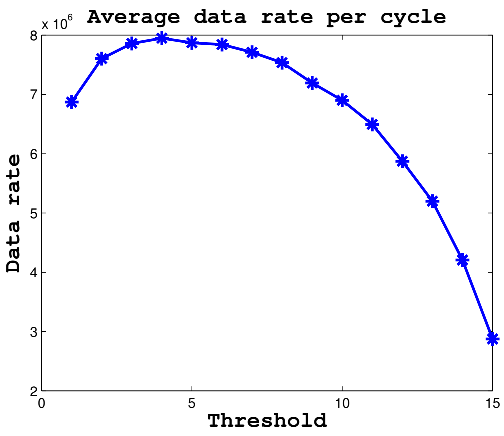

Figure 3.8: Heterogeneous exp. channels.

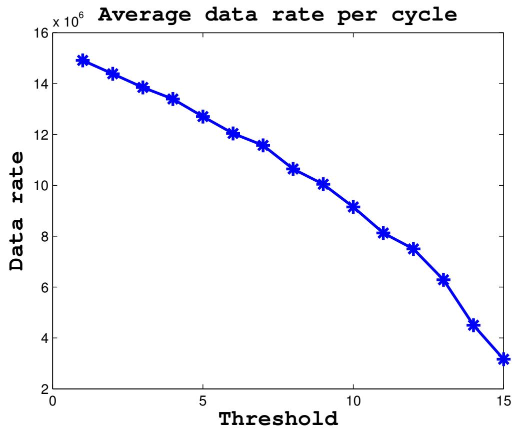

Figure 3.9: Heterogenerous exp. channels with low var.

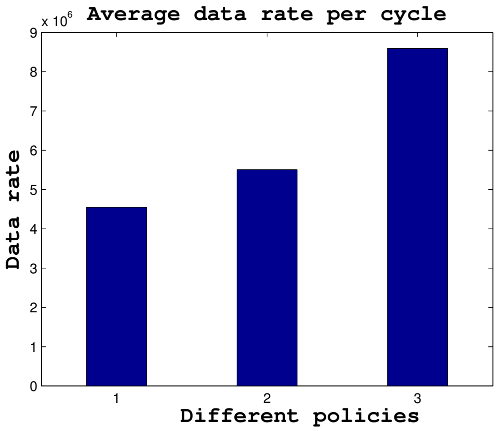

Figure 3.10: Throughput for different policies.

the interval $[1, 200]\ s^{-1}$. There is no initial scanning cost, and the sensing period is $1\,ms$. From Fig. 3.7 we can observe that the ideal threshold value is $L = 7$. This was also expected, because from that point on the condition $T_s \sum_i \lambda_i \geq \frac{1}{2(N-1)}$, does not hold anymore.

Fig. 3.8 illustrates the average throughput for the same simulation scenario as in Fig. 3.7, with the exception that the upper bound of λ's is $50 s^{-1}$. The ideal threshold value is now $L = 4$. From this plot we can observe that the throughput is higher compared to the case of Fig. 3.7, because we have channels with higher durations of their OFF periods. It is also interesting to observe that the threshold value that provides maximum data rates has decreased. This is because of the fact that the variability of channels has been decreased compared to the previous case. Or, in other words, there is less difference between the channels in terms of their utility.

To further enhance our previous claim, let us consider the case when we have only good channels, i.e. channels whose OFF periods are with relatively long durations. For that purpose we will again consider the scenario in which we use 15 channels, with heterogeneous exponentially distributed OFF periods. The inverses of the mean durations for the OFF periods of different channels (λ_i), are drawn from a Bounded Pareto distribution with shape parameter $\alpha = 1.2$,

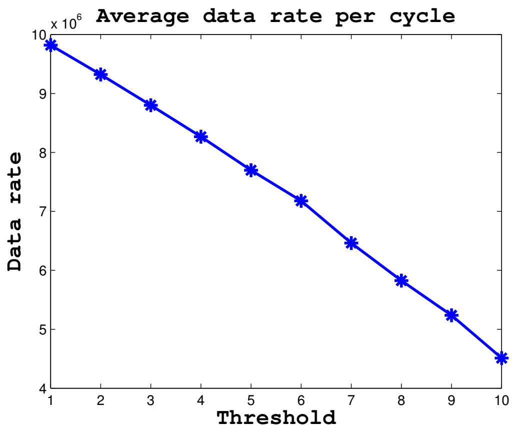

Figure 3.11: Homogeneous uniform.

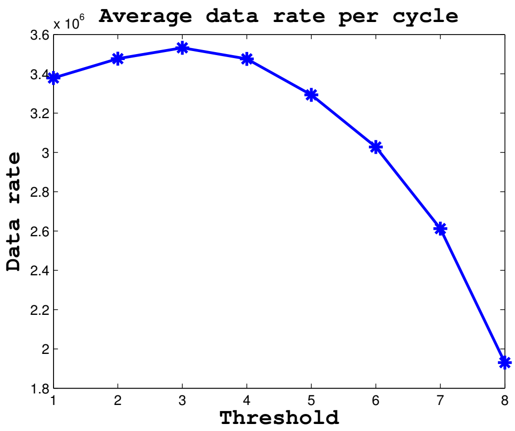

Figure 3.12: Heterogeneous uniform.

lower bound of $0.1\,s^{-1}$, and upper bound equal to $1\,s^{-1}$. We pick λ's from a different distribution to give another proof that our theory is correct. The average throughput for this case is shown in Fig. 3.9. For these parameter values the data rate is decreased by increasing the threshold value. So, the best thing to do is to start scanning after the first lost channel. This is a consequence of the low coefficient of variation for the channels' OFF period durations.

3.4.4 The adaptive algorithm

Having validated our analytical conclusions that an optimal threshold exists when channels are heterogeneous, we now turn our attention to the proposed adaptive (online) algorithm to see if it can further improve performance. The number of channels in use is 15. The rates of the OFF periods are drawn from the uniform distribution in the interval 0.1 to 100 s^{-1}. Fig. 3.10 shows the data rates for three cases:

1. Threshold $L = 1$,
2. The ideal threshold for the offline algorithm, and
3. Online (adaptive) algorithm with the variable threshold.

The (offline) ideal threshold that provides maximum data rate (case 2) is $L = 5$. From Fig. 3.10 we can observe that our proposed online algorithm provides the highest average data rate per cycle. This is a consequence of the fact that we make the decision when to start scanning on the fly, depending on which channel we have lost. For the offline algorithm though, we must make the decision in advance, based on the average characteristics of the pool of channels and the expected quality of the lost ones. It can happen that a channel with average long duration is lost before a bad channel, which gives rise to this difference between the offline and online algorithms. We can also observe that the data rate is lowest if we trigger the scanning immediately.

3.4.5 Generic OFF periods

So far, we have only considered channels with exponential OFF periods in both analysis and simulations. While it is difficult to investigate analytically threshold policies for generic OFF

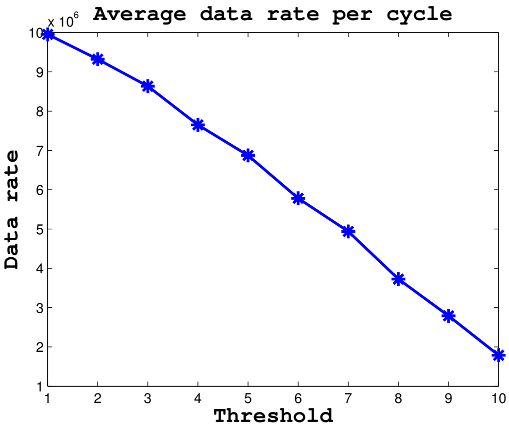

Figure 3.13: Homogeneous Pareto.

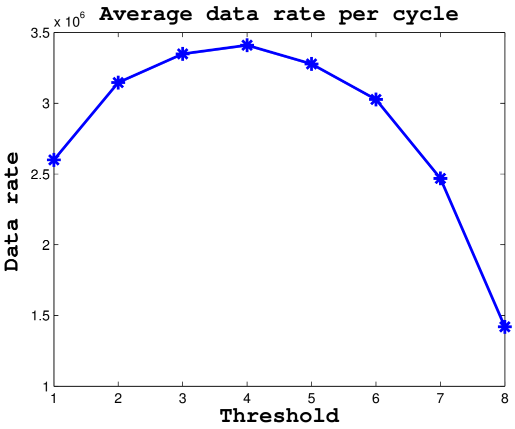

Figure 3.14: Heterogeneous Pareto.

periods, we can do so using simulations. Fig. 3.11 shows the average data rate for different threshold values for homogeneous OFF periods (uniformly distributed) with average duration of 1 s, and average ON duration of 0.1 s. There are 10 channels in use. We can observe that there does not exist a threshold value that provides higher average data rate compared to immediate scanning. This is consistent with the exponential homogeneous scenario, where it is also better to scan immediately. In Fig. 3.12 there are 8 channels in use, each with uniformly distributed OFF periods. Furthermore, this is a heterogeneous scenario where the mean OFF period for each channel can be in the range 0.1 to 100 s^{-1}. For this case, as in the exponential heterogeneous case, an optimal threshold larger than one ($L = 3$) exists. The optimal value of the threshold depends on the failure rate of the distribution for the OFF periods. The uniform distribution has an increasing failure rate, so as time goes on, the probability that a channel in use will be available in the future is lower. Hence, we would expect that the threshold value is not too high ($L = 3$), since we lose quite quickly the good channels.

Fig. 3.13 shows the average throughput for homogeneous Pareto distributed OFF periods with shape parameter $\alpha = 1.2$, and with the rest of parameters identical to the scenario of Fig. 3.11. Here also, we can see that no ideal threshold value higher than 1 exists. This also means that it is not a sufficient condition for a threshold to exist, that the durations of homogeneous OFF periods to be drawn from a distribution with decreasing failure rate. Figure 3.14 shows the throughput for heterogeneous Pareto distributed OFF periods with the same average as in Fig. 3.12, and with identical number of channels ($N = 8$). The ideal threshold value in this case is $L = 4$, which is larger than that in Fig. 3.12. This can be explained as follows. Pareto distribution belongs to the class of distributions with decreasing failure rate, as opposed to the uniform distribution which has increasing arrival rate. This means that as time goes on, the chances to lose a good channel are lower and lower.

3.5 Conclusion

In this chapter, we have analyzed the spectrum scanning process in cognitive radio networks and explored the ways to maximize the average throughput rate. We have introduced the notion of a threshold value as the number of channels we are allowed to lose, before the initiation of the scanning procedure. This threshold value provides the best results in terms of the average data rate. It is proven that no such threshold value exists for the case of homogeneous (i.i.d.) channels if there is no initial cost to be paid at the beginning of the scanning process. However, this value exists for heterogeneous independent channels and its value depends on the variability of the OFF periods of the channels in use. We have also proposed an adaptive algorithm that determines the moments when to stop transmission depending on which channel was lost, and have shown by simulations that this algorithm provides the highest throughput. In future work, we intend to extend our theoretical analysis to the generic OFF periods, as well as to consider joint scanning and sequencing optimization.

Chapter 4

Stay or Switch? Analysis and Comparison of Interweave and Underlay Spectrum Access in Cognitive Radio Networks

Cognitive Networks have been proposed to opportunistically discover and exploit licensed spectrum bands, in which the secondary users' (SU) activity is subordinated to primary users (PU). Depending on the nature of the interaction between the SU and PU, there are two frequently encountered types of spectrum access: *underlay* and *interweave*. While a lot of research effort has been devoted to each mode, there is no clear consensus about which type of access performs better in different scenarios and for different metrics. To this end, in this chapter we approach this question analytically, and provide closed-form expressions that allow one to compare the performance of the two types of access under a common network setup. We focus on two key metrics, delay and throughput, which we analyze using queueing theory and renewal-reward theory. This allows an SU to decide when one type of access technique or the other would provide better performance, as a function of the metric of interest and key network parameters. What is more, based on this analysis, we propose dynamic (hybrid) policies, that can decide at any point to switch from the one type of access to the other, offering up to another 50% of additional performance improvement, compared to the optimal "static" policy in the scenario at hand. We provide extensive validation results using a wide range of realistic simulation scenarios.

4.1 Introduction

Lately, we are witnessing a tremendous increase in the number of data-enabled wireless devices (smartphones, tablets, etc.) as well as in the applications and services that they provide. Coupled with the equally large market growth envisioned for the numerous small and large "things" requiring wireless connectivity [46], this creates a huge pressure on wireless network operators, and a resulting increase in spectrum demand.

Because of this, and due to the static spectrum allocation policies followed by authorities worldwide, spectrum scarcity has become a major problem in today's wireless industry. Nevertheless, measurements of the utilization of licensed wireless spectrum in fact reveal that the

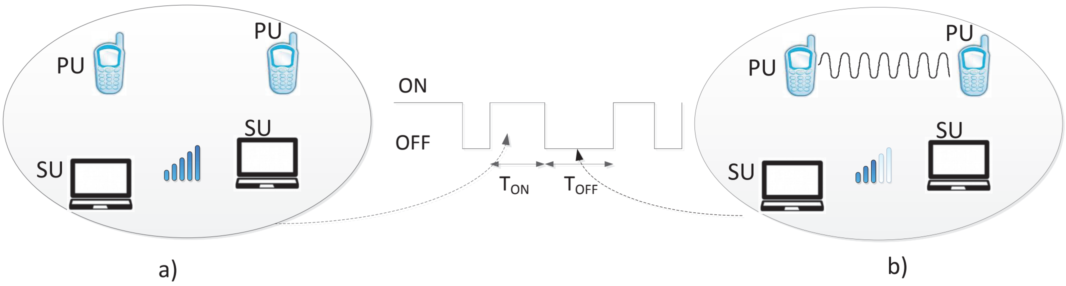

Figure 4.1: The illustration of the underlay mode with: a) idle PU (high periods), b) active PU (low periods).

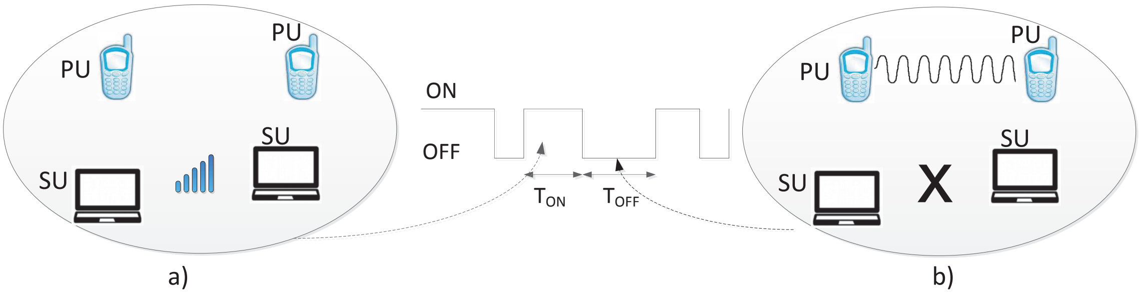

Figure 4.2: The illustration of the interweave mode with: a) idle PU (high periods), b) active PU (low periods).

available spectrum is rather under-utilized, exhibiting high variability across space, frequency and time [1].

To address this issue, dynamic spectrum access techniques have recently been proposed, with cognitive radio (CR) as its key technology [5]. In a cognitive network, there exist licensed users, known as primary users (PU), which are provided the spectrum from the regulation authority, as well as unlicensed users that are known as cognitive or secondary users (SU) utilizing the spectrum opportunistically. Cognitive users are subordinated to primary users' activity. Hence, they have to adapt their transmission parameters, so that there are no impairments on PU Quality of Service (QoS).

One of the main functions of CRs is spectrum access [5]. Spectrum access is very important to prevent potential collisions between the SUs and PUs. Spectrum access techniques can be classified as: *underlay*, *interweave*, and *overlay*. In this chapter, we are concerned only with the first two techniques. In the underlay mode (Fig. 4.1), the SU reduces the transmission power when a PU is utilizing a given channel such that the maximum interference level a PU can tolerate is not exceeded. In the interweave mode (Fig. 4.2), the cognitive user can transmit only when there is no PU, with the maximum power in accordance with the spectral mask. Whenever a PU claims a channel back, the SU must immediately cease its transmission and look for another *white space*, i.e. a part of the spectrum that is currently not utilized by its PU. In the overlay mode, the cognitive user serves as a relay to a licensed user and in turn the PU allows it to access to a portion of its spectrum. However, the necessity of complete channel knowledge from both PU and SU increases complexity and makes this mode less attractive.

A large number of works exist for both underlay and interweave access in CRNs [23, 47]. While some arguments for the one or the other exist (often related to the potential harm to PUs [48]), there is little consensus regarding which mode would offer the best performance to SUs.

While the possibility of transmitting without interruptions (usually causing issues to higher layer protocols) is certainly an advantage that underlay access offers, there are also some drawbacks associated with it. First, the user in this mode can transmit with maximum rate only when the PU is silent. These periods of time can be much shorter than the periods with PU, especially when dealing with high duty cycle licensed users. Furthermore, if that PU is located in the vicinity of the SU, the SU's transmission rate might have to be significantly reduced. This could significantly reduce the effective (average) transmission rate and the resulting throughput.

Contrary to this, in interweave mode the SU can look for another idle channel when the PU arrives (possibly with a lower duty cycle) and start transmitting again at full power, possibly improving the average (long-term) throughput. Yet, the intermittent nature of communication relying on interweave access may delay some application flows (e.g. a request to transmit a short file, fetching a web page, etc.) significantly, if they happen to arrive while the SU is scanning for a new available channel. Such delays can be exacerbated not only if the SU resides in a relatively busy part of the spectrum (e.g. urban areas at "peak" hours), but even if the variability of this scanning time is high (e.g. sometimes a new channel is found quickly, but sometimes the SU might be stuck scanning for a long time).

Based on the above discussion, it is obvious that there are a number of tradeoffs involved, and it is not easy to say, a priori, which mode of spectrum access would perform better in a given scenario. The relative performance has a close dependence not only on specific network parameters (e.g. PU duty cycle, allowed transmission power, etc.), but also on the performance metric of interest, the type of SU traffic (sparse, frequent), size of requests, and even higher order statistics of key parameters, such as the time to find a new white space.

To this end, in this chapter we approach this problem using an analytical framework to evaluate the individual performance of underlay and interweave access, as well as to compare them in a range of settings. We focus on two metrics, delay, which we analyze using a queueing theoretic framework, and long-term throughput, where we apply renewal-reward theory. Our contributions can be summarized as follows:
(i) We derive closed-form expressions for the expected delay for underlay and interweave spectrum access as a function of key network parameters (average PU idle time, transmission rates, scanning time statistics), and user traffic statistics (traffic intensity, file size). This allows us to directly compare the performance of the two, and derive the conditions that would make the one or the other preferable. Finally, we also use these insights to propose a "hybrid" policy, that can switch between the two dynamically, in order to further improve performance (Section 4.3);
(ii) We perform the same steps for the case of average (long term) data rate for both modes, namely providing closed-form expressions for individual performances, comparing analytically, and finally optimizing (Section 4.4);
(iii) Using a wide range of realistic simulation scenarios, we validate our analytical predictions extensively, explore the conditions under which underlay or interweave policies perform better, and show when the dynamic policies can indeed offer additional performance improvements (Section 4.5).

4.2 Performance modeling of spectrum access

We are interested in two figures: the *delay* and the *throughput*. For the former one, we assume that traffic flows arrive randomly as a Poisson process with rate λ. The file sizes are assumed to be exponentially distributed. When a file arrives to find another file in the system, it will be queued. We consider First Come First Served (FCFS) order of service. The total time a file spends in the system is the sum of the service and queueing time, and is referred to as the *system time*. We use also the term *transmission delay* interchangeably with system time.

On the other hand, we define throughput as the long term average data rate. For the throughout calculations we assume that there is always backlogged traffic at the SU. This is very reasonable for the high utilization regime (when a user is very active).

4.2.1 Problem setup for underlay access

In underlay access, the SU can transmit at full power when there is no PU communicating on that channel. When the PU resumes its transmission, the SU has to reduce its transmission power, so that there are no impairments on the PU transmission quality. Although the actual power allowed depends on the primary and the interfering channel quality (distance, LoS, etc.), we will assume for simplicity that the SU power can vary between two levels: "high" power when there is no PU, and "low" power when there is PU activity (e.g. perceived as an average value). Our analytical model could also be extended to multiple SU power levels, but the expressions would be considerably more difficult to interpret.

Consider a channel used by one or more PU. The occupancy of that channel can be modeled as an ON-OFF alternating renewal process [20] $\left(T_{ON}^{i}, T_{OFF}^{i}\right),\ i \geq 1$, as shown in Fig. 4.1. ON periods represent the absence of the PU on that channel, while the OFF periods denote the periods of time with active PU. i denotes the number of ON-OFF cycles elapsed until time t. Unless otherwise stated, the duration of any ON period T_{ON} is assumed to be exponentially distributed with parameter η_H, and is independent of the duration of any other ON or OFF period. Similarly, the duration of an OFF period is also assumed to be exponential, but with parameter η_L. While this assumption is necessary for the tractability of the delay analysis, as we shall see, we consider generic ON/OFF periods in the throughput analysis. Generic ON/OFF distributions could be introduced also in our delay analysis by considering phase-type distributions and matrix analytic methods [49]. However, such methods only yield numerical solutions, that do not allow for direct analytical performance comparisons. What is more, simulation results (Section 4.5) suggest that, even for generic ON/OFF period distributions, the accuracy of our predictions is sufficient.

The data transmission rate during the ON periods is denoted with μ_H, while during OFF periods the date rate is μ_L. The actual values for these depend on technology, channel bandwidth, coding and modulation, etc. However, since the allowed transmission power during OFF periods is higher, it holds that $\mu_H > \mu_L$.

4.2.2 Problem setup for interweave access

In the interweave mode, the SU can transmit only when there is no PU activity (ON periods). It is again assumed that the periods with no PU activity are exponentially distributed with parameter η_H, and the data rate is μ_H.

However, after the arrival of a PU in the channel (at the end of that ON period), the SU does not continue transmitting (at lower power), as in the underlay case, but starts looking for another available channel. As soon as it finds one, it resumes transmission at full power (i.e. with rate μ_H). Consequently, we can again model this system with an alternating renewal process. However, OFF periods now correspond to scanning intervals during which no data can be transmitted, i.e. $\mu_L = 0$.[1] Hence, it changes its operation to the *scanning* mode. During the scanning mode (i.e. during an OFF period), the SU moves to a new channel and senses it for some time. If available, it resides there and goes back to the transmission mode. Otherwise, it switches to another frequency and senses another channel, and so on until finally an available channel is found[2], and the transmission process is resumed. Hence, the scanning time corresponding to one channel is actually the sum of the sensing time (T_I) and the switching delay (T_{switch}) introduced. So, the total scanning time can be expressed as

$$T_s = N\left(T_I + T_{switch}\right), \tag{4.1}$$

where N is a random variable denoting the number of channels a SU has to sense until it finds the first available. We assume that sensing time/channel is much shorter than the ON and OFF periods.

The switching delay while moving from the channel with frequency f_s to the channel with frequency f_d can be expressed as [51]

$$T_{switch} = \beta\frac{|f_d - f_s|}{\delta}, \tag{4.2}$$

where β is the delay to move to the first contiguous channel, and is hardware dependent [51], while δ denotes the frequency separation between two neighboring channels.

By looking at Eq.(4.1), we can infer the following. If the probabilities of finding each channel available are independent and almost equal, we can say that the random variable N is geometrically distributed. If we further assume that the switching time is the same when moving from one to another channel, and the well known fact that the geometric distribution can be obtained by rounding the exponential, we can infer that the scanning time can be approximated by an exponential distribution. For that purpose exponential scanning time distribution is assumed first. However, the nature of scanning time distribution depends heavily on the availability of the backup channels and on the frequency distance between them. Under most scenarios (high discrepancy on the availability probabilities between different channels, very low duty cycle of all the channels, the existence of available channels in the more remote parts of the spectrum etc.), the exponential assumption on the scanning time will not hold. For that reason, we also analyze the system for the scenarios when the switching time underlies high and low variance distributions.

Let's assume that the eligible channels have roughly the same duty cycles (% of time the PU is active on a channel) that are very low (sporadic activity of PUs), and that they are

[1] We assume that in both modes a single radio and antenna is used. Hence, a SU can only transmit or scan at any time, but not both. In contrast, to detect that the PU is back, we could just switch the radio periodically to receive mode, take a short time sample (in the order of μs) and do energy detection, to see if there is a PU signal [50]. Since we assume that the sensing time is much shorter than the actual durations of ON and OFF periods, we can safely ignore these sensing periods. However, the switching time is usually much higher than the sensing time (order of *ms* or even seconds) and cannot be ignored.

[2] We assume that the SU chooses channels sequentially from a list [51].

"neighbors" in the frequency context. Under this scenario, we would expect that the time needed to find an available channel is almost constant (most of the time only one channel needs to be sensed) and will not deviate much from the average value, $E[T_s]$. Hence, in this case the scanning time distribution would have a low variance (lower than the exponential distribution). A convenient and generic way to model such low variance distributions is by using a k-stage Erlang distribution [24].

As opposed to the previous scenario, there might be a number of channels with high duty cycles, and there might be one or few channels with much lower duty cycle located further away in the spectrum. Although there is a low probability, it may happen that all these channels located close to each other are busy, and the SU ends up searching the channel that is far away in the spectrum band. Since the switching time is proportional to the frequency difference, the scanning time will be considerably larger. So, there is a chance that some of the scanning time samples will deviate from the average to a considerable extent. Hence, in those cases the scanning time distribution can be considered as heavy-tailed, and the exponential assumption cannot capture that behavior. For that purpose, we model the scanning time with a hyperexponential distribution, in which with a probability p the scanning time will be exponentially distributed with parameter η_L, and with a small probability $1 - p$ it will be exponentially distributed with parameter η_V. Note that $\eta_V << \eta_L$.

To support our aforementioned claims, we consider two scenarios. First, we assume that there is a group of $M = 20$ channels, which are close to each other in the spectrum. The duty cycles of the channels are all low (0.2), PU activities are i.i.d., $T_I = 1$ ms, and $T_{switch} = 10$ ms. Fig. 4.3 shows the complementary cumulative distribution function (CCDF) of the scanning time durations. On the same plot, the CCDFs of the exponential and Erlang distributions for $k = 3$ and $k = 6$, are also shown. The plot demonstrates that the exponential distribution cannot really capture the behavior of the system. Instead, an Erlang distribution needs to be used. Similar conclusion for the inability of the exponential distribution to capture the scanning time can be inferred from Fig. 4.4. As opposed to the previous scenario, in this case the channels have very high duty cycles (0.8), and the switching time between this group and the group with 2 channels (with a duty cycle of 0.2) is 0.5s. The hyperexponential distribution (shown also in the plot) has the following parameters: $p = 0.2$, $\lambda_1 = 90$, $\lambda_2 = 3$. As can be seen, the hyperexponential distribution can capture to a better extent this scenario.

The above results highlight the need to consider more general distributions for scanning times directly in our analysis. Surprisingly, as we shall see, closed-form results for the system delay can still be found for such generic scanning times.

Before proceeding any further, we summarize in Table 4.1 some useful notation that will be used throughout the rest of the chapter.

4.3 Delay analysis of underlay and interweave access

In this section we will derive the formulas for the average file delay for interweave and underlay access. For the former one, we perform the analysis over different scanning time distributions: exponential, k-stage Erlang and hyper exponential. For that purpose we use 2D Markov chain models, and the Probability Generating Functions (PGF) approach to derive the delay.

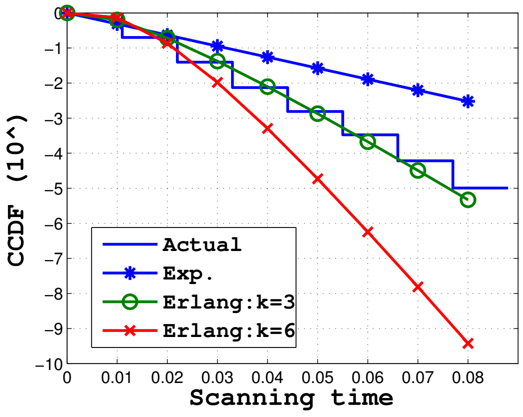

Figure 4.3: The distribution of the scanning time for low PU duty cycles.

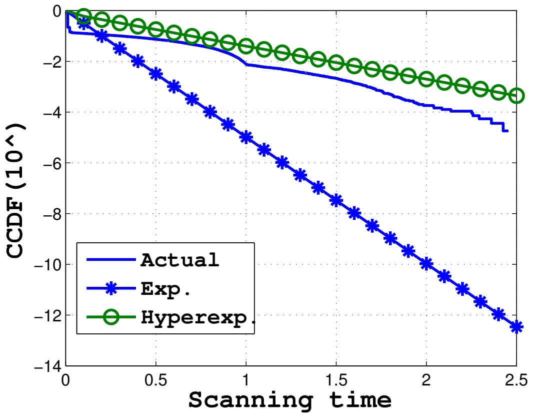

Figure 4.4: The distribution of the scanning time for two group of channels.

Table 4.1: Variables and Shorthand Notation.

Variable	Definition/Description
T_{ON}	Duration of PU idle periods
T_{OFF}	Duration of PU busy periods or scanning time
λ	Average file arrival rate at the mobile user
$\pi_{i,L}$	Probability of finding i files in the OFF (low) period
$\pi_{i,H}$	Probability of finding i files in the ON (high) period
$\pi_{i,V}$	Probability of finding i files in the OFF (V-state) period
$\eta_H(\eta_L)$	The rate of leaving the ON (OFF) state
η_V	The rate of leaving the V-state
$\mu_H(\mu_L)$	The service rate while in a high (low) state
Δ	The average file size
$E[T]$	The average system (transmission) time

4.3.1 Delay analysis for interweave access

Exponential scanning time. Due to the assumptions made in Section 4.2.2 about the exponential distributions, we can model our system with a 2D Markov chain, shown in Fig. 4.5. Each state in this chain indicates the number of files present in the system and the presence (lower states) or absence (upper states) of the PU. $\pi_{i,L}$ ($\pi_{i,H}$) denotes the stationary probability of finding i files when there is (not) a PU active on that channel. The transition rates η_H and η_L are the parameters of the exponentially distributed ON and OFF periods. While in the upper parts of the chain there are transitions between states (i, H) and $(i-1, H)$ with rates equal to μ_H, in the lower part (corresponding to the active PU), there is no transition going from state (i, L) to $(i-1, L)$. This is a consequence of the inability of the SU to transmit while scanning. The transition rate from low to high periods is η_L, with exponential scanning time of average duration $E[T_s] = \frac{1}{\eta_L}$.

Theorem 5. *The average file delay in a cognitive radio network with interweave spectrum access*

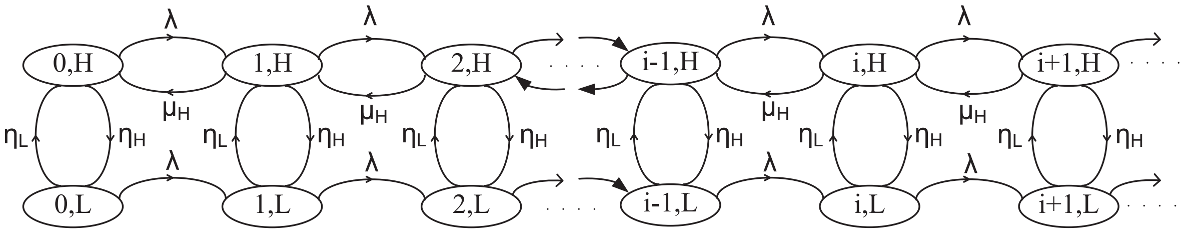

Figure 4.5: The 2D Markov chain for exponentially distributed scanning time.

and exponentially distributed scanning time is

$$E[T_{exp}] = \frac{\eta_H(\eta_H+\mu_H)(E[T_s])^2 + 2\eta_H E[T_s] + 1}{(1+\eta_H E[T_s])(\mu_H - \lambda - \lambda\eta_H E[T_s])}. \tag{4.3}$$

Proof. Writing the balance equations for this chain, we have

$$\begin{aligned}
\pi_{0,L}\left(\lambda+\eta_L\right) &= \eta_H \pi_{0,H} && (4.4)\\
\pi_{i,L}\left(\lambda+\eta_L\right) &= \lambda\pi_{i-1,L} + \eta_H\pi_{i,H},\ i \geq 1 && (4.5)\\
\pi_{0,H}\left(\eta_H+\lambda\right) &= \eta_L\pi_{0,L} + \mu_H\pi_{1,H} && (4.6)\\
\pi_{i,H}\left(\lambda+\mu_H+\eta_H\right) &= \lambda\pi_{i-1,H} + \eta_L\pi_{i,L} + \mu_H\pi_{i+1,H},\ i \geq 1 && (4.7)
\end{aligned}$$

We define the probability generating functions for this chain as

$$G_L(z) = \sum_{i=0}^{\infty} \pi_{i,L} z^i, \text{and } G_H(z) = \sum_{i=0}^{\infty} \pi_{i,H} z^i, |z| \leq 1, z \in C.$$

Multiplying Eq.(4.5) with z^i and adding it to Eq.(4.4), we obtain

$$(\lambda+\eta_L)\sum_{i=0}^{\infty}\pi_{i,L}z^i = \eta_H\sum_{i=0}^{\infty}\pi_{i,H}z^i + \lambda\sum_{i=1}^{\infty}\pi_{i-1,L}z^i, \tag{4.8}$$

which leads to

$$\left[\lambda(1-z)+\eta_L\right] G_L(z) = \eta_H G_H(z). \tag{4.9}$$

Similarly, multiplying Eq.(4.7) with z^i and summing with Eq.(4.6), we get

$$(\eta_H+\lambda)\sum_{i=0}^{\infty}\pi_{i,H}z^i + \mu_H\sum_{i=1}^{\infty}\pi_{i,H}z^i = \eta_L\sum_{i=0}^{\infty}\pi_{i,L}z^i + \lambda\sum_{i=1}^{\infty}\pi_{i-1,H}z^i + \mu_H\sum_{i=0}^{\infty}\mu_H\pi_{i+1,H}z^i. \tag{4.10}$$

Eq.(4.10) results in

$$\left[\lambda z(1-z) + \mu_H(z-1) + \eta_H z\right] G_H(z) - \eta_L z G_L(z) = \mu_H\pi_{0,H}(z-1). \tag{4.11}$$

Solving the system of equations Eq.(4.9) and Eq.(4.11) leads to

$$G_L(z) = \frac{\mu_H\pi_{0,H}(z-1)}{\left\{\frac{1}{\eta_H}\left[\lambda z(1-z)+\mu_H(z-1)+\eta_H z\right]\left[\lambda(1-z)+\eta_L\right] - z\eta_L\right\}}, \tag{4.12}$$

$$G_H(z) = \frac{1}{\eta_H}\left[\lambda(1-z)+\eta_L\right]G_L(z). \tag{4.13}$$

The only unknown in Eq.(4.12) is $\pi_{0,H}$ (the stationary probability of SU having zero files while there is no PU activity). To find it, we proceed as following. First, we write the balance equation across the vertical cut between states (i, L) and (i, H) on one side, and $(i, L+1)$ and $(i, H+1)$ on the other. This gives

$$\lambda\pi_{i,L} + \lambda\pi_{i,H} = \mu_H\pi_{i+1,H}. \tag{4.14}$$

After summing over all the values of i, we have

$$\lambda = \mu_H\left[G_H(1) - \pi_{0,H}\right]. \tag{4.15}$$

In Eq.(4.15), $G_H(1) = \sum\limits_{i=0}^{\infty}\pi_{i,H}$ is the probability of finding the system in the high state. So, for the zero probability we have

$$\pi_{0,H} = \frac{\mu_H G_H(1) - \lambda}{\mu_H}. \tag{4.16}$$

Replacing $z = 1$ into Eq.(4.9) gives $G_L(1) = \frac{\eta_H}{\eta_L}G_H(1)$. It is also evident that $G_L(1)+G_H(1) = 1$, resulting in

$$G_H(1) = \frac{1}{1+\frac{\eta_H}{\eta_L}}. \tag{4.17}$$

Replacing Eq.(4.17) into Eq.(4.16) enables us to find $\pi_{0,H}$:

$$\pi_{0,H} = \frac{1}{1+\frac{\eta_H}{\eta_L}} - \frac{\lambda}{\mu_H}. \tag{4.18}$$

After finding $\pi_{0,H}$ and replacing it into Eq.(4.12), and the later into Eq.(4.13) we find $G_L(z)$ and $G_H(z)$ in closed form.

The next step is to find the average number of files in the system. It is the sum of the derivatives of partial PGFs at point $z = 1$, i.e.

$$E[N] = E[N_L] + E[N_H] = G_L^{'}(1) + G_H^{'}(1). \tag{4.19}$$

Differentiating Eq.(4.12) with respect to z we have

$$G_L^{'}(z) = \frac{\mu_H\pi_{0,H}F(z) - \mu_H\pi_{0,H}(z-1)F^{'}(z)}{F^2(z)}. \tag{4.20}$$

In Eq.(4.20), $F(z) = A(z)B(z) - \eta_L z$, where $A(z) = \frac{\lambda z(1-z)+\mu_H(z-1)+\eta_H z}{\eta_H}$, and $B(z) = \lambda(1-z)+\eta_L$.

It can easily be proven that Eq.(4.20) is of the form $\frac{0}{0}$ at $z = 1$. After applying L'Hôpital's rule twice, we get

$$G_L^{'}(z) = \frac{-\mu_H\pi_{0,H}F^{''}(z) + \mu_H\pi_{0,H}F^{'''}(z)(1-z)}{2F^{'}(z)^2 + 2F(z)F^{''}(z)}. \tag{4.21}$$

Based on Eq.(4.21) we can now get

$$E[N_L] = \lim_{z\to 1} G_L^{'}(z) = \frac{-\mu_H\pi_{0,H}F^{''}(1)}{2F^{'}(1)^2}. \tag{4.22}$$

After some algebra, we obtain

$$E[N_L] = \frac{\lambda\mu_H\pi_{0,H}\left(\eta_L + \mu_H + \eta_H - \lambda\right)}{\eta_H\left[\frac{1}{\eta_H}(\mu_H + \eta_H - \lambda)\eta_L - \lambda - \eta_L\right]^2}. \tag{4.23}$$

The next step is to find $E[N_H]$ (the average number of files while being in a high state). For that purpose, Eq.(4.13) is differentiated, giving

$$G_H^{'}(z) = \frac{1}{\eta_H}\left\{-\lambda G_L(z) + \left[\lambda(1-z) + \eta_L\right]G_L^{'}(z)\right\}. \tag{4.24}$$

Since $E[N_H] = \lim_{z\to 1} G_H^{'}(z)$, substituting $z = 1$ into Eq.(4.24) results in

$$E[N_H] = \frac{1}{\eta_H}\left[-\lambda G_L(1) + \eta_L G_L^{'}(1)\right]. \tag{4.25}$$

In Eq.(4.25), $E[N_L] = G_L^{'}(1)$, and $G_L(1) = \frac{\eta_H E[T_s]}{1+\eta_H E[T_s]}$. So, Eq.(4.25) reduces to

$$E[N_H] = -\frac{\lambda E[T_s]}{1+\eta_H E[T_s]} + \frac{1}{\eta_H E[T_s]}E[N_L]. \tag{4.26}$$

After some algebra, we can find that the average number of files in the system is $E[N] = E[N_L] + E[N_H]$.

Finally, using Little's law $E[N] = \lambda E[T]$ [20], we obtain the average file delay in the interweave access as in Eq.(4.3). □

Low variability scanning time. In the previous section we have derived the average file delay for exponentially distributed scanning times. However, as explained in Section 4.2.2, there exist some cases when the scanning time can have less variability than the exponential distribution. To capture this low variability an Erlang k-stage distribution is assumed. Our system can still be modeled with a 2D Markov chain, as depicted in Fig. 4.6. However, a transition from a low state (i.e. scanning) to a high state (i.e. finding and using a new available channel) would now have to go through an additional $k-1$ intermediate states (vertically), as opposed to going directly to the high state as in the exponential case (see Fig. 4.5). The transition rate between these states is $k\eta_L$. Since there are k stages, the average scanning time is $E[T_s] = k\frac{1}{k\eta_L} = \frac{1}{\eta_L}$. It is not possible to make a transition backwards while in the scanning mode (no transmission).

In general, it is very difficult to solve these kind of Markov chains analytically, and one needs to use numerical, matrix-analytic methods [49]. However, numerical methods do not provide any insight on the nature of the solution and its dependency on certain parameters.

Interestingly, due to the particular structure of the MC at hand, we are nevertheless able to derive a closed form analytical expression. Although there are more than two states in the "vertical" direction, we can still write down the balance equations and follow our approach to solve a system of $k+1$ equations in the partial probability generating functions. This is the main difference with the scenario for scanning times that are exponential.

The following theorem gives the expected delay in this scenario.

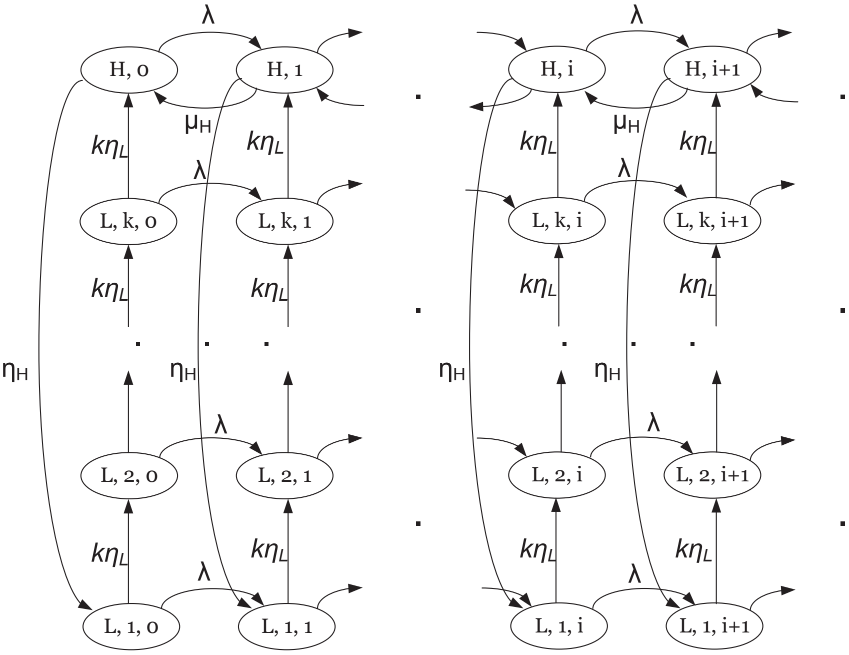

Figure 4.6: The 2D Markov chain for Erlang-distributed scanning time.

Theorem 6. *The average file delay in the interweave access with Erlang distributed scanning time is given by*

$$E[T_{erl}] = \frac{\eta_H \left[\eta_H + \frac{(k+1)}{2k}\mu_H\right](E[T_s])^2 + 2\eta_H E[T_s] + 1}{(1+\eta_H E[T_s])(\mu_H - \lambda - \lambda\eta_H E[T_s])}. \tag{4.27}$$

Proof. The balance equations for this chain are

$$\begin{aligned}
\pi_{L,1,0}(\lambda + k\eta_L) &= \eta_H \pi_{H,0} && (4.28)\\
\pi_{L,1,i}(\lambda + k\eta_L) &= \eta_H \pi_{H,i} + \lambda\pi_{L,1,i-1},\ i \geq 1 && (4.29)\\
\pi_{L,j,0}(\lambda + k\eta_L) &= k\eta_L \pi_{L,j-1,0},\ 2 \leq j \leq k && (4.30)\\
\pi_{L,j,i}(\lambda + k\eta_L) &= k\eta_L \pi_{L,j-1,i} + \lambda\pi_{L,j,i-1},\ i \geq 1, 2 \leq j \leq k && (4.31)\\
\pi_{H,0}(\lambda + \eta_H) &= k\eta_L \pi_{L,k,0} + \mu_H \pi_{H,1} && (4.32)\\
\pi_{H,i}(\lambda + \eta_H + \mu_H) &= k\eta_L \pi_{L,k,i} + \lambda\pi_{H,i-1} + \mu_H \pi_{H,i+1},\ i \geq 1 && (4.33)
\end{aligned}$$

We define the partial probability generating functions for the scanning phases as

$$G_{L,j}(z) = \sum_{i=1}^{\infty} \pi_{j,i} z^i,\ j = 2, \ldots, k, |z| \leq 1. \tag{4.34}$$

and for the transmission phase

$$G_H(z) = \sum_{i=1}^{\infty} \pi_{H,i} z^i,\ |z| \leq 1. \tag{4.35}$$

Further, we continue with multiplying Eq.(4.29), Eq.(4.31), and Eq.(4.33) with z^i and adding each of them to Eq.(4.28), Eq.(4.30) and Eq.(4.32), respectively. After that, we obtain the following system of equations with partial probability generating functions as unknowns

$$[\lambda(1-z)+k\eta_L]\,G_{L,1}(z)=\eta_H G_H(z) \tag{4.36}$$

$$[\lambda(1-z)+k\eta_L]\,G_{L,j}(z)=\eta_L G_{L,j-1}(z) \tag{4.37}$$

$$[\lambda z(1-z)+\mu_H(z-1)+\eta_H z]\,G_H(z)=k\eta_L z G_{L,k}(z)+\mu_H\pi_{H,0}(z-1) \tag{4.38}$$

Eq.(4.36) can be expressed as

$$G_{L,1}(z)=\frac{\eta_H}{[\lambda(1-z)+k\eta_L]}G_H(z). \tag{4.39}$$

From Eq.(4.37), for the states $j=2,\ldots,k$ there are recursions involved. The partial PGF for the jth states can be written as

$$G_{L,j}(z)=\frac{k\eta_L}{[\lambda(1-z)+k\eta_L]}G_{L,j-1}(z),\ j\geq 2, \tag{4.40}$$

and after using this recursion we get

$$G_{L,j}(z)=\frac{(k\eta_L)^{j-1}\eta_H}{[\lambda(1-z)+k\eta_L]^j}G_H(z),\ j\geq 2. \tag{4.41}$$

Solving the system of equations Eq.(4.36)-Eq.(4.38), for $G_H(z)$ we obtain

$$G_H(z)=\frac{\mu_H\pi_{H,0}(z-1)}{\lambda z(1-z)+\mu_H(z-1)+\eta_H z-\frac{\eta_H z}{\left[1+\frac{\lambda}{k\eta_L}(1-z)\right]^k}}. \tag{4.42}$$

Replacing Eq.(4.42) into Eq.(4.39) and Eq.(4.41), we have the solutions for all the PGFs. However, there is an unknown component $\pi_{H,0}$ appearing in those expressions. We can find it by taking the vertical cut between the states corresponding to i and $i+1$ files

$$\lambda\left(\sum_{j=1}^{k}\pi_{L,j,i}+\pi_{H,i}\right)=\mu_H\pi_{H,i+1}, \tag{4.43}$$

which after some rearrangements yields to

$$\pi_{H,0}=G_H(1)-\frac{\lambda}{\mu_H}. \tag{4.44}$$

The stationary probability of finding the system in the idle PU state, $G_H(1)$, can be found as follows. Eq.(4.39), for $z=1$ reduces to $G_{L,1}=\frac{\eta_H}{k\eta_L}G_H(1)$. Similarly, for the other scanning phases the following result can be derived

$$G_{L,j}(1)=G_{L,j-1}(1),\ j\geq 2. \tag{4.45}$$

Obviously, it holds that

$$G_{L,1}(1) + \ldots + G_{L,k}(1) + G_H(1) = 1, \tag{4.46}$$

which after solving gives

$$G_H(1) = \frac{1}{1 + \frac{\eta_H}{\eta_L}}. \tag{4.47}$$

The term $\pi_{H,0}$ is obtained after replacing Eq.(4.47) into Eq.(4.44) as

$$\pi_{H,0} = \frac{1}{1 + \eta_H \eta_L} - \frac{\lambda}{\mu_H}. \tag{4.48}$$

After finding all the partial PGFs, we move further in finding the average file delay. For that purpose, first we need to find the average number of files in the system, which is given as

$$E[N] = E[N_1] + \ldots + E[N_k] + E[N_H]. \tag{4.49}$$

We find $E[N_H] = G_H^{'}(1)$ as follows. Differentiating Eq.(4.42) with respect to z gives

$$G_H^{'}(z) = \frac{\mu_H \pi_{H,0} F(z) - \mu_H \pi_{H,0}(z-1)F^{'}(z)}{F(z)^2}, \tag{4.50}$$

with $F(z) = \lambda z(1-z) + \mu_H(z-1) + \eta_H z - \frac{\eta_H z}{[1+\lambda k \eta_L(1-z)]^k}$. It can be easily proven that $F(z) = 0$ making both the numerator and denominator equal to 0. Hence, we need to apply L'Hôpital's rule twice. After that, we obtain

$$E[N_H] = \lim_{z \to 1} G_H^{'}(z) = -\frac{\mu_H \pi_{H,0} F^{''}(1)}{2F^{'}(1)^2}, \tag{4.51}$$

where $F^{'}(1) = \mu_H - \lambda - \lambda\frac{\eta_H}{\eta_L}$, and $F^{''}(1) = \lambda\left[2 + \frac{\eta_H}{\eta_L}\left(\frac{\lambda}{\eta_L} + \frac{\lambda}{k\eta_L} + 2\right)\right]$. Next, we need to determine the PGFs related to the scanning part. As shown earlier, the following relation holds

$$G_{L,j}(z) = \frac{(k\eta_L)^{j-1}\eta_H}{[\lambda(1-z) + k\eta_L]^j} G_H(z),\ j \geq 1. \tag{4.52}$$

Differentiating Eq.(4.52) with respect to z yields

$$G_{L,j}^{'}(z) = (k\eta_L)^{j-1}\eta_H \cdot \frac{[\lambda(1-z) + k\eta_L]^j G_H^{'}(z) + j\lambda G_H(z)[\lambda(1-z) + k\eta_L]^{j-1}}{[\lambda(1-z) + k\eta_L]^{2j}}. \tag{4.53}$$

Since $E[N_{L,j}] = G_{L,j}^{'}(1)$, for $j = 1, \ldots, k$, after some algebra we obtain

$$E[N_{L,j}] = \frac{\eta_H}{\eta_L} E[N_H] + \frac{\lambda \eta_H}{(k\eta_L)^2} \cdot j \cdot G_H(1),\ j \geq 1. \tag{4.54}$$

Replacing Eq.(4.51) and Eq.(4.54) into Eq.(4.49), we find the average number of files in the system as

$$E[N] = \left(1 + \frac{\eta_H}{\eta_L}\right) E[N_H] + \lambda \eta_H \frac{k+1}{2k}\frac{1}{\eta_L^2} G_H(1). \tag{4.55}$$

Replacing Eq.(4.47) and Eq.(4.51) into Eq.(4.55), we can find the average number of files in the system. Finally, using Little's law $E[T] = \lambda E[N]$ and expressing the scanning time as $E[T_s] = \frac{1}{k\eta_L}$, we obtain Eq.(4.27). □

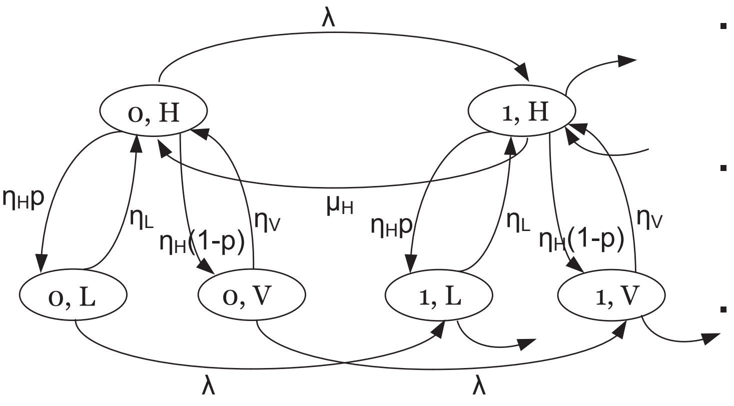

Figure 4.7: The 2D Markov chain for hyperexponentially distributed scanning time.

By carefully comparing Eq.(4.3) and Eq.(4.27) one can notice that the average delay for exponential scanning time is always higher compared to the delay induced in the case with Erlang scanning time, since $\frac{k+1}{2k} < 1, \forall k > 1$. This is in accordance with queueing system experience, where higher variability usually reduces performance.

High variability scanning time. Finally, we proceed with the case of high variability scanning time, which is modeled by an hyperexponential distribution with two branches that will be mapped into two separate states (denoted with the index L and V). The 2D Markov chain for this model is shown in Fig. 4.7. While being in the scanning phase, the SU can be either in one of the (i, L) states (short time of finding an available channel), or in one of the (i, V) states (long time until an available channel is found). The average scanning time in this setup is $E[T_s] = \frac{p}{\eta_L} + \frac{1-p}{\eta_V}$. In order to maintain the same $E[T_s]$ as before, but with much higher variability, we choose a very low value for $1-p$ (e.g. lower than 0.05) and $\eta_V << \eta_L$.

Once more, the structure of this chain allows us to avoid numerical, matrix-analytic methods, and instead apply the methodology of PGFs to derive a closed form expression, given in the following theorem.

Theorem 7. *The average file delay in interweave access with hyperexponential scanning time is given by*

$$E[T_{hyp}] = \frac{(\eta_H E[T_s])^2 + \eta_H \mu_H \left(\frac{p}{\eta_L^2} + \frac{1-p}{\eta_V^2}\right) + 2\eta_H E[T_s] + 1}{(1 + \eta_H E[T_s])(\mu_H - \lambda - \lambda \eta_H E[T_s])}. \tag{4.56}$$

Proof. As a first step, we need to write the balance equations for this chain:

$$\pi_{0,L}(\lambda + \eta_L) = p\eta_H \pi_{0,H} \tag{4.57}$$

$$\pi_{i,L}(\lambda + \eta_L) = \lambda \pi_{i-1,L} + p\eta_H \pi_{i,H},\ i \geq 1 \tag{4.58}$$

$$\pi_{0,V}(\lambda + \eta_V) = (1-p)\eta_H \pi_{0,H} \tag{4.59}$$

$$\pi_{i,V}(\lambda + \eta_V) = \lambda \pi_{i-1,V} + (1-p)\eta_H \pi_{i,H},\ i \geq 1 \tag{4.60}$$

$$\pi_{0,H}(\eta_H + \lambda) = \eta_L \pi_{0,L} + \eta_V \pi_{0,V} + \mu_H \pi_{1,H} \tag{4.61}$$

$$\pi_{i,H}(\lambda + \mu_H + \eta_H) = \lambda \pi_{i-1,H} + \eta_L \pi_{i,L} + \eta_V \pi_{i,V} + \mu_H \pi_{i+1,H},\ i \geq 1. \tag{4.62}$$

The probability generating function for the V-state is defined as

$$G_V(z) = \sum_{i=0}^{\infty} \pi_{i,V} z^i, |z| \leq 1. \tag{4.63}$$

As in the previous subsections, we multiply Eq.(4.58), Eq.(4.60), and Eq.(4.62) with z^i and add them to Eq.(4.57), Eq.(4.59) and Eq.(4.62), respectively. After some modifications we get the following system of equations with partial PGFs as unknowns

$$[\lambda(1-z) + \eta_L]\, G_L(z) = p\eta_H G_H(z), \tag{4.64}$$

$$[\lambda(1-z) + \eta_V]\, G_V(z) = (1-p)\eta_H G_H(z), \tag{4.65}$$

$$[\lambda z(1-z) + \eta_H z + \mu_H(z-1)]\, G_H(z) = \eta_L z G_L(z) + \eta_V z G_V(z) + \mu_H \pi_{0,H}(z-1). \tag{4.66}$$

Expressing $G_L(z)$ and $G_V(z)$ through $G_H(z)$ in Eq.(4.64) and Eq.(4.65), and replacing them afterward into Eq.(4.66) results in

$$G_H(z) = \frac{\mu_H \pi_{0,H}(z-1)}{\lambda z(1-z) + \eta_H z + \mu_H(z-1) - \frac{\eta_L z p \eta_H}{\lambda(1-z)+\eta_L} - \frac{\eta_V z(1-p)\eta_H}{\lambda(1-z)+\eta_V}} \tag{4.67}$$

Obtaining the expression for $G_H(z)$, we replace it into Eq.(4.64) and Eq.(4.65) to get the expressions for $G_L(z)$ and $G_V(z)$, respectively. The only unknown out of these equations is the zero probability for the high state, $\pi_{0,H}$. It is derived the same way as before (using a vertical cut between neighboring triplets of states), and is given as

$$\pi_{0,H} = G_H(1) - \frac{\lambda}{\mu_H}. \tag{4.68}$$

Similar approach, as before, is used in determining $G_H(1)$. Replacing $z = 1$ into Eq.(4.65), we have

$$G_V(1) = (1-p)\frac{\eta_H}{\eta_V} G_H(1). \tag{4.69}$$

In the same way, we replace $z = 1$ into Eq.(4.64), leading to

$$G_L(1) = p\frac{\eta_H}{\eta_L} G_H(1). \tag{4.70}$$

We next substitute Eq.(4.69) and Eq.(4.70) into $G_H(1) + G_L(1) + G_V(1) = 1$. We get

$$G_H(1) = \left[\frac{p\eta_H}{\eta_L} + \frac{(1-p)\eta_H}{\eta_V} + 1\right]^{-1}. \tag{4.71}$$

To get $E[N_H] = G_H^{'}(1)$, we need to differentiate Eq.(4.67). After using twice L'Hôpital's rule, we obtain

$$E[N_H] = \frac{-\mu_H \pi_{0,H} F^{''}(1)}{2F^{'}(1)^2}, \tag{4.72}$$

where $F^{'}(1) = F_1^{'}(1) + F_2^{'}(1) + F_3^{'}(1)$, and $F^{''}(1) = -2\lambda - \frac{2\lambda p \eta_H(\lambda+\eta_L)}{\eta_L^2} - \frac{2\lambda(1-p)\eta_H(\lambda+\eta_V)}{\eta_V^2}$.

The "low component" of the average number of files is $E[N_L] = G_L^{'}(1)$. The first derivative of $G_L(z)$ is found from Eq.(4.64), and is

$$G_L^{'}(z) = \frac{p\eta_H G_H^{'}(1)\left[\lambda(1-z)+\eta_L\right] + p\eta_H\lambda G_H(z)}{\left[\lambda(1-z)+\eta_L\right]^2}. \tag{4.73}$$

So, for $E[N_L]$ we have

$$E[N_L] = \frac{p\eta_H\eta_L E[N_H] + p\eta_H\lambda G_H(1)}{\eta_L^2}, \tag{4.74}$$

with $E[N_H]$ given by Eq.(4.72).

Pursuing the same process for the V-states we have as follows. First, from

$$G_V(z) = \frac{(1-p)\eta_H}{\lambda(1-z)+\eta_V} \cdot G_H(z), \tag{4.75}$$

we get the first derivative as

$$G_V^{'}(z) = \frac{(1-p)\eta_H G_H^{'}(z)\left[\lambda(1-z)+\eta_V\right] + \lambda(1-p)\eta_H G_H(z)}{\left[\lambda(1-z)+\eta_V\right]^2}, \tag{4.76}$$

and at point $z = 1$, the average number of files for the V-state is

$$E[N_V] = G_V^{'}(1) = \frac{(1-p)\eta_H E[N_H] + \lambda(1-p)\eta_H G_H(1)}{\eta_V^2}. \tag{4.77}$$

The average number of files in the system is $E[N] = E[N_H] + E[N_L] + E[N_V]$. Applying the Little's law $E[N] = \lambda E[T]$, we obtain the average file delay as in Eq.(4.56). □

Let's consider the term in brackets in the numerator of Eq.(4.56). It can be rewritten as

$$\frac{p}{\eta_L^2} + \frac{1-p}{\eta_V^2} = \frac{1}{\eta_V^2} + p\left(\frac{1}{\eta_L^2} - \frac{1}{\eta_V^2}\right) = \frac{1}{\eta_V^2} + \left(\frac{1}{\eta_L} + \frac{1}{\eta_V}\right)\cdot p\left(\frac{1}{\eta_L} - \frac{1}{\eta_V}\right). \tag{4.78}$$

Since the average scanning time is given as

$$E[T_s] = \frac{p}{\eta_L} + \frac{1-p}{\eta_V} = \frac{1}{\eta_V} + p\left(\frac{1}{\eta_L} - \frac{1}{\eta_V}\right).$$

From the last equation, we have

$$p\left(\frac{1}{\eta_L} - \frac{1}{\eta_V}\right) = E[T_s] - \frac{1}{\eta_V}.$$

Replacing the last relation into Eq.(4.78), we obtain

$$\begin{aligned}\frac{p}{\eta_L^2} + \frac{1-p}{\eta_V^2} &= \frac{1}{\eta_V^2} + \left(\frac{1}{\eta_L} + \frac{1}{\eta_V}\right)\left(E[T_s] - \frac{1}{\eta_V}\right) > \frac{1}{\eta_V^2} + E[T_s]\left(E[T_s] - \frac{1}{\eta_V}\right) \\ &= \left(E[T_s]\right)^2 + \frac{1}{\eta_V}\left(\frac{1}{\eta_V} - E[T_s]\right) > \left(E[T_s]\right)^2. \end{aligned} \tag{4.79}$$

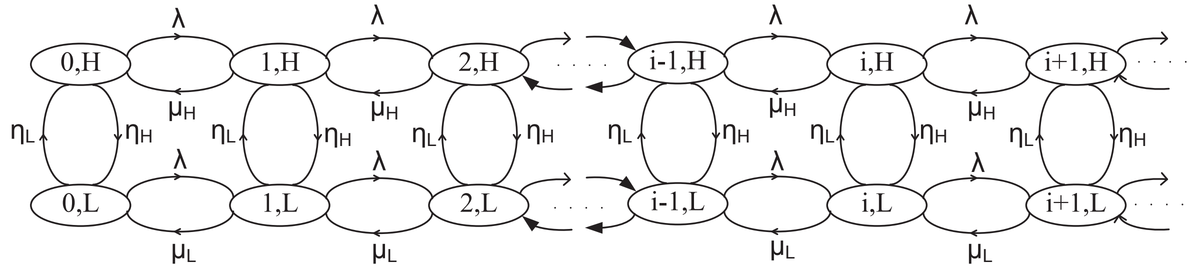

Figure 4.8: The 2D Markov chain for the underlay model.

Since $p < 1$, it holds that $\frac{1}{\eta_L} + \frac{1}{\eta_V} > \frac{p}{\eta_L} + \frac{1-p}{\eta_V} = E[T_s]$. Since the average scanning time is much lower than the average occasional long periods, it holds also that $\frac{1}{\eta_V} > E[T_s]$. Hence, we arrive at Eq.(4.79).

The corresponding term in the numerator of Eq.(4.3) is $(E[T_s])^2$, and the other terms in both the numerator and denominator are identical to the terms of Eq.(4.56). This way we have proven that $E[T_{hyp}] > E[T_{exp}]$, which is expected from queueing systems. So, to sum it up we have shown that the following relation holds $E[T_{hyp}] > E[T_{exp}] > E[T_{erl}]$. This means that the variability of the scanning time plays a crucial role in the average delay in the interweave access mode.

4.3.2 Delay analysis for underlay access

As we have already explained in Section 4.2.1, in the underlay CRN the SU can transmit all the time (both when in high and low states). We can again model the system with a 2D Markov chain, as shown in Fig. 4.8. Note the difference with Fig. 4.5. While in Fig. 4.5 there is no transition backwards in the low states, in the underlay CRN these transitions exist with rate μ_L.

We should mention that $\pi_{0,H}$ ($\pi_{0,L}$) denote, as before, the stationary probability of finding the SU with no files to transmit while being in a high (low) period.

Theorem 8. *The average file delay in the underlay access mode is given by*

$$E[T_u] = \frac{\eta_H + \eta_L + \mu_H(1-\pi_{0,H}) + \mu_L(1-\pi_{0,L}) - \lambda + \frac{\mu_L\mu_H}{\lambda}(\pi_{0,L} + \pi_{0,H} - 1)}{\mu_H\eta_L + \mu_L\eta_H - \lambda(\eta_H + \eta_L)}. \tag{4.80}$$

Proof. Writing the balance equations for this chain gives

$$\begin{aligned}
\pi_{0,L}(\lambda + \eta_L) &= \pi_{1,L}\mu_L + \pi_{0,H}\eta_H & (4.81)\\
\pi_{0,H}(\lambda + \eta_H) &= \pi_{1,H}\mu_H + \pi_{0,L}\eta_L & (4.82)\\
\pi_{i,L}(\lambda + \eta_L + \mu_L) &= \pi_{i-1,L}\lambda + \pi_{i+1,L}\mu_L + \pi_{i,H}\eta_H, (i > 0) & (4.83)\\
\pi_{i,H}(\lambda + \eta_H + \mu_H) &= \pi_{i-1,H}\lambda + \pi_{i+1,H}\mu_H + \pi_{i,L}\eta_L, (i > 0). & (4.84)
\end{aligned}$$

Similarly as before, we define the probability generating functions for both the low and high states as

$$G_L(z) = \sum_{i=0}^{\infty} \pi_{i,L} z^i, \text{and } G_H(z) = \sum_{i=0}^{\infty} \pi_{i,H} z^i, |z| \leq 1.$$

After multiplying Eq.(4.83) and Eq.(4.84) by z^i, adding them to Eq.(4.81) and Eq.(4.82), respectively, and rearranging (in the same direction as we did for interweave CRN), we obtain

$$(\lambda + \eta_L + \mu_L)G_L(z) = \lambda z G_L(z) + \eta_H G_H(z) + \frac{\mu_L}{z}\left(G_L(z) - \pi_{0,L}\right) + \pi_{0,L}\mu_L, \tag{4.85}$$

and

$$(\lambda + \eta_H + \mu_H)G_H(z) = \lambda z G_H(z) + \eta_L G_L(z) + \frac{\mu_H}{z}\left(G_H(z) - \pi_{0,H}\right) + \pi_{0,H}\mu_H. \tag{4.86}$$

Solving the system of equations Eq.(4.85)-(4.86), gives

$$f(z)G_L(z) = \pi_{0,H}\eta_H\mu_H z + \pi_{0,L}\mu_L\left[\eta_H z + (\lambda - z\mu_H)(1-z)\right], \tag{4.87}$$

where

$$f(z) = \lambda^2 z^3 - \lambda(\eta_L + \eta_H + \lambda + \mu_H + \mu_L)z^2 + (\eta_L\mu_H + \eta_H\mu_L + \mu_L\mu_H + \lambda\mu_H + \lambda\mu_L)z - \mu_L\mu_H. \tag{4.88}$$

It can be proven that the polynomial in Eq.(4.88) has only one root in the open interval $(0, 1)$ [52]. This root is denoted as z_0. Setting $z = z_0$ into Eq.(4.87), and after some algebra we get $\pi_{0,L}$ and $\pi_{0,H}$, as

$$\pi_{0,L} = \frac{\eta_H\left(\frac{\eta_H\mu_L + \eta_L\mu_H}{\eta_H + \eta_L} - \lambda\right)z_0}{\mu_L(1-z_0)(\mu_H - \lambda z_0)}, \tag{4.89}$$

$$\pi_{0,H} = \frac{\eta_L\left(\frac{\eta_H\mu_L + \eta_L\mu_H}{\eta_H + \eta_L} - \lambda\right)z_0}{\mu_H(1-z_0)(\mu_L - \lambda z_0)}. \tag{4.90}$$

Finally, using Little's law $E[N] = \lambda E[T]$ [20], we obtain Eq.(4.80). □

4.3.3 Analytical comparison of delays in underlay and interweave mode

Having derived the formulas for the mean delay in underlay and interweave CRNs in Sections 4.3.1 and 4.3.2, we are able to compare the delays incurred in each of them. As could have been noticed, the delay depends on the statistics of the PU activity, data rate, traffic intensity, and scanning time. In a first scenario, we assume that the SU has to decide at the beginning which of the access modes to use: underlay (i.e. always stay on same channel and transmit with the permitted power), or interweave (i.e. become silent whenever a PU arrives on the channel and scan for a new one). We will refer to this simply as "the static policy". While not a real policy per se (in practice, a node will always be able to scan and switch channels eventually), it allows us to gain some insights as to the relative parameters affecting the performance in each case. In Section 4.3.4, we will consider a more realistic, *dynamic policy.*

In general, for interweave access to outperform underlay access, the expected scanning time $E[T_s]$ should be short enough to ensure that the opportunity cost of not transmitting/receiving any data for some time (which is allowed in underlay) is amortized by the quick discovery of a new white space. In Table 4.2, we provide analytical expression for the maximum $E[T_s]$ values for which interweave access has lower delays. As can be seen from Table 4.2, there is a complex dependency on the various system parameters. What is more, this "boundary" point further depends on the variability of the scanning time.

Table 4.2: The analytical comparison of underlay and interweave modes.

T_s	Condition	Notation
Erlang	$E[T_s] < \frac{-B_2+\sqrt{B_2^2-4B_1B_3}}{2B_1}$	$B_1 = 2\eta_H^2 k + \eta_H(k+1)\mu_H + 2kE[T_u]\lambda\eta_H^2$ $B_2 = \eta_H(4k - 2k\mu_H E[T_u] + 4kE[T_u]\lambda)$ $B_3 = 2k(1 - (\mu_H - \lambda)E[T_u])$
Exponential	$E[T_s] < \frac{-A_2+\sqrt{A_2^2-4A_1A_3}}{2A_1}$	$A_1 = \eta_H(\mu_H + \eta_H) + \lambda\eta_H^2 E[T_u] > 0$ $A_2 = 2\eta_H - \eta_H(\mu_H - 2\lambda)E[T_u]$ $A_3 = 1 - (\mu_H - \lambda)E[T_u]$
Hyperexponential	$E[T_s] < \frac{-C_2+\sqrt{C_2^2-4C_1C_3}}{2C_1}$	$C_1 = \eta_L\eta_V\eta_H^2\,(1 + \lambda E[T_u])$ $C_2 = \eta_H\,[\mu_H(\eta_V + \eta_L) + \eta_L\eta_V\,(2 + 2\lambda E[T_u] - \mu_H E[T_u])]$ $C_3 = \eta_L\eta_V - \eta_H\mu_H - \eta_L\eta_V\,(\mu_H - \lambda)\,E[T_u]$

From Table 4.2 we can observe the following relations: $B_2 = 2kA_2$, $B_3 = 2kA_3$, and $B_1 < 2kA_1$. So, for the crossing point of the Erlang distributed scanning time, we have

$$E[T_{s,erl}] < \frac{-B_2 + \sqrt{B_2^2 - 4B_1B_3}}{2B_1} = \frac{-2kA_2 + \sqrt{4k^2A_2^2 - 4B_1 \cdot 2kA_3}}{2B_1} >$$

$$> \frac{-2kA_2 + \sqrt{4k^2A_2^2 - 8kA_1 \cdot 2kA_3}}{2B_1} > \frac{-2kA_2 + \sqrt{4k^2A_2^2 - 16k^2A_1A_3}}{4kA_1} =$$

$$= \frac{-A_2 + \sqrt{A_2^2 - 4A_1A_3}}{2A_1} = E[T_{s,exp}]. \tag{4.91}$$

From Eq.(4.91) we can observe that for interweave to outperform underlay access, in the case of Erlang scanning time, a smaller scanning time is needed. Similar conclusions can be drawn by comparing the parameters of hyperexponential distribution with the two previous ones. We can conclude that the variability of the scanning time has an important impact on the boundary scanning time. The higher the variability of the scanning time is, the lower scanning time is needed for the interweave mode to outperform the underlay access.

4.3.4 The delay minimization policy

In the previous section, we have compared underlay and interweave access, in a "static" context, where the decision between the two is made once, at the beginning. In practice, a node with a cognitive radio will normally be able to choose to stay at the current channel and transmit at low(er) power, or scan for a new white space *at any time.* Such a hybrid policy might lead to a further improvement in performance, if designed properly. We next define such a hybrid policy, identify the conditions under which it offers gains, and derive an optimal switching rule (from one mode to the other).

Definition 1. Delay minimization policy.

- *The SU will reside on the current channel if it is idle (no PU activity) and continue its activity there.*
- *If a PU is detected, the SU will continue transmitting with lower power, until a time t, called the "turning point".*

- *If the PU does not release the channel by time t, then the SU ceases transmission and starts scanning for a new idle channel.*
- *If the PU leaves the channel before t, then the SU resumes transmitting at higher power, and resets the turning point to t time units ahead.*

The above policy is generic. Our goal is to find an optimal value for t. Let us consider some cases, to better understand the tradeoffs involved. First, if the static interweave policy, as described in the previous section is better than the static underlay policy, then it is easy to see that the optimal value of t is 0: it is always better to start scanning immediately when a PU arrives. Hence, we are interested only in cases where the underlay is better *on average* (i.e. the respective condition in Table 4.2 *not* satisfied), but there are instances when the current channel remains busy for too long and it then becomes better to start scanning instead.

In the above context, assume that the PU activity (OFF) periods are exponentially distributed. Assume further that a PU arrived at the current channel and t units have already elapsed and the channel is still busy. Due to the memoryless property of the exponential distribution, the remaining time until the PU leaves is still the same, as in the beginning (when the PU just arrived), i.e. equal to $E[T_{OFF}]$. Hence, if at time 0 it was better to stay on the channel and transmit at lower rate rather than initiate scanning (which is what we assumed above), for any elapsed time t *it is still better to stay on the channel and not start scanning.* A similar conclusion can be drawn for PU activity periods with *increasing failure rate (IFR)*[3], i.e. lower variability than exponential. There, if at $t = 0$ one cannot gain by scanning (i.e. static underlay is better on average), then as t increases, the expected gain from staying in the underlay mode in fact increases.

Hence, we can conclude that a dynamic policy (i.e. an optimal value of t strictly larger than 0) may offer gains *only for PU activity periods with decreasing failure rate (DFR).* There, although at the beginning, when the PU arrives, it might be on average better to do underlay, as time elapses, the expected remaining PU busy time keeps increasing (above the average), until at some point it becomes profitable to stop and scan for a new empty channel. This allows the dynamic policy to outperform any of the static policies, as we show later, by essentially "pruning" the long OFF periods from the underlay mode.

In deriving the condition for dynamic delay policy, we make the assumption that files are not excessively large (average size Δ) and that they can be transmitted in 1-2 ON and OFF periods. We use this approximation to keep the derivation tractable and we will show that it does not affect much the result. With that approach, the total transmission delay is almost equal to the average service time. If by $p_{ON} = \frac{E_{T_{ON}}}{E_{T_{ON}}+E[T_{OFF}]}$ we denote the probability that an arriving file will find the system in an ON state, and by $p_{OFF} = \frac{E_{T_{OFF}}}{E_{T_{ON}}+E[T_{OFF}]}$ the probability of finding the system in an OFF state, then the average service time would be

$$E[S] = p_{ON}E[T_{X,ON}] + p_{OFF}E[T_{X,OFF}]. \tag{4.92}$$

In Eq.(4.92), $E[T_{X,ON}]$ ($E[T_{X,OFF}]$) denotes the average service time of a file that arrives during an ON (OFF) period. Each arriving file (in an ON period) will be partially transmitted during the time T^e_{ON}, which is the remaining (excess) duration of the ON period. Then at the beginning

[3] The distributions with increasing (decreasing) failure rate are those for which $\frac{f(x)}{1-F(x)}$ is an increasing (decreasing) function of x.

of the OFF period, the remaining file size is Δ_0. There are two options for the OFF periods (larger or smaller than t). With probability $P\left[T_{OFF} < t\right] = F_{OFF}(t)$ ($F(t)$ is the cumulative distribution function (CDF) of the OFF periods), the next OFF period will be short enough, so the SU will reside on the current channel and transmit with low power. In that case, the file will be completely transmitted in the next ON period. On the other hand, with probability $P\left[T_{OFF} > t\right] = \bar{F}_{OFF}(t)$, the next OFF period will be larger than the turning point. Note that $P\left[T_{OFF} > t\right] = 1 - F(t) = \bar{F}_{OFF}(t)$ is the complementary cumulative distribution function (CCDF) of the OFF periods. Then, the SU will initiate the scanning procedure and start looking for another available channel (during time T_s). In that period of time, there is no transmission. After finding an available channel, the SU will transmit with rate μ_H. So, the average service time of a file arriving during an ON period is

$$\begin{aligned} E[T_{X,ON}] = E[T_{ON}^{(e)}] &+ F_{OFF}(t)\left\{E[T_{OFF}|T_{OFF} < t] + \frac{\Delta_0 - \mu_L E[T_{OFF}|T_{OFF} < t]}{\mu_H}\right\} \\ &+ \bar{F}_{OFF}(t)\left\{t + E[T_s] + \frac{\Delta_0 - \mu_L t}{\mu_H}\right\} + \Omega_{ON}. \end{aligned} \tag{4.93}$$

Note that in Eq.(4.93), the term Ω_{ON} represents the contribution to the average delay, of the scenarios not included in the other term (the file transmitted during the first ON period, during the first OFF period, or eventually if it needs more ON and OFF periods for the complete transmission). However, as we have assumed that file sizes are exponentially distributed, hence with a low variance, and their sizes are such that in most cases it will suffice 1-2 ON-OFF periods to finish transmission, the term Ω_{ON} is negligible ($\Omega_{ON} -> 0$).

For files arriving in an OFF period again there are two possibilities. They could either arrive to an OFF period whose remaining time, $T_{OFF}^{(e)}$, is shorter than the turning point, or to a periods with excess time larger than t. The probability for the first scenario is $P\left[T_{OFF}^{(e)} < t\right] = F_{OFF}^{(e)}(t)$, while for the second one is $P\left[T_{OFF}^{(e)} > t\right] = \bar{F}_{OFF}^{(e)}(t)$. In the first case, the SU will remain on the current channel and continue its transmission with rate μ_L, and the file will be transmitted during the next ON period (when the rate is μ_H). In the second case, after time t, the SU will initiate the scanning process (no transmission) that will last T_s until an idle channel is found, and then will transmit with rate μ_H. The file will be transmitted during that ON period. So, the average service time of a file arriving during an OFF period can be expressed as

$$\begin{aligned} E[T_{X,OFF}] &= F_{OFF}^{(e)}(t)\left\{E\left[T_{OFF}^{(e)}|T_{OFF}^{(e)} < t\right] + \frac{\Delta - \mu_L E\left[T_{OFF}^{(e)}|T_{OFF}^{(e)} < t\right]}{\mu_H}\right\} \\ &+ \bar{F}_{OFF}^{(e)}(t)\left\{t + E[T_s] + \frac{\Delta - \mu_L t}{\mu_H}\right\} + \Omega_{OFF}. \end{aligned} \tag{4.94}$$

For the same reasons as for Ω_{ON} in Eq.(4.93), the term Ω_{OFF} in Eq.(4.94) can be neglected.

In Eq.(4.93), the following two terms are equivalent to

$$E\left[T_{ON}^{(e)}\right] = \frac{E[T_{ON}^2]}{2E[T_{ON}]}, \tag{4.95}$$

$$E[T_{OFF}|T_{OFF} < t] = \int_0^t \frac{x f_{OFF}(x)}{F_{OFF}(t)} dx. \tag{4.96}$$

Similarly, we have the following relations in Eq.(4.94)

$$F_{OFF}^{(e)}(t) = \int_0^t f_{OFF}^{(e)}(x)dx, \tag{4.97}$$

$$E\left[T_{OFF}^{(e)}|T_{OFF}^{(e)} < t\right] = \int_0^t \frac{xf_{OFF}^{(e)}(x)}{F_{OFF}^{(e)}(t)}dx. \tag{4.98}$$

In the two previous equations we have

$$f_{OFF}^{(e)}(x) = \frac{1 - F_{OFF}(x)}{E[T_{OFF}]}. \tag{4.99}$$

Replacing Eq.(4.93) and Eq.(4.94) into Eq.(4.92), we find the average service time. Since we are assuming that the average transmission delay is almost identical to the average service time, i.e. $E[T] \approx E[S]$. Further, to find the value of turning time that minimizes the delay, we have to find the solution of the following equation

$$\frac{\partial E[T]}{\partial t} = 0. \tag{4.100}$$

So, the solution of Eq.(4.100) is the optimal turning point. For illustration purposes, after solving Eq.(4.100), the following result provides an analytical expression for the case of Pareto distributed OFF periods (with parameters L, α), a popular distribution with decreasing failure rate, and for exponential ON periods.

Result 9. *The optimal turning point, t_{opt}, in a Cognitive Radio Network can be found as a solution of*

$$\begin{aligned} &\frac{\eta_H}{\eta_H + \eta_L} \cdot \frac{\mu_H - \mu_L}{\mu_H} \left(\left(1 - \frac{1}{\alpha}\right) t^{1+\alpha} + \frac{1}{\alpha} L^{\alpha-1} t^2 \right) \\ &+ \frac{\eta_H}{\eta_H + \frac{\mu_H}{\Delta}} \cdot \frac{\eta_L}{\eta_H + \eta_L} \left(\frac{\mu_H - \mu_L}{\mu_H} L^{\alpha} t - E[T_s]\alpha L^{\alpha} \right) = 0. \end{aligned} \tag{4.101}$$

The solution to Eq.(4.101) can be found numerically. Table 4.3 summarizes all the possible scenarios.

Table 4.3: Summary of the delay policies.

Scenario	Optimal dynamic policy
Static interweave better	Static interweave ($t_{opt} = 0$)
Static underlay better + IFR OFF	Static underlay ($t_{opt} = \infty$)
Static underlay better + Exp. OFF	Static underlay ($t_{opt} = \infty$)
Static underlay better + DFR OFF	Dynamic policy with $t_{opt} \in (0, \infty)$

It is interesting to note that, unlike the variability of OFF periods for underlay access, the variability of the scanning time distribution (the "OFF" periods in the interweave mode) does not affect the dynamic policy decisions. It only enters the picture for the comparison between the static underlay and interweave modes (Table 4.2).

It is also positive that in all but few cases the optimal policy is just the static one. This reduces the complexity of the algorithm significantly, as one needs to make this decision only once. In practice, some recalculation of this threshold (and thus the optimal mode) might still be necessary periodically, in order to account for qualitative (e.g. non-stationarity) changes in the behavior of the system and estimated statistics.

4.4 Throughput analysis of underlay and interweave access

Although (per message) delay is often a key performance measure, especially for interactive applications, a number of applications of interest (e.g. file download, peer-to-peer file exchange, one-way streaming, etc.) are often throughput-sensitive. Hence, maximizing the average data rate for these applications is a priority for a SU. Again, there is a trade off between staying on a channel and transmitting at low(er) power and pausing and looking for a new idle channel. When delay is the key metric, it might be better to simply transmit a message immediately, even if at lower rate, in order to not delay the message while scanning (making underlay preferable). However, if one cares more about throughput, an SU might gain more by scanning more aggressively and finding a new high throughput channel (even if some files have to queue in the meantime).

As in the case of delay, we will first compare the performance of the two "static" approaches, where the decision to use underlay or interweave type of access (exclusively), is made at the beginning, based on average statistics. We will then propose a dynamic policy, and consider the conditions under which it further improves the throughput.

4.4.1 Analytical comparison of throughput

Since channel use in both underlay and interweave modes can be modeled with an alternating (ON-OFF) renewal process, as explained earlier, we can use the *renewal-reward* approach to calculate the long-term throughput of each mode, and compare them. In both cases, a cycle consists of one ON and one OFF period, and the reward corresponds to the amount of data sent during one cycle. The respective rewards are shown in Fig. 4.9 (underlay) and Fig. 4.10 (interweave). Two differences can be seen there: (i) there is a non-zero reward during the OFF periods in underlay mode; (ii) the OFF periods in the interweave case depend on the scanning time statistics, rather than the PU activity time.

The following result gives the condition for which the interweave access outperforms the underlay mode.

Theorem 10. *The interweave spectrum access technique outperforms the underlay mode of access in CRN if the following condition is fulfilled*

$$\frac{E[T_s]}{E[T_{OFF}]} < \frac{1 - \frac{\mu_L}{\mu_H}}{1 + \frac{\mu_L}{\mu_H}\frac{E[T_{OFF}]}{E[T_{ON}]}}. \tag{4.102}$$

Proof. If we denote by $R(t)$ the reward earned by time t, the average (long-term) throughput is the mean reward rate $\frac{R(t)}{t}$, which from renewal-reward theory we know it to be [20]

$$\lim_{t\to\infty}\frac{E\left[R\left(t\right)\right]}{t} = \frac{E\left[R\right]}{E\left[T_{cycle}\right]}. \tag{4.103}$$

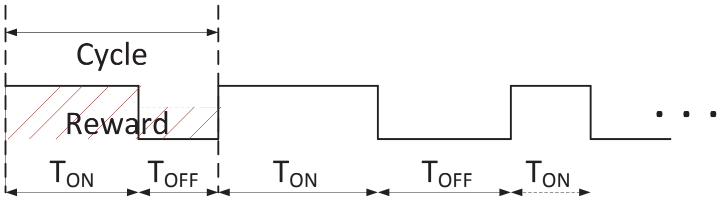

Figure 4.9: The renewals for the underlay CRN case.

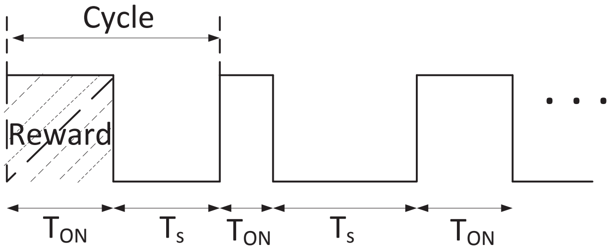

Figure 4.10: The renewals for the interweave CRN case.

We denote by $E[X]$ the average reward rate, that is equivalent to average throughput. The average reward is the amount of transmitted data per cycle, i.e. $E[R_u] = \mu_H E[T_{ON}] + \mu_L E[T_{OFF}]$, and the average cycle duration is $E[T_{u,cycle}] = E[T_{ON}] + E[T_{OFF}]$. Replacing these into Eq.(4.103), we obtain

$$E[X_u] = \frac{\mu_H E[T_{ON}] + \mu_L E[T_{OFF}]}{E[T_{ON}] + E[T_{OFF}]}. \tag{4.104}$$

Following a similar approach for interweave access, the average reward earned during a cycle is $E[R_i] = \mu_H E[T_{ON}]$, and the average cycle duration is $E[T_{i,cycle}] = E[T_{OFF}] + E[T_s]$. The average data rate is then

$$E[X_i] = \frac{E[R_i]}{E[T_{i,cycle}]} = \frac{\mu_H E[T_{ON}]}{E[T_{ON}] + E[T_s]}. \tag{4.105}$$

Comparing Eq.(4.104) and Eq.(4.105), we can find the condition under which it is better, on average, not to reside on the current channel. We solve the inequality $E[X_i] > E[X_u]$, which gives Eq.(4.102). □

According to Eq.(4.102), if, e.g. $E[T_{ON}] = E[T_{OFF}]$ and $\mu_H = 2\mu_L$, the interweave access will perform better for $\frac{E[T_s]}{E_{T_{OFF}}} < \frac{1}{3}$. On the other hand, if the difference between data rates is quite high, e.g. $\mu_H = 5\mu_L$, then Eq.(4.102) gives $\frac{E[T_s]}{E[T_{OFF}]} < \frac{2}{3}$. While in the first case there was a requirement that the scanning time should be smaller than 1/3 of the average OFF period, increasing the ratio between μ_H and μ_L allows higher $E[T_s]$, such that the interweave mode still

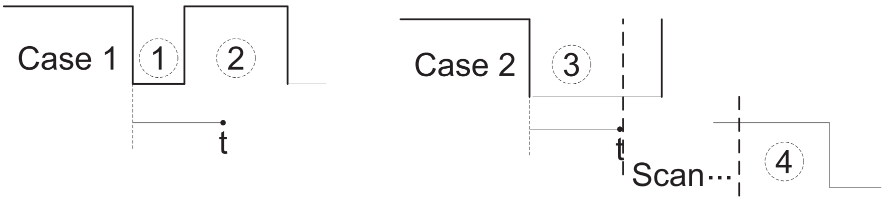

Figure 4.11: The throughput policy.

outperforms the underlay. Similarly, if the ratio $\frac{E[T_{ON}]}{E[T_{OFF}]}$ increases, the relative threshold also increases, as can be seen from Eq.(4.102).

4.4.2 Throughput maximization policy

In the case of throughput maximization as well (as in the case of delay minimization), if the "static" interweave mode of access is better *on average* (in the sense of the previous result), then it is easy to see that there is no need to ever employ the underlay mode. However, if the underlay mode is better *on average*, one might still improve performance by deciding to scan and switch to a new (idle) channel in some instances. In this section, we present such a dynamic policy, that tries to predict the potential gains "on-the-fly" and may choose to transmit at low power or scan at any time. Similarly as in the delay policy, we determine the optimal *turning point* t_{opt} when the SU should switch to the interweave mode, so that the maximum throughput is achieved. This policy is defined as follows:

Definition 2. Throughput maximization policy.

- *The SU will reside on the current channel if it is idle (no PU activity) and continue its activity there.*
- *If a PU is detected, the SU will continue transmitting with lower power, until a time t, called the "turning point".*
- *If the PU does not release the channel by time t, then the SU ceases transmission and starts scanning for a new idle channel.*
- *If the PU leaves the channel before t, then the SU resumes transmitting at higher power, and resets the turning point to t time units ahead.*

Hence, this is the same as the dynamic delay policy. The only difference is in the choice of the optimal turning point t_{opt}, which we will now choose to maximize the average data rate instead. We can again use the renewal-reward approach, with the difference that now the low periods can be either low periods as in the underlay scenario, or a combination of a low period and the scanning time.

As in the case of the dynamic delay policy, the key in the switching decision is in the uncertainty about how much longer the PU activity (i.e. the OFF cycle) will last. Even if we know its expected value (which is the basis for the static policy choice), the variability beyond this mean value plays an important role. It is easy to see that the same arguments, as in the case of delay minimization, can be made to show that any improvement beyond the static policies is only possible if underlay is better *on average*, and OFF periods have a decreasing failure rate. The following theorem thus derives the optimal turning point t_{opt} for Pareto OFF periods with parameters L, α.

Result 11. *The optimal value of the turning point that maximizes the average throughput for Pareto OFF, and exponential ON periods is the solution of*

$$\frac{\mu_H - \mu_L}{\eta_H} t^{\alpha} - \alpha E[T_s] \left(\frac{\mu_H}{\eta_H} - \mu_L L \frac{\alpha}{1-\alpha} \right) t^{(\alpha-1)} - \frac{\mu_L E[T_s] L^{\alpha}}{1-\alpha} = 0. \tag{4.106}$$

Proof. A cycle now consists of the following combination of ON and OFF periods (Fig. 4.11): (1) an OFF period shorter than the turning point $T_{OFF} < t$, and a *regular* ON period (Case 1), corresponding to the case when the SU remains on the current channel and transmits with lower power; (2) an OFF period that is longer than the turning point, and an ON period corresponding to the remaining time of the new channel being idle ($T_{ON}^{(e)}$) (Case 2). The OFF period is then the sum of the deterministic turning point t and the scanning time, T_s. Case 1 can happen with a probability $P[T_{OFF} < t] = F_{OFF}(t)$. Hence, the average amount of transmitted data per cycle and its duration are

$$E[R] = \left(\mu_L E[T_{OFF}|T_{OFF} < t] + \mu_H E[T_{ON}]\right) F_{OFF}(t) + \left(\mu_L t + \mu_H E[T_{ON}^{(e)}]\right) \bar{F}_{OFF}(t), \tag{4.107}$$

$$E[T_{cycle}] = \left(E[T_{OFF}|T_{OFF} < t] + E[T_{ON}]\right) F_{OFF}(t) + \left((t + E[T_s]) + E[T_{ON}^{(e)}]\right) \bar{F}_{OFF}(t). \tag{4.108}$$

The average duration of an OFF period given that it is shorter than t, $E[T_{OFF}|T_{OFF} < t]$, can be found easily due to our knowledge of the channel statistics. It is equal to $E[T_{OFF}|T_{OFF} < t] = \int_0^t \frac{x f_{OFF}(x)}{F_{OFF}(t)} dx$. The term $E\left[T_{ON}^{(e)}\right]$ is given by Eq.(4.95), and for an exponential ON period reduces to $E[T_{ON}]$.

For Pareto distributed OFF periods it holds that

$$f_{OFF}(x) = \frac{\alpha L^{\alpha}}{x^{\alpha+1}}, x \geq L.$$

The CDF of Pareto distribution is

$$F_{OFF}(x) = 1 - \left(\frac{L}{x}\right)^{\alpha}, x \geq \alpha.$$

After some algebra, we can also find the value of the term $E[T_{OFF}|T_{OFF} < t]$ for a Pareto distribution, as

$$E[T_{OFF}|T_{OFF} < t] = \int_0^t \frac{x f_{OFF}(x)}{F_{OFF}(t)} dx = \frac{\alpha L^{\alpha}}{(1-\alpha)\left(1 - \frac{L^{\alpha}}{t^{\alpha}}\right)} \left(t^{1-\alpha} - L^{1-\alpha}\right). \tag{4.109}$$

The average data rate is given by

$$E[X] = \frac{E[R]}{E[T_{cycle}]}. \tag{4.110}$$

The optimal value of the turning point t that maximizes the throughput can be found by solving the equation

$$\frac{\partial E[X]}{\partial t} = 0. \tag{4.111}$$

Solving Eq.(4.111), we find the optimal turning point that maximizes the throughput in our dynamic throughput policy. □

This policy, as we will see in the next section, can indeed improve the long-term throughput. As in the case of the dynamic delay policy, this gain comes with some extra complexity to maintain estimates of PU activity variability on every new channel, in addition to average (i.e. duty cycle) behavior, compared to the simpler "static" policy.

4.5 Simulation results

The first goal of this section is to validate the various analytical expressions we have derived, against simulated scenarios, including ones where one or more of the assumptions do not hold. We will also show the advantages offered by different optimization policies.

In the first scenario, we will assume that the average ON and OFF durations correspond to those measured in [18] and are equal to $E[T_{OFF}] = 10$s ($\eta_L = 0.1$), and $E[T_{ON}] = 5$s ($\eta_H = 0.2$). We will refer to this as the cellular network scenario. In the second scenario, we fit the average ON and OFF durations to the values observed in [19], with $E[T_{OFF}] = 9$s ($\eta_L = 0.11$), and $E[T_{ON}] = 4$s ($\eta_H = 0.25$). We will refer to this as the WiFi scenario. For both scenarios, unless otherwise stated, we assume exponential distributed periods. We consider other distributions later in Sections 4.5.1 and 4.5.2. The data rates for the cellular scenario are $\mu_L = 1.2$ Mbps and $\mu_H = 8$ Mbps. For the WiFi scenario the data rates are $\mu_H = 10$ Mbps, and $\mu_L = 2$ Mbps.[4] Finally, we assume that file arrivals at the SU are Poisson distributed with rate λ, and file sizes exponentially distributed with mean size 125KB.[5]

4.5.1 Validation of the delay models

Fig. 4.12 compares simulation results to our analytical model predictions for the average delay of SU files as the file arrival rate increases. The system parameters correspond to the cellular scenario. As can be seen, our theoretical results match with the results obtained from simulations. As is expected in queueing systems, the delay increases when the arrival rate (and thus the utilization of the system) increases. In the same plot the delay based on the statistics of the WiFi network [19] is shown as well. Again, there is a solid match between theory and simulations. The delay incurred in the cellular scenario is larger. This is because the data rates in the WiFi scenario are considered to be higher and the PU is less active there.

We move next to validating our analytical predictions for the interweave scenario. Fig. 4.13 and 4.14 show the theoretical vs. simulated results for the cellular and WiFi network scenarios for three types of scanning time distributions (exponential, 4-stage Erlang, hyperexponential),

[4] These values are taken to be of the same order of magnitude as the actual values encountered in practice [3]. Although these correspond to PUs, and the actual data rates for SUs depend on the distance of the SUs from the BTS or WiFi AP, channel width, channel conditions, modulation/coding, etc., we assume w.l.o.g. that the data rates of the SU in a WiFi network are higher than in a cellular network.

[5] This value is normalized for the arrival rates considered, to correspond approximately to the traffic intensities reported in [18] and [19]. We have also considered other values with similar conclusions drawn.

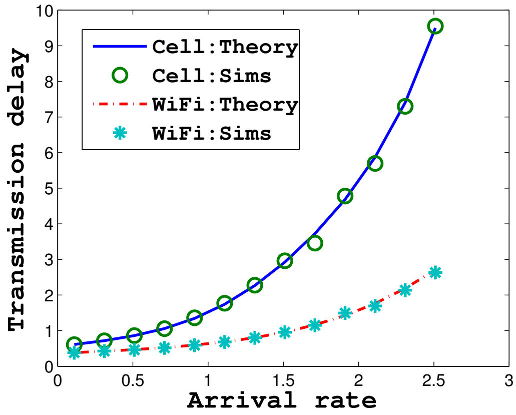

Figure 4.12: The delay for underlay spectrum access.

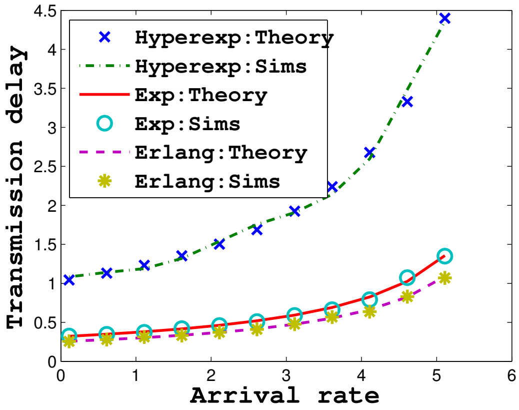

Figure 4.13: The delay for interweave spectrum access in the cellular scenario.

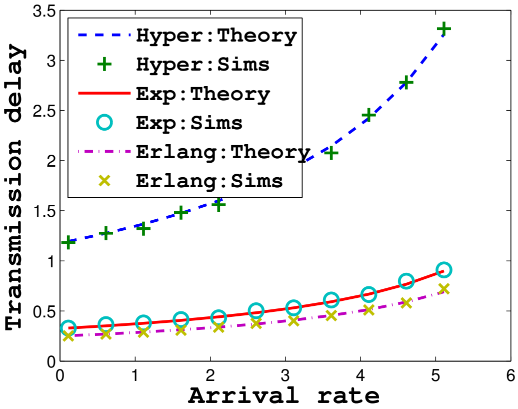

Figure 4.14: The delay for interweave spectrum access in the WiFi scenario.

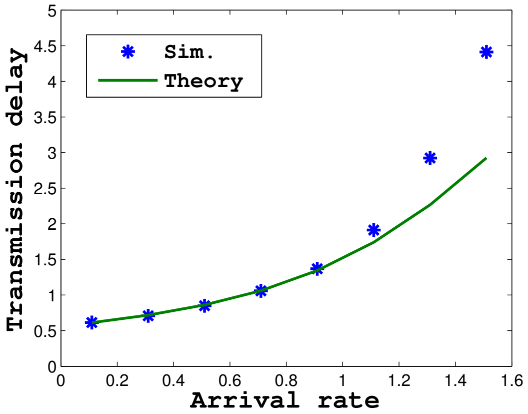

Figure 4.15: The delay for generic underlay spectrum access.

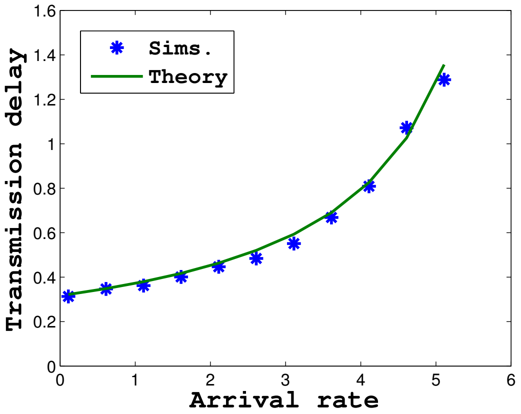

Figure 4.16: The delay for generic interweave spectrum access.

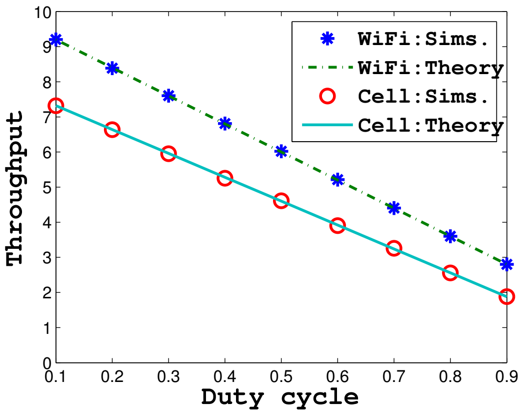

Figure 4.17: The throughput for underlay spectrum access.

all with the same mean $E[T_s] = 1$s. For the hyperexponential scanning time, we take $\eta_L = 1.9$ and $\eta_V = 0.1$. The probability of having a large scanning time (far away channel) is 0.05. The coefficient of variation for this distribution is around 3.

As the plots show, the theory is correct and provides an excellent match with simulations for both the cellular and WiFi networks for different scanning time distributions. Again as for the underlay access, the delay in the WiFi is lower compared to the cellular network, for similar reasons. Another outcome is that the average delay has the lowest value for Erlang distributed scanning time, while the worst performance is achieved for hyperexponential distribution. The above conclusion is in line with our analytical outcome of Section 4.3. As our analysis suggest, higher (lower) variability in scanning time leads to higher (lower) variability in the service time, which leads further to higher (smaller) delays. Observing the curve corresponding to the cellular case in Fig. 4.12 and Fig. 4.13, it can be noticed that the delay in the interweave access is lower than in the underlay. For the cellular scenario with exponential scanning time, and $\lambda = 1$, Table 4.2 suggests that the maximum average scanning time should be 2.8s. In our case $E[T_s]$ is much smaller (1s). Hence, the superiority of the interweave mode.

In the previous scenarios, we have used realistic values for the transmission rates and WiFi availabilities, but we have assumed exponential distributions for ON and OFF periods, according to our model. While the actual distributions are subject to the PU activity pattern, measurement studies [18, 19] suggest these distributions to be "heavy-tailed". It is thus interesting to consider how our model's predictions fare in this (usually difficult) case. To this end, we consider a scenario with "heavy-tailed" ON/OFF distributions (Bounded Pareto-BP), with parameters $L_{ON} = 1.31, H_{ON} = 200, \alpha_{ON} = \alpha_{OFF} = 1.2, L_{OFF} = 2.9, H_{OFF} = 200$. Due to space limitations, we focus on the cellular scenario. The other parameters are the same as for the scenarios of Fig. 4.12 and 4.13. Figure 4.15 compares the average file delay for the underlay access against our theoretical prediction, while Fig. 4.16 does the same for interweave mode. In the later one, the scanning time is exponential with mean 1s. Interestingly, our theory still offers a reasonable prediction accuracy, despite the considerably higher variability of ON/OFF periods in this scenario. Although we cannot claim this to be a generic conclusion for any distribution, the results emphasize the utility of our models in practice.

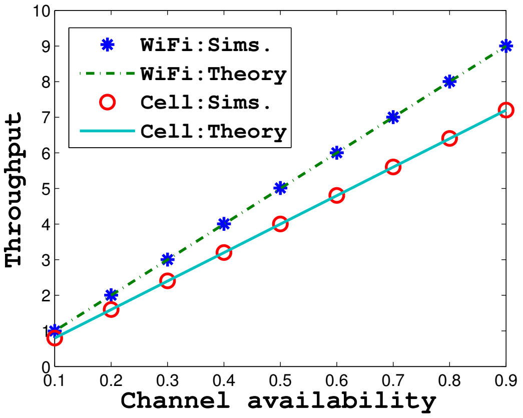

Figure 4.18: The throughput for interweave spectrum access.

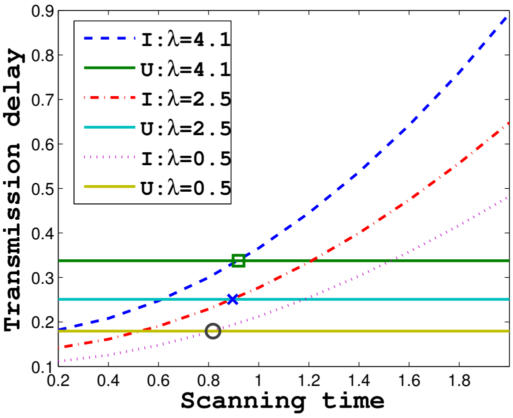

Figure 4.19: The static delay policy for different λ and exp. scanning times.

4.5.2 Validation of the throughput models

Having validated our analytical predictions for the delay in both access modes, we proceed further with validating the throughput models of Section 4.4. As mentioned earlier, these models hold for any generic distribution of ON and OFF periods. First, we will consider the throughput for the underlay mode. For the cellular network, we take the same input parameters for the ON periods and data rates as those presented in Section 4.5.1. Fig. 4.17 illustrates the average throughput for different values of the duty cycle. The first thing to observe is that there is an excellent match between theory and simulations. As can be observed, the throughput decreases as PUs are more active. This decrease is linear to the duty cycle, as suggested by Eq.(4.104). On the same plot we also show analytical and simulation results for the WiFi scenario. The conclusions remain unchanged.

Next, we continue with the validation for the interweave access. For both scenarios, the average scanning time is 0.5 s, and is Bounded Pareto distributed. The other parameters remain exactly the same as in the previous corresponding plots. Fig. 4.18 shows the average throughput against the channel availability for both networks. As opposed to the underlay model validation, we use the *channel availability* on the x-axis. It denotes the % of time the SU can transmit, and is equal to $\frac{E[T_{ON}]}{E[T_{ON}]+E[T_s]}$. Now, duty cycle does not mean much, since the SU does not reside in a single channel and also different channels have different duty cycles. However, the results in the two figures can be easily related by comparing the value for duty cycle X (in Fig. 4.17) to the respective availability value $1-X$ (in Fig. 4.18). The throughput values in Fig. 4.17 are up to some point higher than those of Fig. 4.18, which is in line with the results of Eq.(4.104) and Eq.(4.105).

4.5.3 Delay minimization policies

In this section, we would like to perform a more detailed comparison of the underlay and interweave access modes. Our first goal is to examine the "static" version of the two policies, and validate our analytical predictions regarding when the one or the other will perform better. Our second goal is to consider the dynamic policy, and see if and when it can outperform both simple

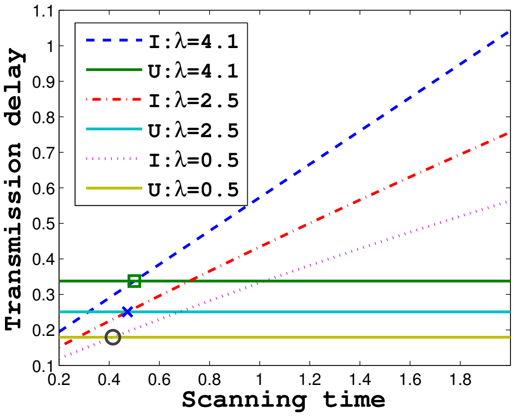

Figure 4.20: The static delay policy for different λ and hyperexp. scan. times.

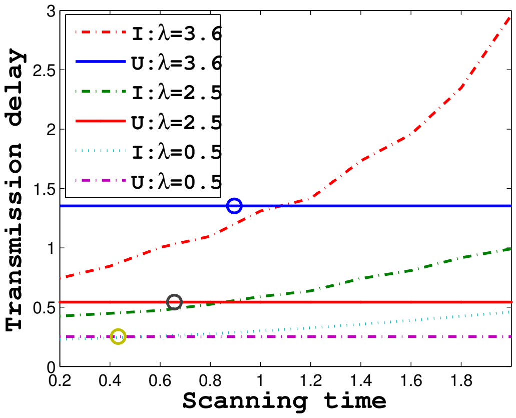

Figure 4.21: The generic static delay policy for different λ.

policies.

As another interesting scenario, we consider the underlay access with parameters: $\eta_H = 0.1$, $\eta_L = 1$, $\mu_H = 10$, $\mu_L = 0.5$. Fig. 4.19 shows the average file delay (denoted as I) against different average scanning times (exp. distributed), for three different traffic intensities (low, medium, high). On the same plot, for each traffic intensity the corresponding underlay delay (denoted as U) is shown as well. Finally, the theoretical maximum values for the expected scanning times (Table 4.2), for which the interweave access mode outperforms underlay access are depicted with small circles. For the sparse traffic case, the interweave starts to become better for scanning times lower than 0.8s. The first thing to observe is that the predicted maximum value for the scanning time (i.e. the crossing point) is correct. The second important outcome is that this boundary is higher when the load increases. This is due to the fact that for higher loads the queueing delay is the largest delay component. Hence, it is worth waiting some time, find an idle channel and then get rid of the queued data at higher rate. We have also noticed that increasing the load further leads to smaller and smaller increases of this crossing point.

Next, we consider the hyperexponential distribution for the scanning times with parameters $\eta_L = 6$, $\eta_V = 0.4$, and the probability p taking values such that a given average scanning time is achieved. The coefficient of variation observed is in the range (2,2.5). The other parameters are identical as for the previous scenario. Fig. 4.20 shows the average delay. Due to the higher variance of the scanning time, the crossing point between underlay and interweave are lower compared to the scenario of Fig. 4.19.

The previous results validate the correctness of our analytical expression for the crossing point between underlay and interweave delay. To further corroborate the utility of this expression, we consider next a scenario with generic distributions for the ON/OFF periods. We consider again three traffic intensities (sparse, medium, heavy). The other parameters are $\mu_H = 5$ Mbps, $\mu_L = 2$ Mbps, $\eta_H = 0.1$, $\eta_L = 1$. The ON periods are subject to a Bounded Pareto distribution with parameters $L = 2.375$, $H = 1000$, $\alpha = 1.2$. Fig. 4.21 shows the average file delay for different exponential scanning times for the three different regimes of traffic intensity vs. the corresponding underlay scenario. We also show (small circles) the theoretical maximum scanning time that still provides better performance for the interweave mode. As can be seen

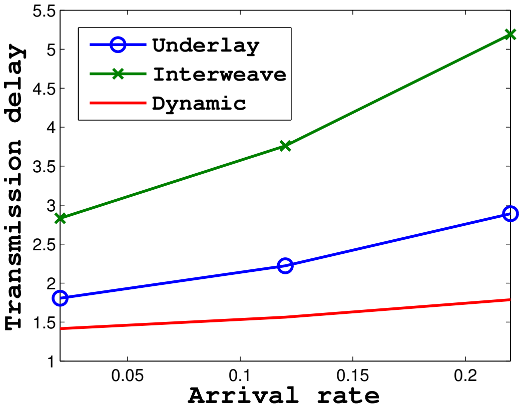

Figure 4.22: The dynamic delay policy for Pareto OFF periods.

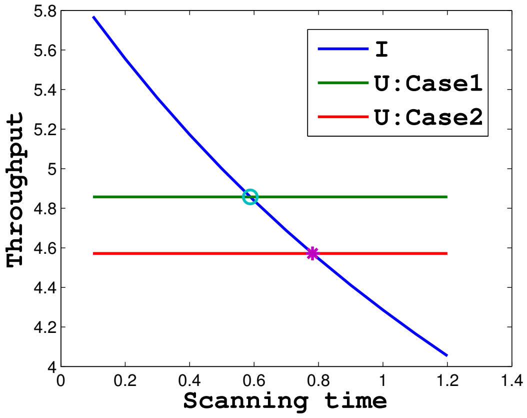

Figure 4.23: The throughput static policy for two different low period data rates.

from Fig. 4.21, for sparse traffic, our theoretical result can predict accurately the bound, despite the different assumptions on the distributions. The worst case mismatch can be observed for the highest arrival rate considered (load close to capacity), and is in the order of 10%. We can conclude that our analytical expression related to the relative performance of the two access modes, can relatively accurately predict which policy would lead to better delays, even if key system parameters, like the variability of channel availability durations, depart from our theory assumptions.

Dynamic delay policy. As discussed in Section 4.3.4, the dynamic policy can offer additional performance benefits, when the PU activity periods (i.e. the OFF periods in the underlay mode) are subject to a probability distribution with a decreasing failure rate (i.e. with very high variability). We consider a scenario where the low (OFF) periods have Bounded Pareto distribution, with parameters $L = 0.2$, $H = 100$, $\alpha = 1.2$. The average scanning time is $E[T_s] = 1$s. The other parameters are the same as for the cellular scenario. Fig. 4.22 shows the average delay vs. the arrival rate. According to the static policy, the underlay mode is better than the interweave. However, the best result is achieved with the dynamic policy, which offers an additional delay reduction of 20-50%.

4.5.4 Throughput maximization policies

In this last part of the simulations section, we turn our attention to the throughput optimization policies. We would like to perform a more detailed comparison of the two modes. First, we examine the “static” version of the two policies, and validate our analytical predictions regarding when the one or the other will perform better. After that, we show that the dynamic policy can outperform both simple policies. Both ON and OFF periods are exponentially distributed with the following parameters: $\eta_H = 0.4$, $\eta_L = 1$, $\mu_H = 6$ Mbps, $\mu_L = 2$ Mbps (Case 1). We compare it with the interweave scenario with $\eta_H = 0.4$, and exponential scanning times. Fig. 4.23 illustrates the average data rate for different average scanning times. The performance of the underlay mode remains unchanged, as it does not depend on scanning time. Case 2, shown also on the plot, corresponds to the underlay access with $\mu_L = 1$ Mbps.

The theoretical maximum values for $E[T_s]$, up to which the interweave access mode provides

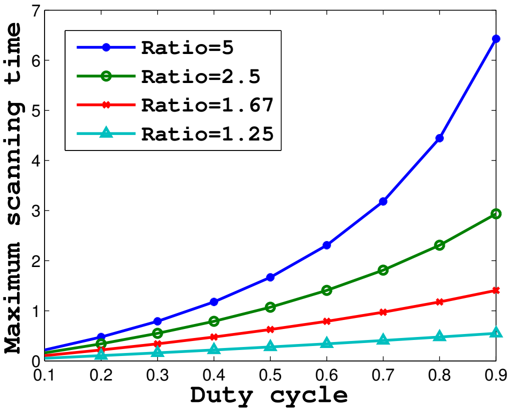

Figure 4.24: Max. Scanning times for different ratios of data rates.

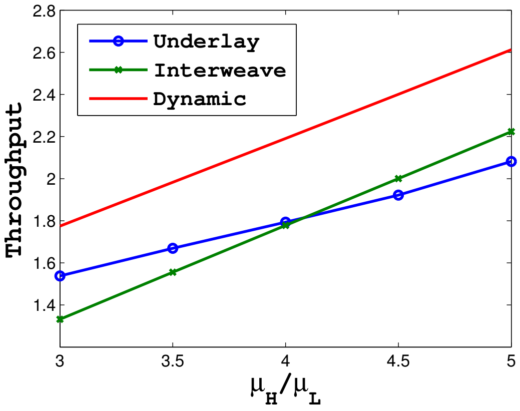

Figure 4.25: The throughput dynamic policy for Pareto OFF periods.

better throughput than underlay access are also shown there as circles. As can be seen, there is a perfect match. Observing Fig. 4.23, one can see that it is better to stay in the underlay mode if the scanning time is larger than 0.6s for Case 1, and 0.8 s for Case 2. However, for lower values of the scanning time, interweave access should be the preferred choice by the SU. When the channel conditions are worse for the SU (related to its vicinity to the SU, higher power of the PU etc.), this maximum $E[T_s]$ increases as well. As expected, the average throughput is higher for Case 1, due to the higher data rates in the low periods.

After validating our throughput results, we will consider another scenario that provides further insights on the crossing point of $E[T_s]$. For that purpose, we consider an identical scenario as before in terms of the network availability, but with different data rates. The ON period data rate is 5 Mbps now. Fig. 4.24 illustrates the maximum scanning time vs. the duty cycle for the underlay CRN model. We consider four cases with different data rate-OFF periods (according to respective ratios). From the plot it is evident that the maximum scanning time changes almost linearly with the duty cycle in the good channels ($\mu_H \approx \mu_L$), while for worse channel conditions when the PU is present, this curve is not linear any more.

Throughput dynamic policy. As in the case of delay, our theory suggested that a dynamic throughput policy could perform better than either underlay or interweave, but only when the OFF periods are subject to decreasing failure rate distributions. We consider the scenario with Pareto OFF periods (the parameters are identical to those of Fig. 4.22). The average duration of the ON and OFF periods, as well as for the scanning time are the same as the previous case. Fig. 4.25 shows the average throughput for different policies. The dynamic policy further increases the throughput by additional 20%. This is consistent with our claims in Section 4.4.2 that the throughput can be improved further when underlay is better. Since the OFF periods are Pareto distributed, some of them will be very long and will have decisive impact on the average data rate. Hence, finding another idle channel will improve the performance.

4.6 Related Work

There has been a significant amount of research in the areas of interweave and underlay access for CRN. The underlay CRN have been the focus of [47,53,54]. So far, more attention was paid to the interweave techniques [23,35,48,50]. All these focus on algorithmic design.

Some analysis for underlay CRN was performed in [55–57]. Authors in [55] analyze the performance of SU networks based on the interference temperature model, and compare the performance of a single and multiple SU networks. In [56], the effect of different interference thresholds on the transmission rate is studied together with the impact of the SU on the PU. The impact of imperfect channel channel state information on the relay selection schemes for underlay CRN is studied in [57]. Some of the works that deal with performance analysis related to interweave access in CRN are [21, 22, 29, 30, 58, 59]. In [58] and [59], the capacity of the voice service is analyzed. Furthermore, authors in [58] propose two cognitive MAC protocols for channel access. In [59], two call admission control algorithms are proposed.

The majority of performance analysis work in underlay CRN are information-theoretic considering only the capacity, and mostly based on interference temperature [13, 55, 56]. These works are orthogonal to ours, since they are concerned with the allowed transmission power (and thus the data rate) by estimating the channel quality between PUs and SUs. We can use those results to determine the data rate that we use in our analysis. In contrast, much fewer studies exist about the per file/flow delay in such systems. In [60], authors propose an M/G/1 queueing system with finite buffer and timeout, and derive different metrics, among which the delay as well. However, their results can be obtained only numerically and as such are difficult to be interpreted and used in solving different optimization problems. A similar conclusion can be drawn for [61]. The authors there also model the SU activity with an M/G/1 queue. However, they do not show how to find the second moment of the service time in the P-K formula. On the other hand, we propose an analytical queueing model that leads to a closed-form expression for delay and throughput, which not only provides more insight into the effect of system parameters, but also allows to analytically compare and optimize the policies. Although in our analysis for underlay CRN we are restricted to the exponential distributions for analytical tractability, we show by simulations that our model can predict accurately the delay even when dealing with heavy-tailed distributions.

As far as interweave CRN is concerned, there exist more analytical works that aim to derive the average packet delay. Most of the works model the PU activity with stochastic ON-OFF process. Some recent work [21,29,30] have capitalized on the measurement-based study of [26], in which the Poisson approximation seems to be decent for call arrivals, but call duration is generically distributed. These works model SUs together with PUs, as an M/G/1 system with priorities and preemption. M/G/1 systems with priorities have been long analyzed (see e.g. [62]). Nevertheless, there are some important caveats in the above models. *First*, they consider the problem in the packet level and model the problem as an M/G/1 system with preemptive-resume, while in reality SU packets will collide with a PU when it reclaims channel back, and have to retransmit in the next available period. On the other hand, our model can capture both the resume and retransmitting feature of real wireless systems. *Second*, the M/G/1 systems with priorities can capture only exponential scanning times. Our interweave model holds for generic scanning times. We do not need the Poisson assumption for the PU traffic, as opposed to the priority models, since in our model the time between two PU arrivals is the sum of an exp. (ON period) and a generic (scanning time) random variable, which is generic.

In [63], a model that is able to capture generic PU activity for interweave CRN is proposed with the delay as the only metric considered. Here we analyze both access modes, with delay and throughput being our metrics of interest. However, it is assumed that after losing a channel the SU waits in the same channel. Here we assume that the SU, as soon as the channel it is using is reclaimed back by the PU, ceases its transmission and initiates the scanning procedure to find another available channel, and model with different distributions the scanning time. Here we analyze both access modes, with delay and throughput being our metrics of interest.

While certainly there is a relatively large number of papers proposing models for underlay and interweave modes separately, there are only few studies comparing them. Simply taking one model for the underlay and one for the interweave from another reference is not enough, due to the mismatch in assumptions. Very few studies exist that directly compare the performance between the access modes. Comparing results from different papers is not straightforward due to different assumptions, non-closed form expressions, etc. In this chapter, we propose models that enable us to do analytical comparisons between the modes. A study closer to ours in its aim is [64], where a hybrid CR system is investigated, in which a SU probabilistically changes its mode of operation for throughput optimization. However, the delay metric is not considered there, and the arrival process at the PU is quite restrictive (Bernoulli). On the other hand, we propose policies that are able to optimize both the delay and throughput, and our models hold for generic PU arrivals. The authors in [65] have done performance analysis for the three spectrum access techniques (interweave, underlay and overlay). They provide theoretical limits for all the schemes assuming power constraints and additive white Gaussian noise. Nevertheless, they consider only theoretical capacities, without considering other metrics of interest, and there is no analytical comparison. The authors compare the performance only by simulating few scenarios. We, on the other hand, provide analysis for the delay and throughput for underlay as well as for interweave access under a wide variety of scanning distributions.

Summarizing, the main novelties of this chapter compared to different related works revolve around the following key points: (i) we make a direct analytical comparison of interweave and underlay access; (ii) we consider both key metrics of interests, delay and throughput; (iii) we provide closed form expressions for all cases; (iv) we use our results to propose optimal hybrid policies for both metrics.

4.7 Conclusion

In this chapter, we have proposed queueing analytic models for the delay analysis of interweave and underlay spectrum access, and we validated them against realistic scenarios. Besides that, we have proposed models relying on renewal-reward theory to determine the average throughput in both modes. We have also provided the bounds on the scanning times for which the interweave access outperforms the underlay. To further improve the performance in terms of both the throughput and delay, we have proposed dynamic policies. In future work, we plan to extend our model to capture generic file sizes.

Part II

Modeling, Analysis and Optimization of Mobile Data Offloading

Chapter 5

Performance Analysis of On-the-Spot Mobile Data Offloading

An unprecedented increase in the mobile data traffic volume has been recently reported due to the extensive use of smartphones, tablets and laptops. This is a major concern for mobile network operators, who are forced to often operate very close to (or even beyond) their capacity limits. Recently, different solutions have been proposed to overcome this problem. The deployment of additional infrastructure, the use of more advanced technologies (LTE), or offloading some traffic through Femtocells and WiFi are some of the solutions. Out of these, WiFi presents some key advantages such as its already widespread deployment and low cost. While benefits to operators have already been documented, it is less clear how much and under what conditions the user gains as well. Additionally, the increasingly heterogeneous deployment of cellular networks (partial 4G coverage, small cells, etc.) further complicates the picture regarding both operator- and user-related performance of data offloading. To this end, in this chapter we propose a queueing analytic model that can be used to understand the performance improvements achievable by WiFi-based data offloading, as a function of WiFi availability and performance, user mobility and traffic load, and the coverage ratio and respective rates of different cellular technologies available. We validate our theory against simulations for realistic scenarios and parameters, and provide some initial insights as to the offloading gains expected in practice.

5.1 Introduction

Lately, an enormous growth in the mobile data traffic has been reported. This increase in traffic demand is due to a significant penetration of smartphones and tablets in the market, as well as Web 2.0 and streaming applications which have high-bandwidth requirements. Furthermore, Cisco [7] reports that by 2017 the mobile data traffic will increase by 13 times, and will climb to 13.2 exabytes per month, with approximately 5.2 billion users. Mobile video traffic will comprise 66% of the total traffic, compared to 51% in 2012 [7].

This increase in traffic demand is overloading the cellular networks (especially in metro areas) forcing them to operate close to their capacity limits causing a significant degradation of user experience. Possible solutions to this problem could be the complete upgrade to LTE or LTE-advanced, as well as the deployment of additional network infrastructure [8]. However, these solutions may not be cost-effective from the operators' perspective: they imply an increased cost (for power, location rents, deployment and maintenance), without similar revenue increases, due to flat rate plans, and the fact that a small number of users consume a large amount of traffic (3% of users consume 40% of the traffic [10]). As a result, LTE has only been sporadically deployed, and it is unclear whether providers will choose to upgrade their current deployments, and if in fact the additional capacity would suffice [8, 9].

A more cost-effective way of alleviating the problem of highly congested mobile networks is by offloading some of the traffic through Femtocells (SIPTO, LIPA [11, 66]), and the use of WiFi. In 2012, 33% of total mobile data traffic was offloaded [7]. Projections say that this will increase to 46% by 2017 [7]. Out of these, data offloading through WiFi has become a popular solution. Some of the advantages often cited compared to Femtocells are: lower cost, higher data rates, lower ownership cost [8], etc. Also, wireless operators have already deployed or bought a large number of WiFi access points (AP) [8]. As a result, WiFi offloading has attracted a lot of attention recently.

The current approach to offloading is that of *on-the-spot offloading*: when there is WiFi availability, all data is sent over WiFi, otherwise traffic is transmitted over the cellular network. This is easy to implement, as smart phones currently are only able to use one interface at a time[1]. More recently, *delayed offloading* has been proposed [12, 68]: if there is currently no WiFi availability, (some) traffic can be delayed up to a chosen time threshold, instead of being sent immediately over the cellular interface. If up to that point, no AP is detected, the data is transmitted over the cellular network. Nevertheless, delayed offloading is still a matter of debate, as it is not known to what extent users would be willing to delay a packet transmission. It also requires disruptive changes in higher layer protocols (e.g. TCP) [69].

As a result, in this chapter, we will focus on on-the-spot offloading. Although on-the-spot offloading is already used and there is some evidence that it helps reduce the network load [7,8], it is not clear how key factors such as WiFi availability, average WiFi and cellular performance, and existence of advanced cellular technologies (e.g. LTE) affect this performance. What is more, it is still a matter of debate if offloading offers any benefits to the user as well (in terms of performance, battery consumption, etc.). Studies suggest that such benefits might strongly depend on the availability and performance of WiFi networks, or type of user mobility [12,68], etc. Finally, the increasingly heterogeneous deployment of current and future cellular environments, with partial coverage by different technologies (GPRS, EDGE, HSPA/HSPDA, LTE) and a growing number of "small cells" (e.g. femto, pico) utilized, further complicates these questions.

To this end, in this chapter we propose a queueing analytic model for performance analysis of on-the-spot mobile data offloading. Our contributions can be summarized as follows:

- We consider a simple scenario where the user can choose between WiFi and a single cellular technology, and derive general formulas, as well as simpler approximations, for the expected delay and offloading efficiency (Section 5.2).
- We generalize our analysis to the case where multiple cellular technologies (and respective rates) are available to a user, e.g. depending on her location, and/or different rates are offered by the same technology (e.g. rate adaptation, indoor/outdoor, etc.) (Section 5.3).
- We validate our model in scenarios where most parameters of interest are taken from real measured data, and which might diverge from our assumptions (Section 5.4).

[1] Using both interfaces in parallel, as well as per flow offloading (IFOM) are currently being considered in 3GPP [11,67].

- We use our model to provide some preliminary answers to the questions of offloading efficiency and delay improvements through WiFi-based offloading (Section 5.4.5).

We discuss some related work in Section 5.5, and conclude in Section 5.6.

Before proceeding to our model and results, we would like to stress that the proposed model and analysis clearly is not meant to capture all the details of the various wireless access technologies, as such a task would be beyond the scope of a single paper, and would be too complex to yield useful insights. Nevertheless, we believe it is an important first step towards analytical tools that can be used to understand the performance of offloading in current and future network deployments, and to allow operators to make informed design decisions.

5.2 Analyzing Offloading in Homogeneous Cellular Networks

5.2.1 Problem Setup

We consider a mobile user that enters and leaves zones with WiFi coverage. The average time until a user enters such a zone, and the time she stays there, depend on the user's mobility (e.g. pedestrian, vehicular), the environment (e.g. rural, urban), the access point (AP) density, etc. E.g., increasing AP density would (i) increase the total WiFi coverage time, (ii) shorten the transition times between WiFi covered areas. On the other hand, higher user speeds would also shorten the expected time until one finds WiFi connectivity again, but would also lead to shorter connections (with a given AP).

In this chapter, we will first assume a "simple" cellular coverage environment, where the user has access to a single cellular data access technology, in addition to WiFi. Without loss of generality, we assume that there is always cellular network coverage (we allow coverage holes in the next section). Whenever there is coverage by some WiFi AP, all traffic will be switched over to WiFi. As soon as the WiFi connectivity is lost, the traffic will be transmitted through the cellular network.

Definition 3. *[WiFi Availability Model] We model the WiFi network availability as an ON-OFF alternating renewal process [20].*

- *Time consists of ON-OFF cycles,* $\left(T_{ON}^{(i)}, T_{OFF}^{(i)}\right), i \in \mathcal{N}^+$*, as shown in Fig. 5.1. ON periods represent the presence of WiFi connectivity, while during OFF periods only cellular access is available.*
- $T_{ON}^{(i)} \sim$ *Exponential*(η_w)*, and independent from other (ON or OFF) periods. The data transmission rate during these periods is denoted with* μ_w*.*
- $T_{OFF}^{(i)} \sim$ *Exponential*(η_c)*, and independent from other (ON or OFF) periods, with an offered data rate equal to* μ_c*.*

Transmission rates in practice: Transmission rates in real networks are not stable and are affected by signal quality (e.g. through rate adaptation), as well as the presence of other users. As a result the actual rate might change from ON period to ON period (or OFF to OFF), because e.g. the newly encountered WiFi AP is more congested or is further away. Consequently the above nominal rates μ_c (μ_w) correspond to the effective rate allocated by the AP or BS to the user (e.g. based on channel quality, other users' presence, and the respective MAC protocol)

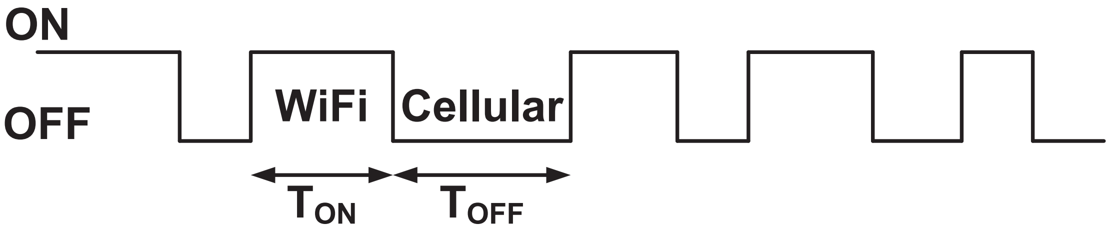

Figure 5.1: The WiFi network availability model.

and is an *average* over all OFF (ON) periods (which might be known, for example, from general city-wide or time-of-day statistics). We validate this assumption against arbitrary, randomly generated rates as well in Section 5.4, and also generalize our analytical result to scenarios with more than 2 rates offered, in the next section.

Next, we will assume that while a user is moving between areas with and without WiFi, she also generates new "data requests". Each such request generates a *flow* of data, e.g. corresponding to a file upload, data synchronization (e.g. DropBox, Flickr, Google+), etc. in the uplink direction, or a web page access, file download, news feed, etc. in the downlink direction. We will use the terms "file", "flow", and "message" interchangeably, to refer to a concatenation of packets corresponding to the same application request (e.g., a downloaded file, a photo uploaded on Facebook), and which must be uploaded (downloaded) in completion for the application request to be satisfied. Identifying and delimiting flows is a well researched problem and beyond the scope of this chapter (see e.g. [70]). We assume the following simple model for the arrival and service of new flows at the user.

Definition 4. *[User Traffic and Service Model] Without loss of generality, we focus in one direction (uplink or downlink) and consider the following simple queueing model.*

- *New data requests arrive as a Poisson process with rate λ.*
- *The size of the flow/file corresponding to a data request is random and exponentially distributed.*
- *A flow arriving to find another flow currently in transmission will queue (the size of the queue assumed to be infinite), and will be served in a First Come First Served (FCFS) order.*
- *The service rate for the flow at the head of the queue will be equal to the WiFi rate (cellular rate), if the system is currently during an ON (OFF) period.*
- *A switch in connectivity might sometimes occur while a file transmission is ongoing. We assume that network transitions do not cause any interruptions to the traffic flow in service (i.e. the transmission continues with the new rate immediately), and ignore vertical handover delays, if any. We discuss later how such interruptions can be included in the model.*

The above two definitions capture our basic model for the offloading problem. We would like to stress that we do not claim that the actual availability periods or flow sizes are exponentially

Table 5.1: Variables and Shorthand Notation.

Variable	Definition/Description
T_{ON}	Duration of ON (WiFi) periods
T_{OFF}	Duration of periods (OFF) without WiFi connectivity
λ	Average packet (file) arrival rate at the mobile user
$\pi_{i,c}$	Stationary probability of finding i files in cellular state
$\pi_{i,w}$	Stationary probability of finding i files in WiFi state
π_c	Probability of finding the system under cellular coverage only
π_w	Probability of finding the system under WiFi coverage
η_w	The rate of leaving the WiFi state
η_c	The rate of leaving the cellular state
μ_w	The service rate while in WiFi state
μ_c	The service rate while in cellular state
$E[S]$	The average service time
$E[T]$	The average system (transmission) time
$\rho = \lambda E[S]$	Average user utilization ratio
OE	The offloading efficiency
μ_k	The service rate while in network type k
η_k	The rate of leaving the type k network

distributed (in fact, evidence for the contrary exists from measurement studies). The above assumptions are only made to keep our analysis tractable. In Section 5.2.6, we show how to extend our model to generic file size distributions. One could also try to extend our framework to arbitrary ON and OFF distributions that can be approximated by Coxian distributions [71], fitting the first three moments to the real duration of the ON (OFF) period. In this case, there would be more states along one dimension in the Markov chain, and one could employ matrix-analytic methods [49]. However, these would offer little analytical insight. For this reason, we will test instead our model and its predictions against scenarios with realistic ON/OFF distributions in Section 5.4. Before proceeding further, we summarize in Table 6.1 some useful notation that will be used throughout the rest of the chapter.

5.2.2 Delay Analysis

From the problem description, it is easy to see that the above setup corresponds to a single server queueing system whose rate varies according to an external (random) process. We are interested in the total time a new user flow spends in the system (service+queueing) until it is served. It is a key metric of user experience that we would like to analyze. This is often referred to as the *system time* in queueing theory. However, we will also use the term *transmission delay* interchangeably.

Given the assumptions in the previous subsection, the on-the-spot offloading system can be modeled with a 2D Markov chain, as shown in Fig. 5.2.

$\pi_{i,w}$ denotes the stationary probability of having WiFi coverage *and* finding i flows in the system queue (i.e. $i-1$ waiting and one being transmitted over WiFi).

$\pi_{i,c}$ denotes the stationary probability of having only cellular coverage *and* i flows in the

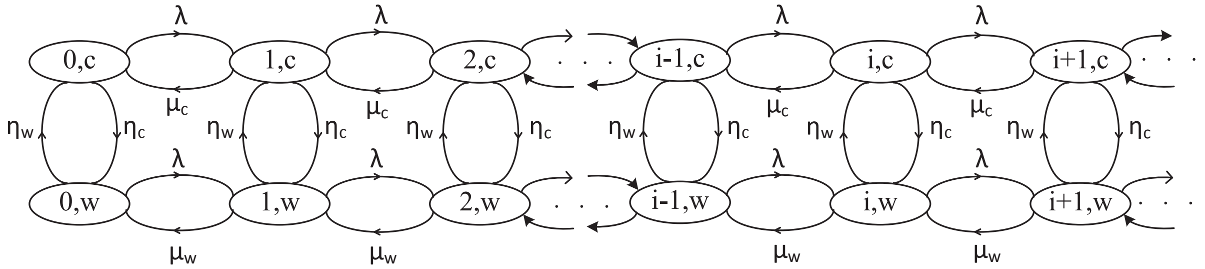

Figure 5.2: The 2D Markov chain for the simple on-the-spot mobile data offloading scenario of Def. 3 and 4.

system.

Writing the balance equations for this chain gives

$$\begin{aligned}
\pi_{0,c}(\lambda+\eta_c) &= \pi_{1,c}\mu_c + \pi_{0,w}\eta_w && (5.1)\\
\pi_{0,w}(\lambda+\eta_w) &= \pi_{1,w}\mu_w + \pi_{0,c}\eta_c && (5.2)\\
\pi_{i,c}(\lambda+\eta_c+\mu_c) &= \pi_{i-1,c}\lambda + \pi_{i+1,c}\mu_c + \pi_{i,w}\eta_w, (i>0) && (5.3)\\
\pi_{i,w}(\lambda+\eta_w+\mu_w) &= \pi_{i-1,w}\lambda + \pi_{i+1,w}\mu_w + \pi_{i,c}\eta_c, (i>0) && (5.4)
\end{aligned}$$

We define the probability generating functions (PGF) for both the cellular and WiFi

$$G_c(z) = \sum_{i=0}^{\infty} \pi_{i,c} z^i, \text{and } G_w(z) = \sum_{i=0}^{\infty} \pi_{i,w} z^i, |z| \le 1.$$

We can rewrite Eq.(5.1) and Eq.(5.3) as

$$\begin{aligned}
\pi_{0,c}(\lambda+\eta_c+\mu_c) &= \pi_{0,w}\eta_w + \pi_{1,c}\mu_c + \pi_{0,c}\mu_c\\
\pi_{i,c}(\lambda+\eta_c+\mu_c) &= \pi_{i-1,c}\lambda + \pi_{i,w}\eta_w + \pi_{i+1,c}\mu_c, (i>0) && (5.5)
\end{aligned}$$

We multiply each of the equations from Eq.(5.5) by z^i and sum over all i's. After some calculus this yields

$$(\lambda+\eta_c+\mu_c)G_c(z) = \lambda z G_c(z) + \eta_w G_w(z) + \frac{\mu_c}{z}\left(G_c(z) - \pi_{0,c}\right) + \pi_{0,c}\mu_c. \quad (5.6)$$

Repeating the same process with Eq.(5.2) and Eq.(5.4), we get

$$(\lambda+\eta_w+\mu_w)G_w(z) = \lambda z G_w(z) + \eta_c G_c(z) + \frac{\mu_w}{z}\left(G_w(z) - \pi_{0,w}\right) + \pi_{0,w}\mu_w. \quad (5.7)$$

Equations Eq.(5.6) and Eq.(5.7) define a system of equations in $G_c(z)$ and $G_w(z)$, from which we can get

$$f(z)G_c(z) = \pi_{0,w}\eta_w\mu_w z + \pi_{0,c}\mu_c\left[\eta_w z + (\lambda z - \mu_w)(1-z)\right], \quad (5.8)$$

where

$$f(z) = \lambda^2 z^3 - \lambda(\eta_c+\eta_w+\lambda+\mu_w+\mu_c)z^2 + (\eta_c\mu_w + \eta_w\mu_c + \mu_c\mu_w + \lambda\mu_w + \lambda\mu_c)z - \mu_c\mu_w. \quad (5.9)$$

It can be proven that the polynomial in Eq.(5.9) has only one root in the open interval $(0, 1)$ [52]. This root is denoted as z_0. Setting $z = z_0$ into Eq.(5.8) gives

$$\pi_{0,w}\eta_w\mu_w z_0 + \pi_{0,c}\mu_c\left[\eta_w z_0 + \lambda z_0(1 - z_0) - \mu_w(1 - z_0)\right] = 0.$$

Writing the balance equation through a vertical cut between states containing $i - 1$ and i files, we will obtain

$$\lambda(\pi_{i-1,c} + \pi_{i-1,w}) = \mu_c\pi_{i,c} + \mu_w\pi_{i,w}. \tag{5.10}$$

If we take the sum from $i = 1$ to ∞ in Eq.(5.10), we get

$$\lambda = \mu_c(\pi_c - \pi_{0,c}) + \mu_w(\pi_w - \pi_{0,w}), \tag{5.11}$$

where $\mu = \pi_c\mu_c + \pi_w\mu_w$, and $\pi_w = \sum_{i=0}^{\infty}\pi_{i,w}$ is the steady-state probability of finding the system in some region with WiFi availability. Using standard Renewal theory [20] we get $\pi_w = \frac{\eta_c}{\eta_c+\eta_w}$. Similarly, for periods with only cellular access we have $\pi_c = \frac{\eta_w}{\eta_c+\eta_w}$.

Solving the system of equations consisting of Eq.(5.11) and the equation preceding Eq.(5.10), we get[2]

$$\pi_{0,c} = \frac{\eta_w(\mu - \lambda)z_0}{\mu_c(1 - z_0)(\mu_w - \lambda z_0)}, \tag{5.12}$$

$$\pi_{0,w} = \frac{\eta_c(\mu - \lambda)z_0}{\mu_w(1 - z_0)(\mu_c - \lambda z_0)}. \tag{5.13}$$

Finally, for $G_c(z)$ and $G_w(z)$ we have from Eq.(5.8) and Eq.(5.6)

$$G_c(z) = \frac{[\eta_w(\mu - \lambda)z + \pi_{0,c}\mu_c(1 - z)(\lambda z - \mu_w)]}{f(z)}, \tag{5.14}$$

$$G_w(z) = \frac{[\eta_c(\mu - \lambda)z + \pi_{0,w}\mu_w(1 - z)(\lambda z - \mu_c)]}{f(z)}. \tag{5.15}$$

We define two new quantities $E[N_c] = \sum_{i=0}^{\infty} i\pi_{i,c}$ and $E[N_w] = \sum_{i=0}^{\infty} i\pi_{i,w}$. Hence, we have $E[N_c] = G_c^{'}(1)$ and $E[N_w] = G_w^{'}(1)$.

It is easy to see then that the average number of files in the system is $E[N] = E[N_c] + E[N_w]$. Replacing $z = 1$ in the derivatives of Eq.(5.14)-(5.15) and summing them up, we get for the average number of files in the system

$$E[N] = \frac{\lambda}{\mu - \lambda} + \frac{\mu_c(\mu_w - \lambda)\pi_{0,c} + (\mu_c - \lambda)(\mu_w(\pi_{0,w} - 1) + \lambda)}{(\eta_c + \eta_w)(\mu - \lambda)}.$$

Finally, using the Little's law $E[N] = \lambda E[T]$ [20], we obtain the average file delay in on-the-spot mobile data offloading:

Result 12. *The average file transmission delay in the data offloading scenario of Definition 3 and 4 is*

$$E[T] = \frac{1}{\mu - \lambda} + \frac{\mu_c(\mu_w - \lambda)\pi_{0,c} + (\mu_c - \lambda)(\mu_w(\pi_{0,w} - 1) + \lambda)}{\lambda(\eta_c + \eta_w)(\mu - \lambda)}. \tag{5.16}$$

[2] Note that μ is *not* the average experienced service rate, except for some limiting cases, e.g. when the system is always backlogged with traffic.

Under the same assumptions, Eq.(5.16) holds for the Processor Sharing (PS) policy as well. To prove that, we need to show that the Markov chain of Fig. 5.2 is the same under the PS policy. As we have the same assumptions, the parameters λ, η_w, and η_c remain unchanged. Now, let's consider the service rates. If there are i files in the system (in either cellular or WiFi state), each one of the i files shares $\frac{1}{i}$ of the resources, i.e. has a service rate of $\frac{1}{i}\mu_{c(w)}$. Since there are i files (with identically exponentially distributed service rates), the transition rate to move from state i to $i-1$ is simply the rate of a minimum of i exponentially distributed random variables, i.e., $i \cdot \frac{1}{i}\mu_{c(w)} = \mu_{c(w)}$. Hence, we have shown that all the transition rates in Fig. 5.2 remain unchanged, and that Eq.(5.16) holds for the PS policy as well.

5.2.3 Low utilization approximation

In the previous subsection we derived a generic expression for the average delay of on-the-spot-offloading. However, the formula in Eq.(5.16) contains a root of a third order (cubic) equation, which while obtainable in closed-form, is quite complex. For this reason, in the remainder of this section we will consider simpler approximations for specific operation regimes. One such scenario of interest is when resources are underloaded (e.g. nighttime, rural areas) and/or traffic is relatively sparse (e.g. background traffic from social and mailing applications, messaging, Machine Type Communication, etc.).

For low utilization, there is almost no queueing. So, we can only consider the service time as an approximation of the total delay. We can thus use a fraction of the original Markov chain of Fig. 5.2 with only 4 states, as shown in Fig. 5.3 (i.e. number of jobs in the system ≤ 1). The service time will depend only on the probability of the flow arriving during a WiFi period (easily found to be $\frac{\eta_c}{\eta_c+\eta_w}$) or cellular only period ($\frac{\eta_w}{\eta_c+\eta_w}$)), and the amount of time to "hit" a 0 state (i.e. $\{0,c\}$ or $\{0,w\}$) from there. The latter can be derived in closed form by a simple application of first step analysis [72]. This gives us a first useful approximation, which becomes exact as $\lambda \to 0$ (the interested reader can find the detailed proof in [73]).

Low utilization approximation. *The average file transmission delay in the on-the-spot mobile data offloading for sparse traffic can be approximated by*

$$E[T] = \frac{(\eta_w+\eta_c)^2 + \eta_c\mu_c + \eta_w\mu_w}{(\mu_c\mu_w + \mu_c\eta_w + \mu_w\eta_c)(\eta_c+\eta_w)}. \tag{5.17}$$

5.2.4 High utilization approximation

Another interesting regime is that of high utilization. As explained earlier, wireless resources are often heavily loaded, especially in urban centers, due to the increasing use of smart phones, tablets, and media-rich applications. Hence, it is of special interest to understand the average user performance in such scenarios. We provide an approximation that corresponds to the region of high utilization ($\rho \to 1$), i.e. for which it holds that

$$\lambda \approx \frac{\eta_w}{\eta_c+\eta_w}\mu_c + \frac{\eta_c}{\eta_c+\eta_w}\mu_w. \tag{5.18}$$

Under this condition[3] the polynomial of Eq.(5.9) becomes

$$f(z) = (z-1)[\lambda^2 z^2 - \lambda(\mu_c+\mu_w+\eta_c+\eta_w)z + \mu_c\mu_w]. \tag{5.19}$$

[3] In the condition for stability of this queueing system the left-hand side of Eq.(5.18) should be smaller than the right-hand side.

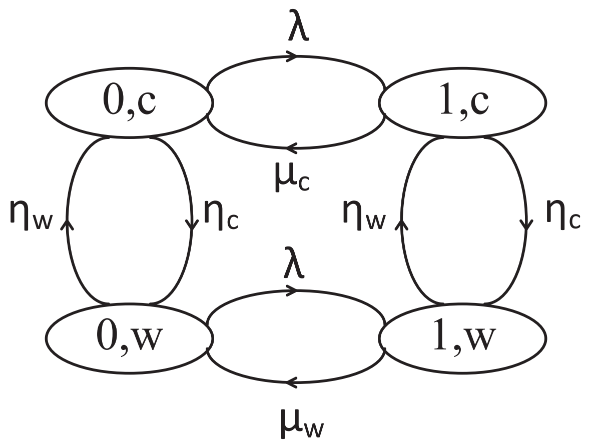

Figure 5.3: The reduced Markov chain for $\rho \to 0$.

The root in the interval $(0, 1)$ of the function (5.19) is

$$z_0 = \frac{(\mu_c + \mu_w + \eta_c + \eta_w) - \sqrt{(\mu_c + \mu_w + \eta_c + \eta_w)^2 - 4\mu_c\mu_w}}{2\lambda},$$

since one other root is 1 and the third one is larger than 1. Hence, we get the following result (which becomes exact as $\rho \to 1$):

High utilization approximation. *The average file transmission delay in the on-the-spot mobile data offloading for a user with heavy traffic can be approximated by*

$$E[T] = \frac{1}{\mu - \lambda}\left(1 - \frac{(\mu_c - \lambda)(\mu_w - \lambda)}{\lambda(\eta_c + \eta_w)}\right) + \frac{z_0}{\lambda(\eta_c + \eta_w)(1 - z_0)}\left(\frac{\mu_w - \lambda}{\mu_w - \lambda z_0}\eta_w + \frac{\mu_c - \lambda}{\mu_c - \lambda z_0}\eta_c\right). \tag{5.20}$$

5.2.5 Moderate utilization approximation

Finally, we can get an approximation for moderate utilization regimes by interpolating function $f(z)$. We only state the result here. Note that unlike the low and high utilization approximations, which become tight in the limit, this is just a heuristic.

Moderate utilization approximation. *The average file transmission delay in the on-the-spot mobile data offloading for moderate traffic can be approximated by*

$$E[T] = \frac{1}{\mu - \lambda} + \frac{\mu_c(\mu_w - \lambda)\pi_{0,c} + (\mu_c - \lambda)(\mu_w(\pi_{0,w} - 1) + \lambda)}{\lambda(\eta_c + \eta_w)(\mu - \lambda)}, \tag{5.21}$$

where $\pi_{0,c}$ and $\pi_{0,w}$ are given by Eq.(5.12)-(5.13), and $z_0 = \frac{1}{\varepsilon}\frac{\mu_c\mu_w}{\eta_c\mu_w + \eta_w\mu_c - \lambda(\eta_c + \eta_w) + \mu_c\mu_w}$. ε is a fitting parameter that takes values in the range 1.4-1.6, and can be fine-tuned empirically for better results.

5.2.6 Generic file size distribution approximation

Our analysis so far considers exponentially distributed flow sizes. Yet, for some traffic types, heavy-tailed file sizes were reported [74]. Unfortunately, generalizing the above 2D chain analysis for generic files is rather hard, if not impossible. Nevertheless, we can use the M/G/1 Pollaczek-Khinchin (P-K) formula [20] as a guideline to introduce a similar "correction factor" related to smaller/higher file size variability.

Let c_v denote the coefficient of variation[4] for the file size distribution, and $E[T]$ and $E[S]$ denote the system and service time, respectively, for exponentially distributed packet sizes (as derived before). The average system time for the generic packet size distributions in the ordinary M/G/1 can be written through the delay of the M/M/1 system as[5]

$$E[T_g] = E[S] + E[T_Q] \cdot \frac{1 + c_v^2}{2}, \tag{5.22}$$

where $E[T_Q] = E[T] - E[S]$ is the average queueing time for the exponentially distributed packet sizes. After some very simple calculus steps performed on Eq.(5.22), we obtain the following approximation for generic file sizes.

Result 13. *The average file transmission delay in the on-the-spot mobile data offloading for generic file size distributions can be approximated by*

$$E[T_g] = \frac{1 - c_v^2}{2} E[S] + \frac{1 + c_v^2}{2} E[T]. \tag{5.23}$$

5.2.7 Offloading efficiency

Finally, an important parameter that can quantitatively characterize data offloading is the *offloading efficiency OE*, defined as the ratio of the amount of transmitted data through WiFi against the total amount of transmitted data. Higher offloading efficiency means better performance for both the client and operator. Knowing this parameter is especially important when it comes to calculating how much a user will have to pay, knowing that the charges for using Internet access are not the same for WiFi as are for cellular network.

Let's denote with ρ_w the percentage of time there is at least one file in the system while there is WiFi coverage. It holds that $\rho_w = \pi_w - \pi_{0,w} = G_w(1) - G_w(0)$. Similarly, $\rho_c = G_c(1) - G_c(0)$ denotes the percentage of time the system is busy while having no WiFi connectivity.

Next, we define t_w (t_c) as the total time (in long-run) during which data are sent through the WiFi (cellular) interface. Actually, it is the sum of the durations of the ON (OFF) periods when there was at least one file in the system. The amount of data transmitted over the WiFi network is $\mu_w t_w$, and over the cellular network it is $\mu_c t_c$. Since we define the offloading efficiency to be the percentage of data transmitted through the WiFi network, we have the following expression for it:

$$OE = \frac{\mu_w t_w}{\mu_w t_w + \mu_c t_c}. \tag{5.24}$$

Due to the ergodicity of our 2D Markov Chain (this is a necessary condition to have a stationary distribution), it is easy to see that

$$\frac{t_w}{t_c} = \frac{\rho_w}{\rho_c}. \tag{5.25}$$

[4] The coefficient of variation of a random variable X is defined as $c_v = \frac{\sqrt{E[X^2]-(E[X])^2}}{E[X]}$ [20].

[5] Eq.(5.22) represents the exact P-K formula for an M/G/1 system [20] with "ordinary" service rate (the average service rate does not change over time). We are using it here as an approximation for a queueing system with variable service rate. As will be shown in Section 5.4, the accuracy of this approximation result depends on the coefficient of variation of the packet size. For lower c_v, it is almost exact.

Replacing Eq.(5.25) into Eq.(5.24), we obtain the following result.

Result 14. *The offloading efficiency in an on-the-spot mobile data offloading system is*

$$OE = \frac{\mu_w}{\mu_w + \mu_c \frac{G_c(1)-G_c(0)}{G_w(1)-G_w(0)}}. \tag{5.26}$$

5.3 Analyzing Offloading in Heterogeneous Cellular Networks

In the previous section, we considered a simpler scenario with two networks (WiFi and Cellular) and respective rates. Nevertheless, in current and future networks different areas might be covered by different cellular network technologies (GPRS, EDGE, HSPA, LTE), and multiple "short-range" options might exist in addition to WiFi (e.g. femto- or other small-cell technologies). Coverage holes might also exist. Finally, even within the same technology (e.g. WiFi), the rate might be different across different access points. While such scenarios could still be emulated by the basic model, by "absorbing" these difference across ON and OFF periods into an average rate, as explained earlier, it would be interesting to see whether we can extend our basic model to better predict performance in more complex scenarios, where a mobile user might be switching between a number of different technologies and/or rates during a large time window.

5.3.1 The model

As earlier in Section 5.2.1, we again consider a mobile user moving between locations with different network characteristics. However, now there are M possible options a user can encounter, rather than just 2. To keep discussion simple, in the remainder we will assume that these different options correspond to different technologies, such as WiFi, 2.5G, 3G, 3G+, 4G, etc., or even no network at all. Note however that the analysis to follow can be also applied to a scenario where, e.g. we have a number of different rates a user could experience within the same technology. This just requires increasing M, and updating the transition rates between different states accordingly (e.g. to some long-term statistics available).

We will assume that whenever there is WiFi coverage, all the traffic will be transmitted through the corresponding AP. When the WiFi connectivity is lost, the traffic will be sent (received) through the cellular technology available. In case there are multiple options in terms of the network coverage, we assume that the user will switch to the technology that offers the highest rates. Nevertheless, this could be a policy matter that will not affect our analysis. As before, we will assume that network transitions do not cause any impairments to the flow. At the end of this section, we briefly discuss how the generic model could be extended to include also possible interruptions caused from switching to a different technology.

We model the network availability with the multilevel scheme as shown in Fig. 5.4. The duration of each period is exponentially distributed with rate $\eta_i, i = 1, \ldots, M$. Each period T_i corresponds to the time duration during which the mobile user is communicating through the same access network technology without interruption. We call these periods *levels* or *phases.* The period durations of any level are mutually independent, and at the same time independent of the durations of periods of other levels. The data transmission rates during periods with different connectivity are denoted as $\mu_i, i = 1, \ldots, M$. As before, the traffic arrival process is Poisson and file sizes are exponentially distributed. The scheduling discipline is the same as before (FCFS).

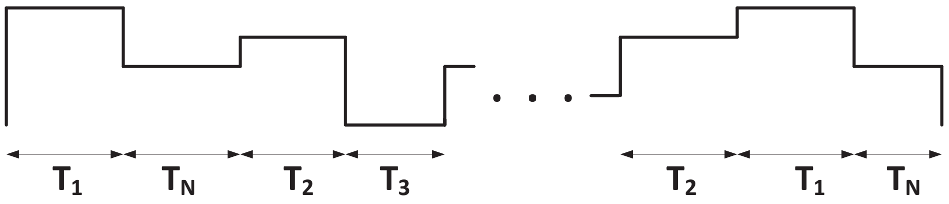

Figure 5.4: The multi-network availability model.

5.3.2 Analysis

In this subsection we derive the expression for the average file delay in a multilevel on-the-spot offloading system. Based on the assumptions we have made, our system can be modeled with a 2D Markov chain that is bounded in one dimension (the dimension that represents the number of levels). This is shown in Fig. 5.5. The possible level transitions are shown only for the state $(0,1)$ and partially for $(j,2)$ to avoid making the figure look more complex. $\pi_{k,i}$ denotes the stationary probability of being at level i (where i corresponds to areas served by a given technology with rate μ_i), and having k files in the buffer. There are a number of possible transitions now corresponding to the following events:

new arrival: From any state, the chain moves to the right (horizontally) with rate λ.

flow finishes transmission: From any state $\{k,i\}$ $(k>0)$, the chain moves to the left (horizontally) with rate μ_i.

change in the network connectivity: The chain moves (vertically) from level i to another level j (transition to all the levels possible with no exception) with rate $\eta_{i,j}, i \neq j$.

If we denote with η_i the (total) rate at which a user leaves an area covered by technology (or rate) i, it holds that $\eta_i = \sum\limits_{j=1}^{M} \eta_{i,j}$, for $i \neq j$.

We start by writing the balance equations for this Markov chain as

$$(\lambda + \eta_i)\,\pi_{0,i} = \mu_i \pi_{1,i} + \sum_{j=1}^{M} \eta_{j,i}\pi_{0,j}, \tag{5.27}$$

for $i = 1,\ldots,M$, $k=0$, and

$$(\lambda + \mu_i + \eta_i)\,\pi_{k,i} = \lambda \pi_{k-1,i} + \mu_i \pi_{k+1,i} + \sum_{j=1}^{M} \eta_{j,i}\pi_{k,j}, \tag{5.28}$$

for $i = 1,\ldots,M$ and $k>0$. Summing Eq.(5.27) and Eq.(5.28) multiplied by z^k, and then summing up over all k, we get

$$\lambda\sum_{k=0}^{\infty}\pi_{k,i}z^k + \mu_i\sum_{k=1}^{\infty}\pi_{k,i}z^k + \sum_{k=0}^{\infty}\pi_{k,i}z^k\sum_{j=1}^{M}\eta_{i,j} = \lambda\sum_{k=1}^{\infty}\pi_{k-1,i}z^k + \mu_i\sum_{k=1}^{\infty}\pi_{k,i}z^{k-1} + \sum_{j=1}^{M}\eta_{j,i}\sum_{k=0}^{\infty}\pi_{k,j}z^k. \tag{5.29}$$

We define the PGF for each level as

$$G_i(z) = \sum_{k=0}^{\infty}\pi_{k,i}z^k,\ |z| \leq 1,\ i = 1,\ldots,M. \tag{5.30}$$

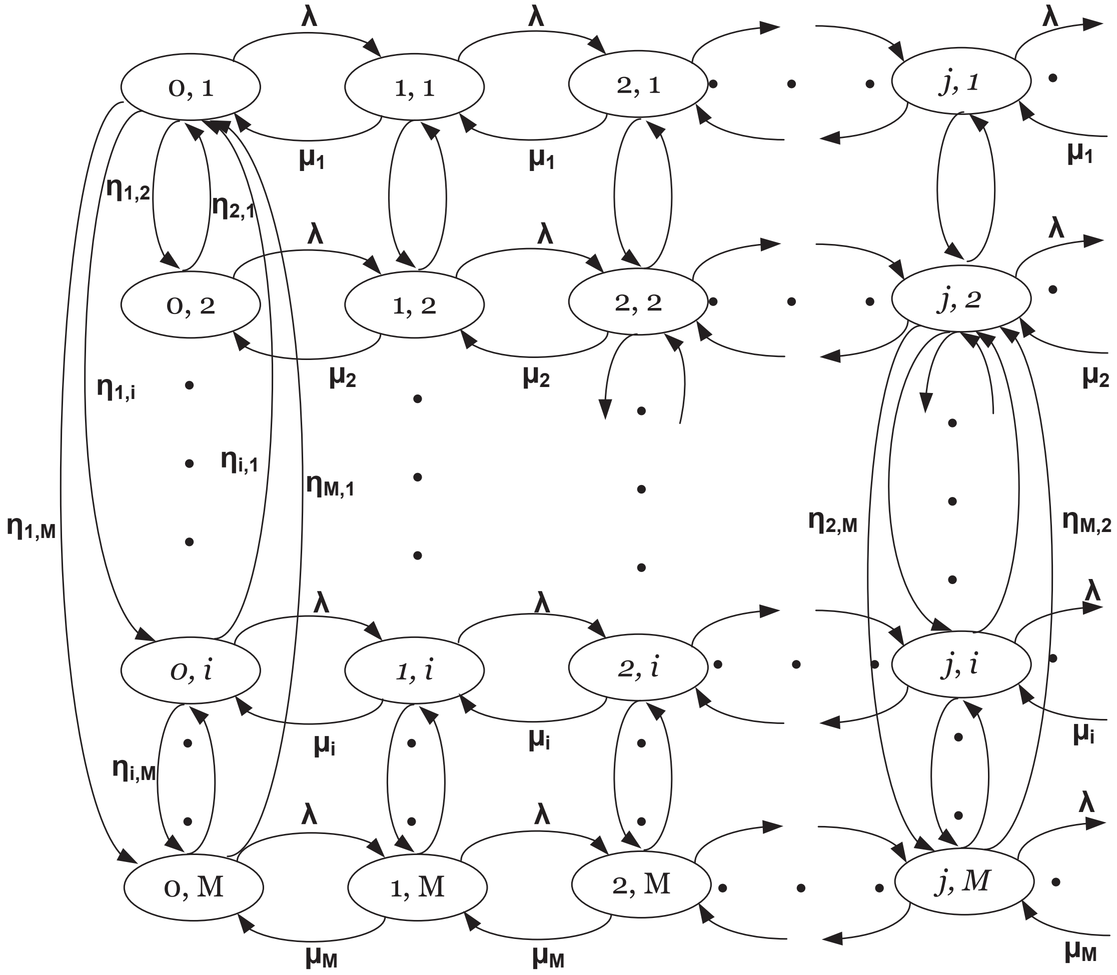

Figure 5.5: The 2D Markov chain for multilevel on-the-spot mobile data offloading model.

Eq.(5.29) is now transformed to

$$\lambda G_i(z) + \mu_i \left[G_i(z) - \pi_{0,i}\right] + G_i(z) \sum_{j=1}^{M} \eta_{i,j} = \lambda z G_i(z) + \frac{\mu_i}{z} \left[G_i(z) - \pi_{0,i}\right] + \sum_{j=1}^{M} \eta_{j,i} G_j(z). \quad (5.31)$$

After performing some algebra we have

$$\left[\lambda z(1-z) + \mu_i(z-1) + \eta_i z\right] G_i(z) - \sum_{j=1}^{M} \eta_{j,i} z G_j(z) = \mu_i(z-1)\pi_{0,i}, i = 1, \dots, M. \quad (5.32)$$

In Eq.(5.32), after introducing the substitution

$$f_i(z) = \lambda z(1-z) - \mu_i(1-z) + \eta_i z, \quad (5.33)$$

we obtain the following equation

$$\mathbf{F}(z)\mathbf{g}(z) = (z-1)\boldsymbol{\theta}, \quad (5.34)$$

where

$$\mathbf{F}(z) = \begin{bmatrix} f_1(z) & -\eta_{2,1}z & -\eta_{3,1}z & \dots & -\eta_{M,1}z \\ -\eta_{1,2}z & f_2(z) & -\eta_{3,2}z & \dots & -\eta_{M,2}z \\ \vdots & \vdots & \vdots & \dots & \vdots \\ -\eta_{1,M}z & -\eta_{2,M}z & -\eta_{3,M}z & \dots & f_M(z) \end{bmatrix},$$

$$\mathbf{g}(z) = \begin{bmatrix} G_1(z) \\ G_2(z) \\ \vdots \\ G_M(z) \end{bmatrix}, \boldsymbol{\theta} = \begin{bmatrix} \mu_1\pi_{0,1} \\ \mu_2\pi_{0,2} \\ \vdots \\ \mu_M\pi_{0,M} \end{bmatrix}.$$

Applying Cramer's rule to Eq.(5.34), we obtain

$$\left|\mathbf{F}(z)\right| G_i(z) = \left|\mathbf{F}_i(z)\right| (z-1). \quad (5.35)$$

$|\mathbf{F}_i(z)|$ is the determinant obtained after replacing the ith column of $|\mathbf{F}(z)|$ with $\boldsymbol{\theta}$. As can be observed from Eq.(5.33), at point $z = 1$, $f_1(1) = \eta_1$. Also, at this same point the sum of the elements in rows 2 to M of the first column $(-\eta_{1,2}z - \eta_{1,3}z - \dots - \eta_{1,M}z)$ represents the sum of transition rates out of state 1 multiplied by -1. If we substract row 1 from the sum of the other rows, we have 0 at the element $\{1,1\}$ of the determinant $|\mathbf{F}_i(z)|$. Similar conclusions can be drawn for the other elements of the first row. Hence, we can obtain an equivalent determinant $|\mathbf{F}(z)|$ with all the elements of the first row equal to 0. So, $z = 1$ is one root of this determinant. Hence, we can write

$$\left|\mathbf{F}(z)\right| = (z-1)Q(z). \quad (5.36)$$

Replacing Eq.(5.36) into Eq.(5.35) we get

$$Q(z)G_i(z) = \left|\mathbf{F}_i(z)\right|. \quad (5.37)$$

In order to find the partial probability generating functions $G_i(z)$, first we need to find the zero probabilities $\pi_{0,1}, \pi_{0,2}, \ldots, \pi_{0,M}$. To do this, we proceed in the following way. First, we find the roots of $Q(z)$. Since our system is of order $M > 2$, these solutions can be obtained only numerically. The polynomial $Q(z)$ is of degree $2M-1$. However, only $M-1$ of its roots lie in the interval $(0,1)$ (which is our interval of interest)[6]. We denote these roots as $z_1, \ldots, z_{M-1}$. Since $G_i(z) \neq 0$ (all the probabilities $p_{k,i}$ are positive), then from Eq.(5.37), we have that $|\mathbf{F}_i(z_j)| = 0, i = 1, \ldots, M, j = 1, \ldots, M-1$. However, from Eq.(5.37) we can observe that, for each z_j, and any pair $1 \leq i, l \leq M$, $\frac{|\mathbf{F}_i(z_j)|}{|\mathbf{F}_l(z_j)|} = const$. This means that for each z_j we have M homogeneous linear equations that differ from each other only by a constant factor. Hence, $|\mathbf{F}_i(z_j)| = 0$ gives only one independent equation for each root z_j. Given that there are $M-1$ different roots z_j, it turns out that there are in total $M-1$ independent equations. Since we have M unknown probabilities $\pi_{0,1}, \pi_{0,2}, \ldots, \pi_{0,M}$, and only $M-1$ equations, we cannot obtain unique solutions for these probabilities. So, we need another condition that relates these zero probabilities, and that is independent of the other $M-1$ equations.

Let's consider the vertical cut between states k and $k+1$. The balance equation through this cut is

$$\lambda(\pi_{k,1} + \pi_{k,2} + \ldots + \pi_{k,M}) = \mu_1 \pi_{k+1,1} + \ldots + \mu_M \pi_{k+1,M}. \tag{5.38}$$

Summing over all k yields

$$\lambda \sum_{i=1}^{M} \pi_i = \mu_1(\pi_1 - \pi_{0,1}) + \ldots + \mu_M(\pi_M - \pi_{0,M}). \tag{5.39}$$

$\pi_i = \sum_{k=0}^{\infty} \pi_{k,i}$ denotes the percentage of time the system is in level i. Eq.(5.39) can be rewritten as

$$\mu - \lambda = \sum_{i=1}^{M} \mu_i \pi_{0,i}, \tag{5.40}$$

where $\mu = \sum_{i=1}^{M} \mu_i \pi_i$ is the average service rate of the system. Eq.(5.40) is the Mth equation of the system we need to solve in order to get the zero probabilities. However, we do need to determine the probabilities π_i first.

We can find π_i by following a standard embedded MC approach for the (collapsed) chain with only M states (corresponding to the M levels). If we define $q_{i,j}$, the transition probabilities in the embedded chain, as $q_{i,j} = \frac{\eta_{i,j}}{\eta_i}$, then

$$\pi_i = \frac{\frac{r_i}{\eta_i}}{\sum_{i=1}^{M} \frac{r_i}{\eta_i}}, \tag{5.41}$$

where r_i are the solutions to the global balance equations for the embedded Discrete Time Markov Chain (DTMC): $\sum_{i=1}^{M} r_i = 1$, and $r_j = \sum_{i=1}^{M} r_i q_{i,j}$.

[6] The proof to this claim is rather long and complicated, and due to space limitations we do not show it here. It was proven by Mitrani and Itzhak in [75], pp. 632-634.

Replacing Eq.(5.41) into Eq.(5.40), we have the Mth equation of our system. Now, solving that system we get all the zero probabilities. The partial PGFs are found from Eq.(5.37) as

$$G_i(z) = \frac{|\mathbf{F}_i(z)|}{Q(z)}, i = 1, \ldots, M. \tag{5.42}$$

The average number of files in the system is

$$E[N] = \sum_{i=1}^{M} G_i^{'}(1). \tag{5.43}$$

Using Little's law $E[N] = \lambda E[T]$, we get the following result:

Result 15. *The average file delay in a multilevel on-the-spot offloading system is given by*

$$E[T] = \frac{1}{\lambda} \sum_{i=1}^{M} \left(\frac{|\mathbf{F}_i(z)|}{Q(z)} \right)'_{z=1}. \tag{5.44}$$

We conclude this section by discussing how the M-level model could be extended to handle interruptions of a flow, due to switching from one technology to another. If we assume that a flow is delayed a bit due to this interruption, but then resumes over the new network, this could be captured, for example, by introducing $2M$ levels, instead of M that we have now. Each of the current levels would have a corresponding quasi-level, into which the system would switch first. These levels correspond to the time needed to resume transmission. After staying for some time in one of those levels, the state of the system would move to a "real" level, in which the communication would be reestablished. While being in the quasi-level state, the data rate is 0, and from it the system can only move to the corresponding level, attached to the quasi-level state. The analysis would then be the same as that described before.

5.4 Simulation results

5.4.1 Basic model validation

In this section we will validate our theory against simulations for a wide range of traffic patterns, different values of file sizes and different average WiFi availability periods and availability ratios. We define the WiFi availability ratio as $AR = \frac{E[T_{ON}]}{E[T_{ON}]+E[T_{OFF}]} = \frac{\eta_c}{\eta_w+\eta_c}$. Unless otherwise stated, the durations of WiFi availability and unavailability periods will be drawn from independent exponential distributions with rates η_w and η_c, respectively. We mainly focus on two scenarios, related to the user's mobility. The first one considers pedestrian users with data taken from [12]. Measurements in [12] report that the average duration of WiFi availability period is 122 min, while the average duration with only cellular network coverage is 41 min (we use these values to tune η_w and η_c). The availability ratio reported is 75%. The second scenario corresponds to vehicular users, related to the measurement study of [68]. An availability ratio of 11% has been reported in [68]. For more details about the measurements we refer the interested reader to [12] and [68]. Finally, unless otherwise stated, file/flow sizes are exponentially distributed, and file arrivals at the mobile user is a Poisson process with rate λ.

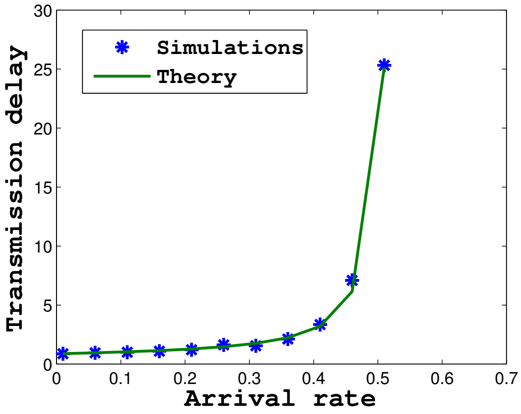

Figure 5.6: Pedestrian user.

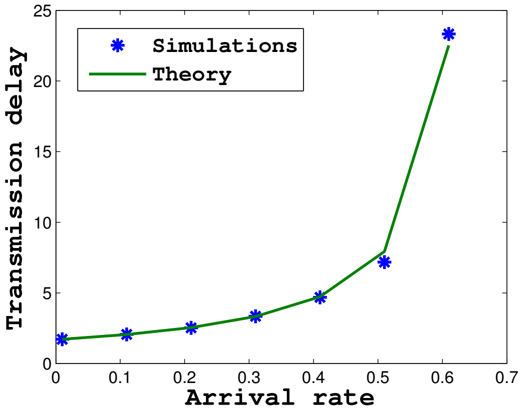

Figure 5.7: Vehicular user.

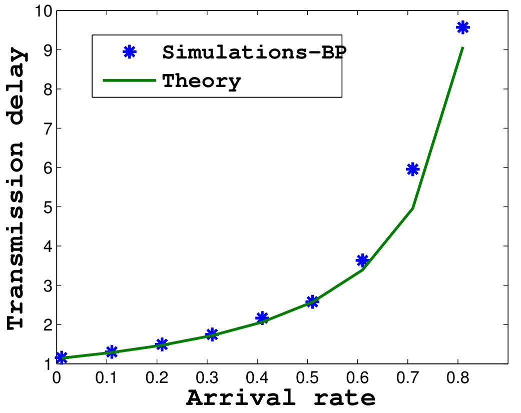

Figure 5.8: BP vehicular periods.

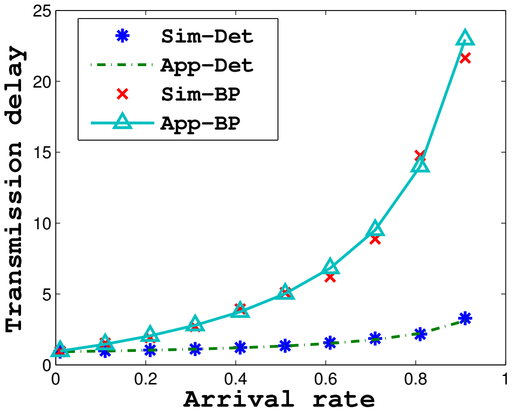

Figure 5.9: Generic flow sizes.

5.4.1.1 Validation of the main delay result

We first validate our model and 2-level result (Eq.(5.16)) against simulations for the two mobility scenarios mentioned (pedestrian and vehicular). The data rate for WiFi is assumed to be 2 Mbps (this is close to the average data rate obtained from measurements with real traces in [76]), and we assume that the cellular network is 3G, with rate 500 kbps. The mean flow size is assumed to be 125 kB[7].

Fig. 5.6 shows the average file transmission delay (i.e. queueing + transmission) for the pedestrian scenario, for different arrival rates. The range of arrival rates shown corresponds to a server utilization of 0-0.9. We can observe, in Fig. 5.6, that there is a good match between theory and simulations. Furthermore, the average file transmission delay is increased with the arrival rate, as expected, due to queueing effects. Fig. 5.7 further illustrates the average file transmission delay for the vehicular scenario. We can observe there that the average transmission time is larger than in Fig. 5.6. This is reasonable, due to the lower WiFi availability, resulting in most

[7] This value is normalized for the arrival rates considered, to correspond to the traffic intensities reported in [68]. We have also considered other values with similar conclusions drawn.

of the traffic being transmitted through the slower cellular network interface. Once more, we can observe a good match between theory and simulations.

5.4.1.2 Validation against non-exponential ON-OFF periods

In the previous scenarios, we have used realistic values for the transmission rates and WiFi availabilities, but we have assumed exponential distributions for ON and OFF periods, according to our model. While the actual distributions are subject to the user mobility pattern, a topic of intense research recently, initial measurement studies ([12,68]) suggest these distributions to be "heavy-tailed". It is thus interesting to consider how our model's predictions fare in this (usually difficult) case. To this end, we consider a scenario with "heavy-tailed" ON/OFF distributions (Bounded Pareto-BP). Due to space limitations, we focus on the vehicular scenario. The shape parameters for BP ON and OFF periods are $\alpha = 0.59$ and $\alpha = 0.64$, respectively. We consider a cellular rate of 800 kbps. We change the value of the data rate to see that our analysis holds for other values as well. Fig. 5.8 compares the average file delay for this scenario against our theoretical prediction. Interestingly, our theory still offers a reasonable prediction accuracy, despite the considerably higher variability of ON/OFF periods in this scenario[8]. While we cannot claim this to be a generic conclusion for any distribution and values, the results underline the utility of our model in practice.

5.4.1.3 Validation of non-exp flow sizes

To conclude our validation, we finally drop the exponential flow assumption as well, and test our generic file size results of Eq.(5.23). Fig. 5.9 compares analytical and simulation results for deterministic, and Bounded Pareto distributed files sizes (shape parameter $\alpha = 1.2$ and $c_v = 3$). Mean file size is in both cases 125KB, and the rest of the parameters correspond to the vehicular scenario (exp. ON and OFF periods). We observe that higher size variability further increases delay, as expected. Somewhat more surprisingly, the observed accuracy in both cases is still significant, despite the heuristic nature of the approximation and the complexity of the queueing system.

5.4.1.4 Validation of approximations

Having validated the main result of Eq.(5.16) we now proceed to validate the various simpler approximations we have proposed in Section 5.2. We begin with the low utilization approximation of Section 5.2.3 with $AR = 0.75$ (similar accuracy levels have been obtained with other values). Fig. 5.10 shows the flow delay for low arrival rates in the range $0.01 - 0.1$, which correspond to a maximum utilization of around 0.1. We can observe that the low utilization approximation provides a good match with the generic result and simulations. As λ increases, the difference between the approximated result and the actual value increases. For $\rho = 0.1$, the approximation error is around 5%. This is reasonable, as we have strictly assumed that there might be at most one file present in the system.

We next consider the high utilization regime and respective approximation (Eq.(5.20)). We consider utilization values of 0.8-0.95. Fig. 5.11 shows the delay for high values of λ, and

[8] This has been the case with additional distributions and values we have tried. We have also observed that the error generally increases (decreases) when the difference between WiFi and cellular rates increases (decreases).

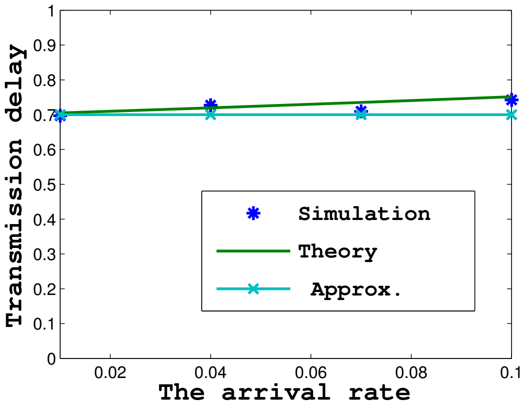

Figure 5.10: The low utilization approx.

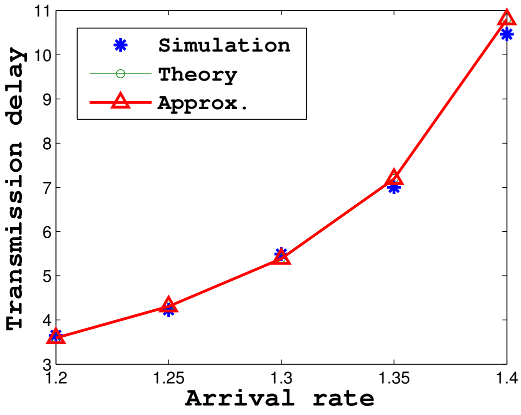

Figure 5.11: The high utilization approx.

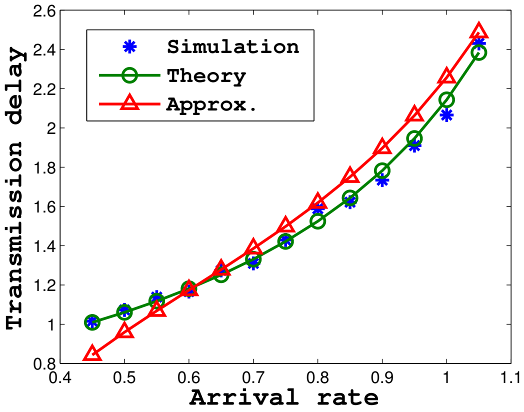

Figure 5.12: The approx. for AR=0.5.

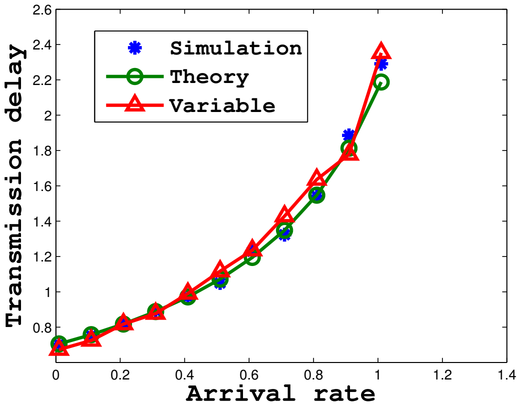

Figure 5.13: Variable WiFi rates.

$AR = 0.5$ (we have again tried different values). We can see there that our approximation is very close to the actual delay and should become exact as ρ goes to 1.

Finally, we consider the approximation result (Eq.(5.21)) for moderate utilization values, in the range $0.3 - 0.7$ ($AR = 0.5$). Fig. 5.12 compares theory and simulations for the delay in this intermediate utilization regime. For moderate ρ, the value of the coefficient ε is 1.5. It is chosen empirically. Although this approximation is heuristic, and does not become exact for any utilization value (unlike the cases of the low/high utilization approximations), we can see that the accuracy is still satisfactory and improves for higher ρ.

5.4.2 Limitations of the 2-level model

So far, we have been assuming constant WiFi data rate in all the regions with WiFi coverage. While in theory it enables analytical tractability, this assumption is rather unrealistic, since the actual rate experienced in different APs will depend on AP load, distance, backhaul technology, etc. Therefore, it is particularly interesting to consider scenarios where the WiFi rate might be different at each connected AP. Specifically, we simulate a scenario where the average data rate over all APs is again 2 Mbps, but the actual rate for each ON (WiFi) period is selected uniformly in the interval 1-3 Mbps. The other parameters remain unchanged. In Fig. 5.13, we compare simulation results for this scenario against our theoretical result (which assumes a constant WiFi rate of 2 Mbps in every AP). From Fig. 5.13 it is evident that WiFi rate variability does not affect significantly the performance, making thus our results applicable in this case as well, despite the variable nature of the WiFi rate.

As we saw in Fig. 5.13, our model can predict with an excellent accuracy the delay even in a system in which the WiFi rate is not constant. Out of that, one might infer that our two-level model is sufficient to provide a high-accuracy approximate analysis for a network technology with any discrete number of data rates. It would be enough to lump all the levels into one level (phase) with a data rate equal to the average data rate of all the other levels. However, this turns out to be incorrect in the general case. The reason why this method gave good match in Fig. 5.13 is because the rates were chosen from a uniform distribution, i.e. the rates were close to each other. We investigate the effect of highly variable rates below.

We consider first the data rates to be drawn from an exponential distribution with the same average as in Fig. 5.13. The same holds for all the other system related parameters as well. Fig. 5.14 shows the average delay for this scenario. As can be seen from there, our theory cannot predict the delay correctly anymore. The discrepancy is even more pronounced when there is higher variability in the data rates. In the same plot we show the delay for data rates drawn from a Pareto distribution with shape parameter $\alpha = 1.2$, and the same average. The delay now is much higher, exceeding 5× the predicted result by our theory.

Having established that the two level model fails in predicting the delay for variable rates that show a tendency of being quite dispersed from the mean, we move on with investigating the effect of availability ratio to the delay. Fig. 5.15 shows the average delay vs. arrival rate for $AR = 0.75$ and uniform, exponential and Pareto distributed data rates. The other parameters are the same as for Fig. 5.14. It is also shown the delay curve from our theory with constant rates. Here also, our theory can give accurate result when data rates underlie uniform distribution, but fails when it comes to the higher variability data rates.

Finally, we consider the effect of larger difference between the WiFi and cellular rates. For that purpose we consider a scenario with cellular rate of 0.5 Mbps, instead of 1 Mbps, and with the same average WiFi rate (2 Mbps). The availability ratio is 0.5. Fig. 5.16 illustrates the average delay. In this case also, the same conclusions hold as before.

So, we can say that the accuracy of our model holds in scenarios where the data rates are relatively close to each other, and it does not depend heavily on the availability ratio nor on the cellular rate. When it comes to data rates that are subject to higher variability, we need the M-level model of Section 5.3.

To further enhance our claims about the necessity of using the M-level model, we consider scenarios with multiple access technologies (WiFi, 3G, HSPA, LTE). The corresponding parameter values are given in Table 5.2. The values taken belong to the range of intervals given in [3, 4]. The other system parameters are the same as in the previous considered scenarios. We lump the cellular network levels into one single level with average duration equal to the sum of their individual average durations (15 s), and average data rate equal to the weighted average of the corresponding rates (3.5 Mbps). Then, we use our 2-level model to find the average delay. Fig. 5.17 illustrates the average delay vs. the arrival rate. On the same plot, we show the actual simulated delay. As can be seen from Fig. 5.17, the 2-level model cannot capture the case with multiple heterogeneous networks, as the prediction is very far from the actual delay. Hence, for

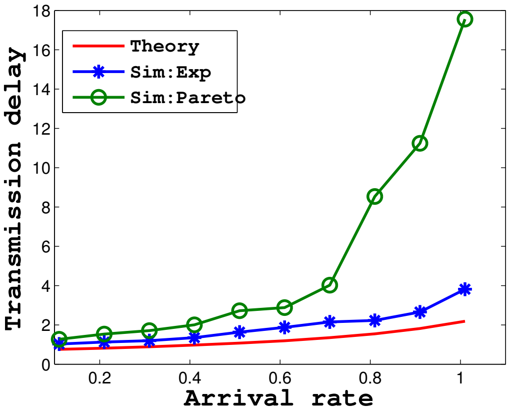

Figure 5.14: Variable rates for $AR = 0.5$.

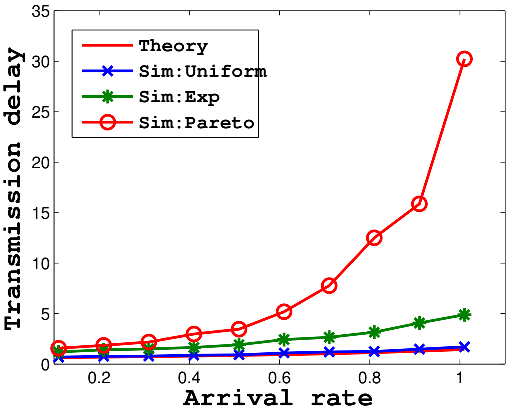

Figure 5.15: Variable rate with $AR = 0.75$.

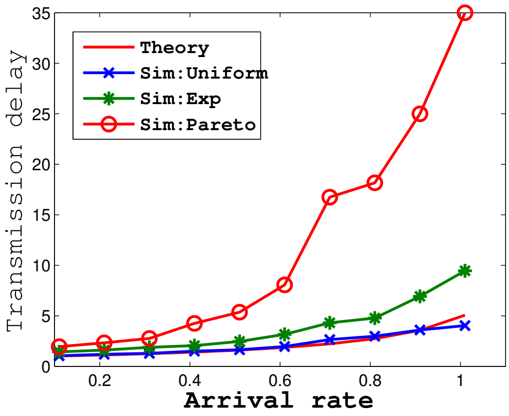

Figure 5.16: Variable rate case with lower cellular rate.

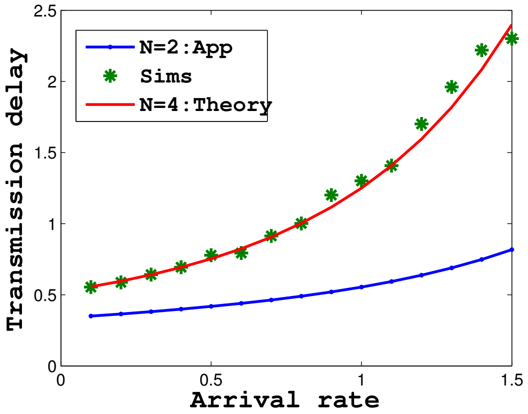

Figure 5.17: The approximation with the 2 level model.

such cases we need the M-level model.

5.4.3 M-level model validation

Next, we consider scenarios with multiple access technologies (WiFi, 3G, HSPA, LTE), or even without network coverage at all. Namely, there are operators that might offer 4G coverage only in some regions, while in the others they offer only 3G. There might also exist regions with little or no coverage at all (in sparse populated areas). In the following we will see how our multilevel theory of Section 5.3 will cope with the actual (simulated) scenarios. Unless otherwise stated, the data rates and average durations are given in Table 5.2.

First, we focus on the scenario when there are 3 possible network choices: WiFi, 3G and LTE. The policy here is that WiFi is the network with absolute priority. When there is no WiFi coverage, LTE has priority over 3G. We assume that there is always 3G network availability. There is an equal probability to move to any other access technology after leaving the current network. Since there are only 3 possible levels, this probability is equal to 0.5. Flows are

Table 5.2: The parameters for different access technologies [3, 4].

Technology	Data rate	Average duration
WiFi	2 Mbps	10 s
3G	1 Mbps	3 s
HSPA	1.5 Mbps	5 s
LTE	10 Mbps	5 s
No coverage	0	2 s

exponentially distributed with average size of 125 kB, and the arrival process is Poisson. The availability ratio of the WiFi network (Eq.(5.41)) is found to be 50%.

Fig.5.18 shows the average file delay vs. the arrival rate for this system. As can be seen, our theory matches with simulations. As expected, the delay increases with increasing the traffic arrival rate, due to the queueing effect.

The second scenario shown in Fig. 5.18 corresponds to $M = 4$ possible levels: WiFi, 3G, HSPA, and LTE. The parameters are given in Table 5.2. The probability of moving to any specific level is 1/3 now. The availability ratio now turns out to be 43.5%. There is a nice fit with theory again. The average delay is a bit higher compared to the previous example, since the HSPA data rate is lower compared to WiFi, and the other networks' characteristics are the same.

We also consider the possibility of not having network coverage at all. Now, we have to consider 5 levels (Fig. 5.18). Given that the other 4 levels have the same parameters as above, for the no network availability we choose the average duration to be 2s. The probability of encountering a specific level after leaving the one in use is 0.25. The availability ratio is 40%. Again, there is a match between theory and simulations that shows that our theory is correct. As expected the delay is larger, because there are some time periods when there is no connectivity at all.

Finally, we consider scenarios with lower WiFi availability ratio, and higher LTE duration periods than before (10 s). The availability ratios for $M = 3, 4, 5$ are 40%, 36% and 33%, respectively. The other parameters are exactly the same as before. Fig. 5.19 illustrates the average delay. As can be seen, the delays are much lower now. This comes from the fact that there is a lower degree of WiFi connectivity, and higher degree of LTE coverage. Since the LTE data rates are much higher, the delay is significantly reduced. The delay reduction can exceed 20%. The only advantage in using WiFi offloading under these circumstances lies in the lower prices WiFi operators offer, as opposed to LTE charges.

5.4.4 Non-exponential assumptions

So far, we have validated our model for the exponentially distributed durations of the different levels, as well as for exponentially distributed flow sizes. Next, we drop these assumptions and see how our theory behaves under these more general conditions. First, we keep the exponential assumption on file sizes, and consider heavy-tailed distributions for the durations of the different levels. There are $M = 4$ possible levels. The average durations of the corresponding phases are identical to those of Fig. 5.18, only that now they are Bounded Pareto with shape parameter $\alpha = 1.2$. Fig. 5.20 shows the average file delay. Surprisingly enough, our theory that is valid only for exponentially distributed periods, is able to predict the delay even for heavy-tailed

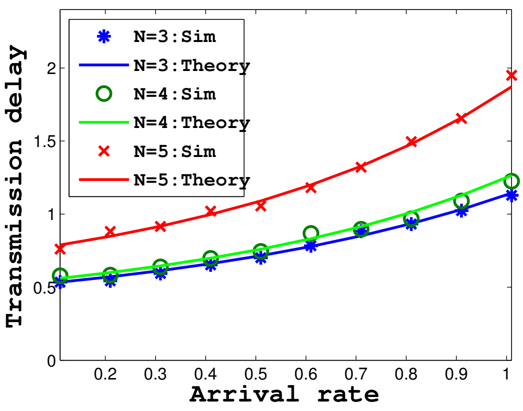

Figure 5.18: The transmission delay.

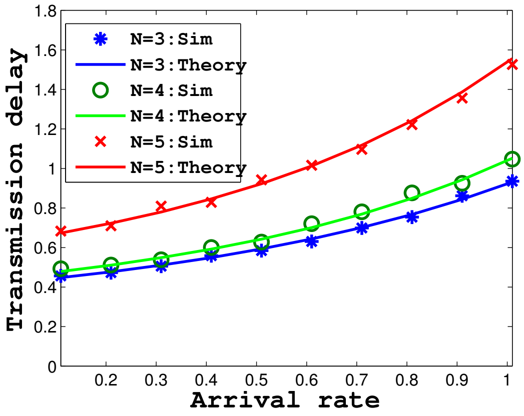

Figure 5.19: The transmission delay.

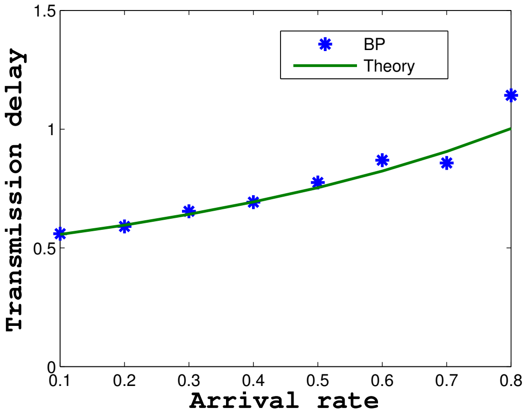

Figure 5.20: Bounded Pareto periods.

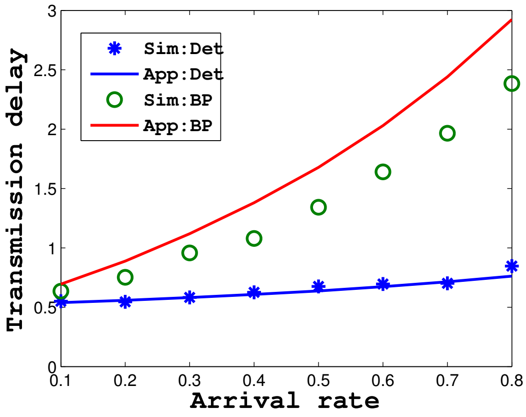

Figure 5.21: Generic flow sizes.

distributions with a remarkable accuracy.

Finally, we drop exponential assumptions for both level durations and flow sizes, and see how our generic flow size distribution approximation (Eq.(5.23)) behaves. We keep other parameters unchanged. The phase durations are Bounded Pareto with identical shape parameter as before ($\alpha = 1.2$). We consider two scenarios in terms of the distribution of flow sizes. While in the first one, all the flows have constant size, in the second one the flows have sizes that are drawn from a Bounded Pareto distribution with parameters $L = 0.24, H = 93, \alpha = 1.2$. It should be mentioned that the average flow size remains unchanged. Fig. 5.21 illustrates the delay for both scenarios. As can be seen, our proposed approximation (although heuristic) can predict the average delay quite satisfactorily, despite the very complex system we are dealing with. This increases the usefulness of our model. Another outcome of the model is that the delay, as in all other queueing systems, is higher for packets with higher variability.

5.4.5 Offloading Gains

We have so far established that our analytical model offers considerable accuracy for scenarios commonly encountered in practice. In this last part, we will thus use our model to acquire some initial insight as to the actual offloading gains expected in different scenarios. The operator's main gain is some relief from heavy traffic loads leading to congestion. The gains for the users are the lower prices usually offered for traffic migrated to WiFi, as well as the potential higher data rates of WiFi connectivity. There are also reported energy benefits associated [77], but we do not consider them here. Specifically, we will investigate the actual gains from data offloading, in terms of average transmission delay (related to user performance) and offloading efficiency (% of total traffic actually sent over WiFi - of interest to both the operator and the user). We consider two key parameters of interest that can affect these metrics: availability ratio and WiFi/cellular rate difference.

We first consider how transmission delay changes as a function of availability ratio, for different traffic intensities: very sparse, relatively sparse ($\rho = 0.15$) and medium ($\approx 40\%$). In the last scenario, however, when the user will be in zones in which it can connect only to lower rate access technologies the intensity would be much higher. Fig. 5.22 shows the average delay vs. AR for those traffic intensities. We can observe that the delay decreases as WiFi availability increases. More data are transmitted through the WiFi network, and hence the delay is lower since we have assumed that, on average, WiFi delivers better rates. A more interesting observation is that the delay improvement with higher WiFi availability values, is considerably more sharp, when the traffic load is higher. While for an arrival rate of $\lambda = 0.01$ the delay difference between the highest and the lowest availability ratios is less than 40%, this value exceeds $2\times$ for medium arrival rates. This seems to imply that denser WiFi deployments do not offer significant performance gains to users in low loaded regions, despite the higher rates offered, but could have a major impact on user experience, in heavily loaded areas.

As mentioned in Section 5.2, offloading efficiency is a very important quantity in characterizing mobile data offloading. Also, one might expect offloading efficiency to simply increase linearly with the availability ratio (i.e. % of data offloaded = % of time with WiFi connectivity). As it turns out, this is not the case. To better understand what affects this metric, we consider the impact of different cellular rates as well as different AR on the offloading efficiency. For the WiFi network we take the data rate to be 2 Mbps, and for the cellular we consider rates of 0.3 Mbps, 0.5 Mbps and 1 Mbps. Fig.5.23 illustrates the offloading efficiency vs. availability ratio for a moderate arrival rate of $\lambda = 0.3$. For comparison purposes we also depict the line $x = y$ (Offloading efficiency = AR). First, as expected, we can observe that offloading efficiency increases with availability ratio, in all scenarios. However, this increase is not linear. More interestingly, the actual offloading efficiencies are always higher than the respective availability ratio, and increase as the difference between the WiFi and the cellular rate increases. For $AR = 0.4$, 75% of the data are offloaded to WiFi when the ratio is 6.67 compared to 50% for a ratio of 2. The reason for this is that, due to the lower cellular rates, traffic arriving during the cellular (only) availability period ends up being transmitted during the next WiFi period due to queueing delays. This effect becomes more pronounced as the rate difference increases. Also, although not shown here, the respective offloading efficiency increases even further as traffic loads increase. Summarizing, these findings are particularly interesting to operators (and users), as they imply that high offloading efficiencies can be achieved for loaded regions, without necessarily providing almost full coverage with WiFi APs.

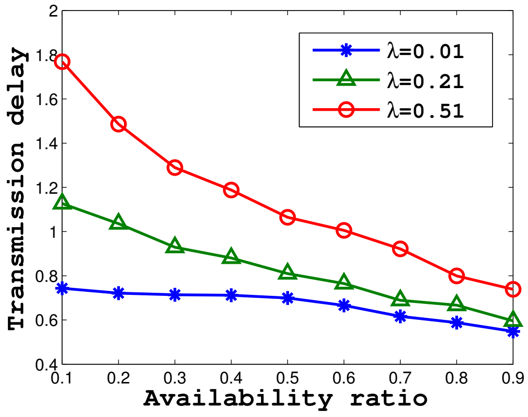

Figure 5.22: Different traffic rates.

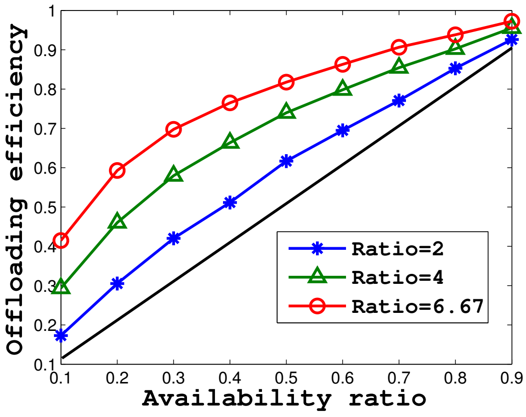

Figure 5.23: Offloading efficiencies.

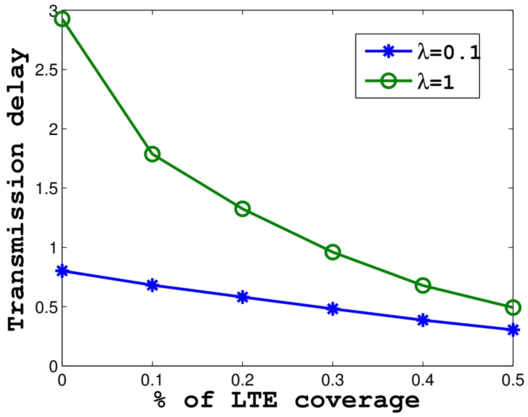

Figure 5.24: The design problem with increasing LTE coverage.

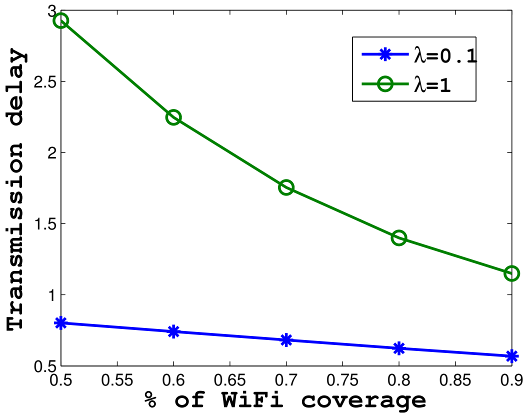

Figure 5.25: The design problem with increasing WiFi coverage.

Finally, let us consider the impact of installing new LTE base stations and WiFi access points on the user experience. At first, we assume that there is a coverage of 50% (the same order as in [12]) with WiFi APs, and that there is no LTE. Next, in the regions with no WiFi coverage, the 3G BTSs are being replaced successively with LTE BTSs. We consider the impact of the deployment of LTE base stations on the sparse and medium-to-high traffic intensities ($\lambda = 0.1$ and $\lambda = 1$). The WiFi data rate is 2 Mbps, while for the 3G and LTE the data rate is 1 Mbps and 10 Mbps, respectively. Fig. 5.24 depicts the dependency of the average file delay on the % of the LTE coverage (not covered by WiFi AP) which have replaced the 3G base stations. On the x-axis the value 0 denotes that there is 50% WiFi coverage, and the regions with no WiFi APs are covered with 3G base stations. The value of 50% on the x-axis refers to the complete replacement of 3G base stations with LTE BTSs (at least in the regions with no WiFi). As can be seen from Fig. 5.24 when it comes to zones with sparse traffic, increasing the number of LTE base stations does not necessarily improve too much the performance of the mobile users. For example, if a mobile operator decides to completely replace the 3G base stations with LTE base stations, the delay will be reduced by less than 2×. Since the deployment and maintenance of the LTE base stations implies an increased cost for the mobile operator, it is not economical to upgrade the network to 4G in the regions with sparse traffic. As opposed to this, when it comes to zones with a large number and very active users, such as city centers, university campuses etc., the full deployment of LTE will reduce the delay for mobile users up to 6× according to our scenario (see Fig. 5.24). Hence, in such regions it is beneficial for both the mobile operator and the users to upgrade the base stations.

Contrary to the previous case, now we decide to deploy additional APs and keep the actual 3G base stations. The other parameters remain the same as before. Fig. 5.25 shows the average delay vs. the WiFi availability ratio for the two traffic intensities (sparse and medium-to-high). We assume that at the beginning there is a coverage of 50% with WiFi APs. If we compare Fig. 5.25 and Fig. 5.24, for sparse traffic, we can notice that the difference in the delay is very low (less than 20%). On the other hand, for dense traffic if we deploy LTE base stations instead of WiFi AP, the delay will be reduced further (reaching the maximum point of reduction of 50%). Although the LTE's BTS coverage area is larger compared to the WiFi AP (more APs will be needed), still the incurred cost for the LTE base stations is much higher compared to the AP deployment, and operators should consider switching to 4G mostly in regions with very high traffic intensity.

5.5 Related Work

Authors in [78] propose to exploit opportunistic communications for information spreading in social networks. Their study is based on determining the minimum number of users that are able to reduce maximally the amount transmitted through the cellular network. A theoretical analysis with some optimization problems of offloading for opportunistic and vehicular communication are given in [79] and [80]. The LTE offloading into WiFi direct is subject of study in [81]. The work in [82] is mainly concerned with studying the conditions under which rate coverage is maximized for random deployment of APs belonging to different networks. Contrary to most of the other works, authors in [83] consider the situation in which cellular operators pay for using the AP from third parties. They use game theory to consider different issues, such as the amount of data and money a cellular operator should pay for utilizing the APs. In [84], a solution for mobile data offloading between 3GPP and non-3GPP access networks is presented. A WiFi based mobile data offloading architecture that targets the energy efficiency for smartphones was presented in [85]. An interesting work on determining the number of WiFi APs that need to be deployed in order to achieve a QoS is presented in [86].

As more related to mobile data offloading are the papers with measurements [12,68]. Authors in [12] have tracked the behavior of pedestrian users and their measurements suggest heavy-tailed periods of WiFi availability. The same holds for the time when there is no WiFi connectivity in the proximity of the mobile user. Similar conclusions for the availability periods are given in [68], where authors conduct measurements for vehicular users. These users are on metropolitan area buses. However, the mean duration of ON and OFF periods are different in the two scenarios of [12] and [68]. This is reasonable given the difference in speeds between vehicular and pedestrian users. The offloading efficiencies reported there are quite different, too. This result comes from different deadlines assumed in the two papers (related to delayed offloading). In addition to the two measurement-based studies [12, 68], already discussed in Section 5.4, there exists some additional interesting work in the area of offloading. Nevertheless, most related work does not deal with performance modeling and analysis of mobile data offloading. In [87], an integrated architecture has been proposed based on opportunistic networking to switch the data traffic from cellular to WiFi networks. The results were obtained from real data traces.

In [88], authors define a utility function related to delayed offloading to quantitatively describe the trade-offs between the user satisfaction in terms of the price that she has to pay and the experienced delay by waiting for WiFi connectivity. The authors use a semi-Markov process to determine the optimal handing-back point (deadline) for three scenarios. However, this analysis does not consider on-the-spot offloading, nor queueing effects. In this chapter, we do take into account the queueing process of the packets at the user. The work in [89] considers the traffic flow characteristics when deciding when to offload some data to the WiFi. However, there is no delay-related performance analysis. A cost based analysis is provided in [90].

The approach we are using here is based on the probability generating functions and is motivated from [75, 91].

To our best knowledge, the closest work in spirit to ours is [76]. The results in [76] are extension of the results in [12] containing the analysis for delayed offloading. The WiFi availability periods, as well as the periods of time when there is only cellular network coverage are modeled with exponential distributions. Also the packet sizes are exponentially distributed. Authors there also use 2D Markov chains to model the state of the system and use matrix-analytic methods to get a numerical solution for the offloading efficiency. However, their model does not apply directly to on-the-spot offloading. Also, they only provide numerical solutions. On the other hand, a performance analysis with closed form results for delayed offloading was provided in [92].

Summarizing, the novelty of our work is along the following dimensions: (i) we deal with on-the-spot offloading, (ii) we provide closed-form results and approximations, (iii) we provide an extension for generic packet size distributions, (iv) we validate our theory against realistic parameter values and distributions, (v) we provide some insight about the offloading gains that are of interest to both users and operators, (v) we generalize our analysis to capture any number of possible network connections.

5.6 Conclusion

In this chapter, we have proposed a queueing analytic model for the performance of on-the-spot mobile data offloading for generic number of access technologies, and we validated it against realistic WiFi network availability statistics. We have provided approximations for different utilization regions and have validated their accuracy compared to simulations and the exact theoretical results. We also showed that our model can be applied to a broader class of distributions for the durations of the periods between and with WiFi availability. Our model can provide insight on the offloading gains by using on-the-spot mobile data offloading in terms of both the offloading efficiency and delay. We have shown that the availability ratio of WiFi connectivity, in conjunction with the arrival rate plays a crucial role for the performance of offloading, as experienced by the user.

Chapter 6

Is it Worth to be Patient? Analysis and Optimization of Delayed Mobile Data Offloading

Operators have recently resorted to WiFi offloading to deal with increasing data demand and induced congestion. Researchers have further suggested the use of "delayed offloading": if no WiFi connection is available, (some) traffic can be delayed up to a given deadline, or until WiFi becomes available. Nevertheless, there is no clear consensus as to the benefits of delayed offloading, with a couple of recent experimental studies largely diverging in their conclusions. Nor is it clear how these benefits depend on network characteristics (e.g. WiFi availability), user traffic load, etc. In this chapter, we propose a queueing analytic model for delayed offloading, and derive the mean delay, offloading efficiency, and other metrics of interest, as a function of the user's "patience", and key network parameters. We validate the accuracy of our results using a range of realistic scenarios, and use these expressions to show how to optimally choose deadlines.

6.1 Introduction

Lately, an enormous growth in the mobile data traffic has been reported. This increase is due to a significant penetration of smartphones and tablets in the market, as well as Web 2.0 and streaming applications which have high-bandwidth requirements. Cisco [7] reports that by 2017 the mobile data traffic will increase by 13 times , and will climb to 13.2 exabytes per month. Mobile video traffic will comprise 66% of the total traffic, compared to 51% in 2012 [7].

This increase in traffic demand is overloading cellular networks, forcing them to operate close to (and often beyond) their capacity limits. Upgrading to LTE or LTE-advanced, as well as the deployment of additional network infrastructure could help alleviate this capacity crunch [8], but reports already suggest that such solutions are bound to face the same problems [9]. Furthermore, these solutions may not be cost-effective from the operators' perspective: they imply an increased cost (for power, location rents, deployment and maintenance), without a similar increase in revenues [10].

A more cost-effective way to cope with the problem of highly congested mobile networks is by offloading some of the traffic through Femtocells (SIPTO, LIPA [11]), and the use of WiFi. In 2012, 33% of the total mobile data traffic was offloaded [7]. Projections say that this will increase to 46% by 2017 [7]. Out of these, data offloading through WiFi has become a popular solution. Some of the advantages often cited compared to Femtocells are: lower cost, higher data rates, lower ownership cost [8], etc. Also, wireless operators have already deployed or bought a large number of WiFi access points (AP) [8].

There exist two types of WiFi offloading. The usual way of offloading is *on-the-spot offloading*: when there is WiFi available, all traffic is sent over the WiFi network; otherwise, *all* traffic is sent over the cellular interface. More recently, "delayed" offloading has been proposed: if there is currently no WiFi availability, (some) traffic can be delayed instead of being sent/received immediately over the cellular interface. In the simplest case, traffic is delayed until WiFi connectivity becomes available. This is already the case with current smartphones, where the user can select to send synchronization or backup traffic (e.g. Dropbox, Google+) only over WiFi. A more interesting case is when the user (or the device on her behalf) can choose a deadline (e.g. per application, per file, etc.). If up to that point no AP is detected, the data are transmitted through the cellular network [12, 68].

We have already analyzed the case of on-the-spot offloading in [73]. Delayed offloading offers additional flexibility and promises potential performance gains to both the operator and user. First, more traffic could be offloaded, further decongesting the cellular network. Second, if a user defers the transmission of less delay-sensitive traffic, this could lead to energy savings [77]. Finally, with more operators moving away from flat rate plans towards usage-based plans [93], users have incentives to delay "bulky" traffic to conserve their plan quotas or to receive better prices [94].

Nevertheless, there is no real consensus yet as to the added value of delayed offloading, if any. Recent experimental studies largely diverge in their conclusions about the gains of delayed offloading [12, 68]. Additionally, the exact amount of delay a flow can tolerate is expected to depend heavily on (a) the user, and (b) the application type. For example, a study performed in [95] suggests that "more than 50% of the interviewed users would wait up to 10 minutes to stream YouTube videos and 3-5 hours for file downloads". More importantly, *the amount of patience will also depend on the potential gains for the user.* As a result, two interesting questions arise in the context of delayed offloading:

- *If deadlines are externally defined (e.g. by the user or application), what kind of performance gains for the user/operator should one expect from delayed offloading and what parameters do these depend on?*

- *If an algorithm can choose the deadline(s) to achieve different performance-cost trade offs, how should these deadlines be optimally chosen?*

The main contributions of this chapter can be summarized as follows: (i) We propose a queueing analytic model for the problem of delayed offloading, based on two-dimensional Markov chains, and derive expressions for the average delay, and other performance metrics as a function of the deadlines, and key system parameters; we also give closed-form approximations for different regimes of interest; (*Section 6.2*) (ii) We validate our results extensively, using also scenarios and parameters observed in real measurement traces that depart from the assumptions made in our model; (*Section 6.3*) (iv) We formulate and solve basic cost-performance optimization problems, and derive the achievable tradeoff regions as a function of the network parameters (WiFi availability, user load, etc.) in hand (*Section 6.4*).

6.2 Analysis of Delayed Offloading

In this section, we formulate the delayed offloading problem, and derive analytical expressions for key metrics (e.g. mean per flow delay). We consider a mobile user that enters and leaves zones with WiFi coverage, with a rate that depends on the user's mobility (e.g. pedestrian, vehicular) and the environment (e.g. rural, urban). Without loss of generality, we assume that there is always cellular network coverage. We also assume that the user generates flows over time (different sizes, different applications, etc.) that need to be transmitted (uploaded or downloaded) over the network[1]. Whenever there is coverage by some WiFi AP, all traffic will be transmitted through WiFi, assuming for simplicity a First Come First Served (FCFS) queueing discipline. When the WiFi connectivity is lost, we assume that flows waiting in the queue and new flows arriving can be delayed until there is WiFi coverage again. However, each flow has a maximum delay it can wait for (a *deadline*), which might differ between flows and users [95]. If the deadline expires before the flow can be transmitted over some WiFi AP, then it is sent over the cellular network[2].

To facilitate the analysis of the above system, we make the following assumptions. We model the WiFi network availability as an ON-OFF alternating renewal process [20] $\left(T_{ON}^{(i)}, T_{OFF}^{(i)}\right), i \geq 1$, as shown in Fig. 6.1. The duration of each ON period (WiFi connectivity), $T_{ON}^{(i)}$, is assumed to be an exponentially distributed random variable with rate η, and independent of the duration of other ON or OFF periods. During such ON periods data can be transmitted over the WiFi network with a rate equal to μ. Similarly, all OFF periods (Cellular connectivity only) are assumed to be independent and exponentially distributed with rate γ, and a data rate that is lower than the WiFi rate[3]. We further assume that traffic arrives as a Poisson process with rate λ, and file sizes are exponentially distributed. Finally, to capture the fact that each file or flow may have a different deadline assigned to it, we assume that deadlines are also random variables that are exponentially distributed with rate ξ.

The above model is flexible enough to describe a large number of interesting settings: high vs. low WiFi availability (by manipulating $\frac{\gamma}{\gamma+\eta}$), low vs. high speed users (low γ, η vs. high γ, η, respectively), low utilization vs. congested scenarios (via λ and μ), etc. However, the assumptions of exponentiallity, while necessary to proceed with any meaningful analysis (as it will be soon made evident), might "hide" the effect of second order statistics (e.g. variability of ON/OFF periods, flow sizes, etc.). To address this, in Section 6.3 we relax most of these assumptions, and validate our results in scenarios with generic ON/OFF periods, generic flow size distributions, and non-exponential deadlines.[4]

Our goal is to analyze this system to answer the following questions: (i) if the deadlines are given (e.g. defined "externally" by the user or application), what is the expected performance

[1] We will use the terms "flow", "file", and "packet" interchangeably throughout the chapter, as the most appropriate term often depends on the application and the level at which offloading is implemented.

[2] In practice the switch in connectivity might sometimes occur while some flow is running. Without loss of generality, we will assume that the transmission is resumed from the point it was interrupted when WiFi was lost. It might continue over the cellular network (vertical handover) or paused until WiFi becomes available again or the deadline expires.

[3] Although this might not *always* be the case, everyday experience as well as a number of measurements [12] suggest this to be the case, on average.

[4] We could further extend our framework to arbitrary ON and OFF distributions using Coxian distributions and matrix-analytic methods [49]. However, the latter are only numerical, reducing their utility. We defer such scenarios and potential closed-form approximations to future work.

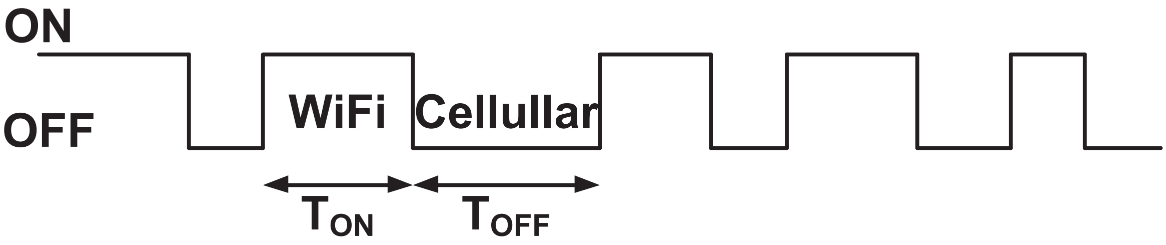

Figure 6.1: The WiFi network availability model.

as a function of network parameters like WiFi availability statistics, and user traffic load? (ii) if the deadlines are "flexible", i.e. the user would like to choose these deadlines in order to optimize his overall performance (e.g. trading off some delay, waiting for WiFi, to avoid the often higher energy and monetary cost of cellular transmission), how should they be chosen?

We will answer the first question in the remainder of this section, and use the derived expressions to provide some answers to the second question, in Section 6.4. Before proceeding, we summarize in Table 6.1 some useful notation. Also, the total time a file spends in the system (queueing+ service time) will be referred to as the *system time* or *transmission delay.*

Table 6.1: Variables and Shorthand Notation.

Variable	Definition/Description
T_{ON}	Duration of ON (WiFi) periods
T_{OFF}	Duration of periods (OFF) without WiFi connectivity
λ	Average packet (file) arrival rate at the mobile user
$\pi_{i,c}$	Stationary probability of finding i files in cellular state
$\pi_{i,w}$	Stationary probability of finding i files in WiFi state
π_c	Probability of finding the system under cellular coverage only
π_w	Probability of finding the system under WiFi coverage
p_r	Probability of reneging
η	The rate of leaving the WiFi state
γ	The rate of leaving the cellular state
μ	The service rate while in WiFi state
ξ	The reneging rate
$E[S]$	The average service time
$E[T]$	The average system (transmission) time
T_d	The deadline time
$\rho = \lambda E[S]$	Average user utilization ratio

6.2.1 Performance of WiFi queue

All files arriving to the system are by default sent to the WiFi interface with a deadline assigned (drawn from an exponential distribution). Files are queued (in FCFS order) if there is another file already in service (i.e. being transmitted) or if there is no WiFi connectivity at the moment,

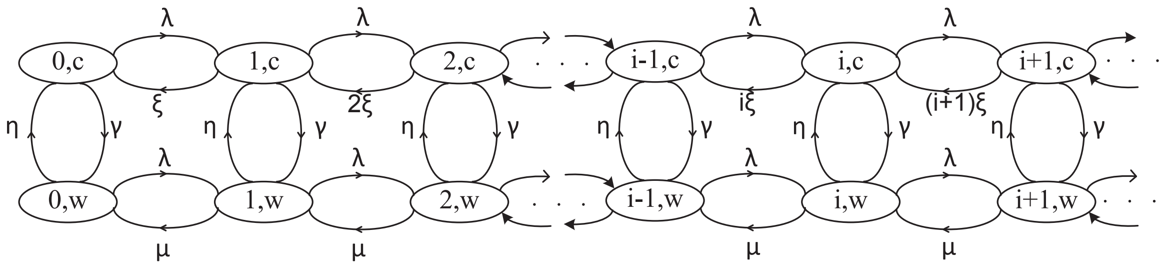

Figure 6.2: The 2D Markov chain for the WiFi queue in delayed offloading.

until their deadline expires. If the deadline for a file expires (either while queued or while at the head of the queue, but waiting for WiFi), the file *abandons* the WiFi queue and is transmitted through the cellular network. These kind of systems are known as queueing systems with *impatient* customers [96] or with *reneging* [97]. Throughout our analysis, we'll assume that files will abandon the queue only during periods without WiFi connectivity[5]. Nevertheless, in Section 6.3 we consider also the deterministic deadlines, i.e. a file will be sent over the cellular network. Our focus here will be on the WiFi queue for two reasons: First, this is the place where files accumulate most of the delay. Second, this is the point where a decision can be made, which will be relevant to the deadline optimization (Section 6.4). For the moment, we can assume that a file sent back to the cellular interface will incur a fixed delay (this might also include some mean queueing delay) that is larger, in general, than the service time over WiFi (i.e. $\frac{file\_size}{\mu}$).

Given the previously stated assumptions, the WiFi queue can be modeled with a 2D Markov chain, as shown in Figure 6.2. States with WiFi connectivity are denoted with $\{i, w\}$, and states with cellular connectivity only with $\{i, c\}$. i corresponds to the number of customers in the system (service+queue). During WiFi states, the system empties at rate μ (since files are transmitted 1-by-1) and during cellular states the system empties at rate $i \cdot \xi$ since any of the i queued packets can renege. The following theorem uses probability generating functions (PGF) to derive the mean system time for this queue. The use of PGFs in 2D Markov chains is known for quite a long time [52], [98], [99].

Theorem 16. *The mean system time for the WiFi queue when delayed mobile data offloading is performed is*

$$E[T] = \frac{1}{\lambda}\left[\left(1 + \frac{\gamma}{\eta}\right)\frac{\lambda - \mu(\pi_w - \pi_{0,w})}{\xi} + \frac{(\lambda - \mu)\pi_w + \mu\pi_{0,w}}{\eta}\right]. \tag{6.1}$$

Proof. Let $\pi_{i,c}$ and $\pi_{i,w}$ denote the stationary probability of finding i files when there is only cellular network coverage, or WiFi coverage, respectively.

[5] In this manner, abandonments are plausibly associated with the accumulated "opportunity cost", i.e. the time spent waiting for WiFi connectivity (the "non-standard" option for transmission). Instead, if WiFi is available, but there are some files in front, it might make no sense to abandon, as queueing delays might also occur in the cellular interface.

Writing the balance equations for the cellular and WiFi states gives

$$\begin{aligned}
(\lambda+\gamma)\pi_{0,c} &= \eta\pi_{0,w}+\xi\pi_{1,c} && (6.2)\\
(\lambda+\gamma+i\xi)\pi_{i,c} &= \eta\pi_{i,w}+(i+1)\xi\pi_{i+1,c}+\lambda\pi_{i-1,c} && (6.3)\\
(\lambda+\eta)\pi_{0,w} &= \gamma\pi_{0,c}+\mu\pi_{1,w} && (6.4)\\
(\lambda+\eta+\mu)\pi_{i,w} &= \gamma\pi_{i,c}+\mu\pi_{i+1,w}+\lambda\pi_{i-1,w} && (6.5)
\end{aligned}$$

The long term probabilities of finding the system in cellular or WiFi state are $\pi_c = \frac{\eta}{\eta+\gamma}$ and $\pi_w = \frac{\gamma}{\eta+\gamma}$, respectively.

We define the probability generating functions for both the cellular and WiFi states as $G_c(z) = \sum_{i=0}^{\infty}\pi_{i,c}z^i$, and $G_w(z) = \sum_{i=0}^{\infty}\pi_{i,w}z^i, |z| \leq 1$. After multiplying Eq.(6.3) with z^i and adding to Eq.(6.2) we obtain

$$(\lambda+\gamma)G_c(z)+\xi\left(1-\frac{1}{z}\right)\sum_{i=1}^{\infty}i\pi_{i,c}z^i = \eta G_w(z)+\lambda z G_c(z). \tag{6.6}$$

The summation in the above equation gives $\sum_{i=1}^{\infty}i\pi_{i,c}z^i = zG_c^{'}(z)$. Hence, after some rearrangements in Eq.(6.6) we obtain

$$\xi(1-z)G_c^{'}(z) = (\lambda(1-z)+\gamma)\,G_c(z)-\eta G_w(z). \tag{6.7}$$

Repeating the same procedure for Eq.(6.4)-(6.5) we get

$$(\lambda+\eta)G_w(z) = \gamma G_c(z)+\lambda z G_w(z)+\mu\left(\frac{1}{z}-1\right)(G_w(z)-\pi_{0,w}),$$

which after some rearrangements yields to

$$((\lambda z-\mu)(1-z)+\eta z)\,G_w(z) = \gamma z G_c(z)-\mu(1-z)\pi_{0,w}.$$

Next, we make two replacements $\alpha(z) = \lambda(1-z)+\gamma$, and $\beta(z) = (\lambda z-\mu)(1-z)+\eta z$. Now, we have the system of equations

$$G_w(z) = \frac{\gamma z G_c(z)-\mu(1-z)\pi_{0,w}}{\beta(z)}, \tag{6.8}$$

$$G_c^{'}(z)-\frac{\alpha(z)\beta(z)-\eta\gamma z}{\xi(1-z)\beta(z)}G_c(z) = \frac{\eta\mu\pi_{0,w}}{\xi\beta(z)}. \tag{6.9}$$

The roots of $\beta(z)$ are

$$z_{1,2} = \frac{\lambda+\mu+\eta\mp\sqrt{(\lambda+\mu+\eta)^2-4\lambda\mu}}{2\lambda}. \tag{6.10}$$

It can easily be shown that these roots satisfy the relation $0 < z_1 < 1 < z_2$. We introduce the function $f(z) = -\frac{\alpha(z)\beta(z)-\eta\gamma z}{\xi(1-z)\beta(z)}$, as the multiplying factor of $G_c(z)$ in the differential equation of Eq.(6.9). Performing some simple calculus operations, the above function transforms into

$$f(z) = -\frac{\lambda}{\xi}+\frac{\gamma}{\xi(1-z)}\left(\frac{\eta z}{\beta(z)}-1\right). \tag{6.11}$$

After some algebra and applying the *partial fraction expansion* the function $f(z)$ becomes

$$f(z) = -\frac{\lambda}{\xi} + \frac{\gamma}{\xi}\left(\frac{M}{z-z_1} + \frac{N}{z_2-z}\right). \tag{6.12}$$

We determine the coefficients M and N in the standard way as $M = \frac{\frac{\mu}{\lambda}-z}{z_2-z}\ |_{z=z_1} = \frac{\frac{\mu}{\lambda}-z_1}{z_2-z_1} = \frac{z_1z_2-z_1}{z_2-z_1} > 0$, and $N = \frac{\frac{\mu}{\lambda}-z}{z-z_1}\ |_{z=z_2} = \frac{\frac{\mu}{\lambda}-z_2}{z_2-z_1} < 0$.

In order to solve the differential equation in system Eq.(6.9) we can multiply it by $e^{\int f(z)dz}$. Hence, we get

$$G_c'(z)e^{\int f(z)dz} + f(z)G_c(z)e^{\int f(z)dz} = \frac{\eta\mu\pi_{0,w}}{\xi\beta(z)}e^{\int f(z)dz}. \tag{6.13}$$

We thus need to integrate the function in Eq.(6.12):

$$\int f(z)dz = -\frac{\lambda}{\xi}z + \frac{\gamma M}{\xi}\ln|z-z_1| - \frac{\gamma N}{\xi}\ln(z_2-z). \tag{6.14}$$

The constant normally needed on the right-hand side of Eq.(6.14) can be ignored in our case. We next raise Eq.(6.14) to the power of e to get

$$e^{\int f(z)dz} = e^{-\frac{\lambda}{\xi}z}\left|z-z_1\right|^{\frac{\gamma M}{\xi}}(z_2-z)^{-\frac{\gamma N}{\xi}}. \tag{6.15}$$

Now, Eq.(6.13) is equivalent to

$$\frac{d}{dz}\left(e^{-\frac{\lambda}{\xi}z}\left|z-z_1\right|^{\frac{\gamma M}{\xi}}(z_2-z)^{-\frac{\gamma N}{\xi}}G_c(z)\right) = \frac{\eta\mu\pi_{0,w}}{\xi\beta(z)}e^{\int f(z)dz} \tag{6.16}$$

We define $k_1(z)$ and $k_2(z)$ as

$$k_1(z) = e^{-\frac{\lambda}{\xi}z}(z_1-z)^{\frac{\gamma M}{\xi}}(z_2-z)^{-\frac{\gamma N}{\xi}}, z \le z_1, \tag{6.17}$$

$$k_2(z) = e^{-\frac{\lambda}{\xi}z}(z-z_1)^{\frac{\gamma M}{\xi}}(z_2-z)^{-\frac{\gamma N}{\xi}}, z \ge z_1. \tag{6.18}$$

Eq.(6.16) now becomes

$$\frac{d}{dz}\left(k_1(z)G_c(z)\right) = \frac{\eta\mu\pi_{0,w}}{\xi\beta(z)}k_1(z), z \le z_1, \tag{6.19}$$

$$\frac{d}{dz}\left(k_2(z)G_c(z)\right) = \frac{\eta\mu\pi_{0,w}}{\xi\beta(z)}k_2(z), z \ge z_1, \tag{6.20}$$

and after integrating we obtain

$$k_1(z)G_c(z) = \frac{\eta\mu\pi_{0,w}}{\xi}\int_0^z \frac{k_1(x)}{\beta(x)}dx + C_1, z \le z_1 \tag{6.21}$$

$$k_2(z)G_c(z) = \frac{\eta\mu\pi_{0,w}}{\xi}\int_{z_1}^z \frac{k_2(x)}{\beta(x)}dx + C_2, z \ge z_1. \tag{6.22}$$

The bounds of the integrals in Eq.(6.21) and Eq.(6.22) come from the defining region of z in Eq.(6.19)-(6.20). We need to determine the coefficients C_1 and C_2 in Eq.(6.21) and Eq.(6.22).

We take $z = 0$ in Eq.(6.21). We have $k_1(0) = z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}}$, and knowing that $G_c(0) = \pi_{0,c}$, we get for $C_1 = \pi_{0,c} z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}}$. In a similar fashion we get for $C_2 = 0$.

Finally, for the PGF in the cellular state we have

$$G_c(z) = \frac{\eta\mu\pi_{0,w}\int_0^z \frac{k_1(x)}{\beta(x)}dx + \xi\pi_{0,c} z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}}}{\xi k_1(z)}, z \leq z_1, \tag{6.23}$$

$$G_c(z) = \frac{\eta\mu\pi_{0,w}\int_{z_1}^z \frac{k_2(x)}{\beta(x)}dx}{\xi k_2(z)}, z \geq z_1. \tag{6.24}$$

In the last two equations, the 'zero probabilities' $\pi_{0,c}$ and $\pi_{0,w}$ are unknown. We can find them in the following way: We know that $\pi_c = \frac{\eta}{\eta+\gamma} = G_c(1) = \frac{\eta\mu\pi_{0,w}\int_{z_1}^1 \frac{k_2(x)}{\beta(x)}dx}{\xi k_2(1)}$. From this we have

$$\frac{\xi k_2(1)}{\eta+\gamma} = \mu\pi_{0,w}\int_{z_1}^1 \frac{k_2(x)}{\beta(x)}dx. \tag{6.25}$$

Similarly, from the boundary conditions in Eq.(6.23) for $z \leq z_1$, we get

$$\eta\mu\pi_{0,w}\int_0^{z_1} \frac{k_1(x)}{\beta(x)}dx + \xi\pi_{0,c} z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}} = 0. \tag{6.26}$$

After solving the system of equations Eq.(6.25) and Eq.(6.26), for the 'zero probabilities' we obtain

$$\pi_{0,w} = \frac{\xi k_2(1)}{(\eta+\gamma)\mu}\frac{1}{\int_{z_1}^1 \frac{k_2(x)}{\beta(x)}dx}, \text{and} \tag{6.27}$$

$$\pi_{0,c} = -\frac{\eta k_2(1)\int_0^{z_1} \frac{k_1(x)}{\beta(x)}dx}{(\eta+\gamma) z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}} \int_{z_1}^1 \frac{k_2(x)}{\beta(x)}dx}. \tag{6.28}$$

The value of the integral $\int \frac{k_1(x)}{\beta(x)}dx$ is always negative, hence $\pi_{0,c}$ is always positive.

By using a vertical cut between any two-pairs of neighboring states in Fig. 6.2 and writing balance equations we have

$$\lambda\pi_{i,c} + \lambda\pi_{i,w} = \mu\pi_{i+1,w} + (i+1)\xi\pi_{i+1,c}. \tag{6.29}$$

Summing over all i yields to

$$\lambda(\pi_c + \pi_w) = \mu(\pi_w - \pi_{0,w}) + \xi\sum_{i=0}^{\infty}(i+1)\pi_{i+1,c}. \tag{6.30}$$

The last equation, obviously reduces to

$$\lambda = \mu(\pi_w - \pi_{0,w}) + \xi E[N_c], \tag{6.31}$$

where $E[N_c] = G_c^{'}(1)$, and $E[N_w] = G_w^{'}(1)$. Eq.(6.31) yields

$$E[N_c] = \frac{\lambda - \mu(\pi_w - \pi_{0,w})}{\xi}. \tag{6.32}$$

So far, we have derived $E[N_c]$ as the first derivative at $z = 1$ of $G_c(z)$. In order to find the average number of files in the system, we need $E[N_w]$ as well. We can get it by differentiating Eq.(6.8)

$$G_w^{'}(z) = \frac{\beta(z)\left(\gamma G_c(z) + \gamma z G_c^{'}(z) + \mu\pi_{0,w}\right)}{\beta^2(z)} - \frac{\beta^{'}(z)\left(\gamma z G_c(z) - \mu(1-z)\pi_{0,w}\right)}{\beta^2(z)}, \tag{6.33}$$

and setting $z = 1$. After some calculus we obtain

$$E[N_w] = \frac{\left(\gamma E[N_c] + \mu\pi_{0,c}\right)\eta - \gamma\pi_c(\mu - \lambda)}{\eta^2}. \tag{6.34}$$

Replacing Eq.(6.32) into Eq.(6.34) we get

$$E[N_w] = \frac{\gamma}{\eta}\frac{\lambda - \mu\left(\pi_w - \pi_{0,w}\right)}{\xi} + \frac{\mu\pi_{0,w}}{\eta} - \frac{\gamma\pi_c(\mu - \lambda)}{\eta^2}. \tag{6.35}$$

The average number of files in the system is

$$E[N] = E[N_c] + E[N_w]. \tag{6.36}$$

Finally, using the Little's law $E[N] = \lambda E[T]$ [20], we obtain the average packet delay in delayed data offloading as in Eq.(6.1).

□

Remark. It can be shown easily that Eq.(6.1) holds also for the Processor haring (PS) service discipline, and even for Last Come First Served (LCFS) non-preemptive order of service.

The above result gives the total expected delay that incoming flows experience in the WiFi queue. For flows that do get transmitted over WiFi (i.e. whose deadline does not expire) this amounts to their total delay. Flows that end up reneging (deadline expires before transmission) must be transmitted through the cellular system and thus incur an additional delay Δ (related to their transmission time over the cellular link, i.e. $\frac{packet\_size}{cellular\_rate}$, and possibly some queueing delay as well). The following Corollary gives the probability of reneging for each.

Corollary 1. *The probability that an arbitrary flow arriving to the WiFi queue will renege, i.e. its deadline will expire before it can be transmitted over a WiFi AP is*

$$p_r = \frac{\lambda - \mu(\pi_w - \pi_{0,w})}{\lambda}. \tag{6.37}$$

In other words, the rate of flows sent back to the cellular network is given by $\lambda \cdot p_r$. This must be equal to $\xi \cdot E[N_c]$, which is the average abandonment rate in Fig. 6.2, i.e. $\lambda p_r = \xi E[N_c]$. Replacing $E[N_c]$ from Eq.(6.32) gives us the above result. This also gives us another important metric, the *offloading efficiency* of our system, namely the percentage of flows that get offloaded over some WiFi network, as $E_{off} = 1 - p_r$.

The above expressions can be used to predict the performance of a delayed offloading system, as a function of most parameters of interest, such as WiFi availability and performance, user traffic load, etc. As we shall see later, it does so with remarkable accuracy even in scenarios where many of the assumptions do not hold. However, Eq.(6.1) cannot easily be used to solve optimization problems related to the deadline (ξ), analytically, as the parameters $\pi_{0,c}$ and $\pi_{0,w}$ involve ξ in a non-trivial way. To this end, we propose next some closed-form approximations for the low and high utilization regimes.

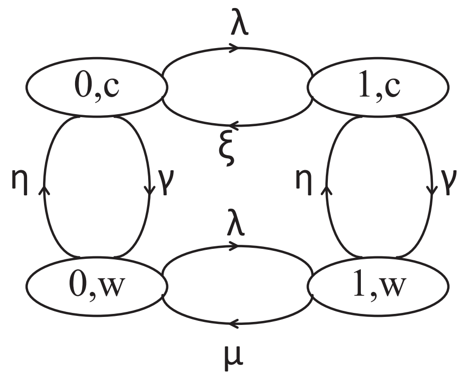

Figure 6.3: The reduced Markov chain for $\rho \to 0$.

6.2.2 Low utilization approximation

One interesting scenario is when resources are underloaded (e.g. nighttime, rural areas, or mostly low traffic users, etc) and/or traffic is relatively sparse (some examples are, background traffic from social and mailing applications, messaging, Machine-to-Machine communication, etc.). For very low utilization, the total system time essentially consists of the service time, as there is almost no queueing, so we can use a fraction of the Markov chain from Fig. 6.2 with only 4 states, as shown in Fig. 6.3 to derive $E[T]$ and p_r. The system empties at either state $(0,w)$ if the packet is transmitted while in WiFi connectivity period or state $(0,c)$, if the packet spends in queue more than the deadline it was assigned while waiting for WiFi availability.

The goal here is to find the average time until a packet arriving in a WiFi or cellular period finishes its service, i.e. the time until the system, starting from the state $(1,c)$ or $(1,w)$ first enters any of the states $(0,c)$ or $(0,w)$. Hence, the average service time is

$$E[S] = \frac{\eta}{\gamma+\eta}E[T_c] + \frac{\gamma}{\gamma+\eta}E[T_w], \tag{6.38}$$

where $E[T_c]$ ($E[T_w]$) is the average time until a packet that enters service during a cellular (WiFi) network period finishes its transmission. This can occur during a different period.

The expression for $E[T_c]$ is equal to

$$E[T_c] = P[I_c = 1]E[T_c|I_c = 1] + P[I_c = 0]E[T_c|I_c = 0], \tag{6.39}$$

where I_c is an indicator random variable having value 1 if the first transition from state $(1,c)$ is to state $(0,c)$. This means that the packet is transmitted during the same cellular period. Otherwise, its value is 0. The probabilities of these random variables are $P[I_c = 1] = \frac{\xi}{\xi+\gamma}$, and $P[I_c = 0] = \frac{\gamma}{\xi+\gamma}$, respectively. For the conditional expectations from Eq.(6.39), we have

$$E[T_c|I_c = 1] = \frac{1}{\xi+\gamma}, \tag{6.40}$$

$$E[T_c|I_c = 0] = \frac{1}{\xi+\gamma} + E[T_w]. \tag{6.41}$$

Eq.(6.40) is actually the expected value of the minimum of two exponentially distributed random variables with rates ξ and γ. Replacing Eq.(6.40) and (6.41) into Eq.(6.39), we get

$$E[T_c] - \frac{\gamma}{\xi+\gamma}E[T_w] = \frac{1}{\xi+\gamma}. \tag{6.42}$$

Following a similar procedure for $E[T_w]$ we obtain

$$E[T_w] - \frac{\eta}{\mu+\eta}E[T_c] = \frac{1}{\mu+\eta}. \tag{6.43}$$

After solving the system of equations Eq.(6.42)-(6.43), we have

$$E[T_w] = \frac{\xi+\gamma+\eta}{\xi\mu+\xi\eta+\mu\gamma}, \tag{6.44}$$

$$E[T_c] = \frac{\mu+\gamma+\eta}{\xi\mu+\xi\eta+\mu\gamma}. \tag{6.45}$$

Now, replacing Eq.(6.44)-(6.45) into Eq.(6.38), we have the average service time, and the low utilization approximation is ($E[T] \approx E[S]$)

$$E[T] = \frac{(\eta+\gamma)^2+\gamma\xi+\eta\mu}{(\xi\mu+\xi\eta+\mu\gamma)(\gamma+\eta)}. \tag{6.46}$$

To find the probability of reneging, we need to know $\pi_{0,c}$. We find it solving the local balance equations for Fig. 6.3. After solving the system we get

$$\pi_{0,c} = \frac{\eta}{\eta+\gamma}\frac{\xi(\mu+\lambda+\eta)+\mu\gamma}{\xi(\mu+\lambda+\eta)+\mu\gamma+\lambda^2+\lambda(\eta+\gamma+\mu)}. \tag{6.47}$$

Replacing Eq.(6.47) and $\pi_c = \frac{\eta}{\eta+\gamma}$ into Eq.(6.37), we get the probability of reneging for low utilization as

$$p_r = \frac{\theta_1\xi}{\theta_2\xi+\theta_3}, \tag{6.48}$$

where $\theta_1 = \frac{\eta(\lambda+\eta+\gamma+\mu)}{\eta+\gamma}$, $\theta_2 = \mu+\lambda+\eta$, and $\theta_3 = \mu\gamma+\lambda^2+\lambda(\eta+\gamma+\mu)$.

6.2.3 High utilization approximation

Another interesting regime is that of high utilization. As explained earlier, wireless resources are often heavily loaded, especially in urban centers, due to the increasing use of smart phones, tablets, and media-rich applications. Hence, it is of special interest to understand the average user performance in such scenarios. Here, we provide an approximation that corresponds to the region of high utilization ($\rho \to 1$).

The expected system time in the WiFi queue for a user with heavy traffic can be approximated by

$$E[T] = \frac{1}{\lambda}\left[\left(1+\frac{\gamma}{\eta}\right)\frac{\lambda-\mu\pi_w}{\eta} + \frac{(\lambda-\mu)\pi_w}{\eta}\right]. \tag{6.49}$$

The approximation Eq.(6.49) comes from Eq.(6.1) by replacing $\pi_{0,w} = 0$.

To find the approximation for the probability of reneging in the high utilization regime we proceed as follows. Since from Eq.(6.37) the only term that depends on ξ is $\pi_{0,w}$ (we will need it later to solve optimization problems), we will not take it equal to 0. We will approximate it by a first order Taylor approximation at $\xi = 1$. For that purpose, we will denote $\pi_{0,w}$ as $\pi_{0,w}(\xi)$. So, we write

$$\pi_{0,w}(\xi) = \pi_{0,w}(1) + (\xi - 1)\pi_{0,w}^{'}(1), \tag{6.50}$$

where $\pi_{0,w}(1) = \frac{A(1)}{(\eta+\gamma)\mu \int_{z_1}^{1} \frac{A(x)}{\beta(x)}dx}$, and $\pi_{0,w}^{'}(1) = \frac{A(1)}{(\eta+\gamma)\mu} \frac{\ln(A(1)e)\int_{z_1}^{1}\frac{A(x)}{\beta(x)}dx - \int_{z_1}^{1}\frac{A(x)\ln A(x)}{\beta(x)}dx}{\left(\int_{z_1}^{1}\frac{A(x)}{\beta(x)}dx\right)^2}$, where $A(x) = \frac{(x-z_1)^{\gamma M}}{(z_2-x)^{\gamma N}}e^{-\lambda x}$.

Hence, the probability of reneging in the high utilization regime can be approximated by

$$p_r = \frac{\lambda - \mu\pi_w}{\lambda} + \frac{\mu}{\lambda}\pi_{0,w}, \tag{6.51}$$

where $\pi_{0,w}$ is given by Eq.(6.50).

6.3 Performance evaluation

In this section we will validate our theory against simulations for a wide range of traffic intensities, different values of file sizes, WiFi availability periods with different distributions, and different deadline times. We define the WiFi availability ratio as $AR = \frac{E[T_{ON}]}{E[T_{ON}]+E[T_{OFF}]} = \frac{\gamma}{\eta+\gamma}$. Unless otherwise stated the durations of WiFi availability and unavailability periods will be drawn from independent exponential distributions with rates η and γ, respectively. The deadlines are exponentially distributed with rate ξ, although we will simulate scenarios with deterministic deadlines as well. We mainly focus on two scenarios, related to the user's mobility. The first one considers mostly pedestrian users with data taken from [12]. Measurements in [12] report that the average duration of WiFi availability period is 122 min, while the average duration with only cellular network coverage is 41 min (we use these values to tune η and γ). The availability ratio is thus 75%. The second scenario corresponds to vehicular users, related to the measurement study of [68]. An availability ratio of 11% has been reported in [68], although not all the details are mentioned there. For more details about the measurements we refer the interested reader to [12] and [68]. Finally, unless otherwise stated, file/flow sizes are exponentially distributed, and file arrival at the mobile user is a Poisson process with rate λ.

6.3.1 Validation of main delay result

We first validate here our model and main delay result (Eq.(6.1)) against simulations for the two mobility scenarios mentioned (pedestrian and vehicular). The data rate for WiFi is assumed to be 1 Mbps. The mean packet size is assumed to be 7.5 MB for the pedestrian scenario and 125 kB for the vehicular scenario.

Fig. 6.4 shows the average file transmission delay (i.e. queueing + transmission) for the pedestrian scenario, for two different average deadline times of $T_{d1} = 1$ hour ($\xi_1 = 1/3600s^{-1}$) and $T_{d2} = 2$ hours ($\xi_2 = 1/7200s^{-1}$), respectively. The range of arrival rates shown corresponds to a server utilization of 0-0.9. We can observe from Fig. 6.4 that there is a good match between theory and simulations. Furthermore, the average file transmission delay is increased by increasing the arrival rate, as expected, due to queueing effects. On the other hand, the

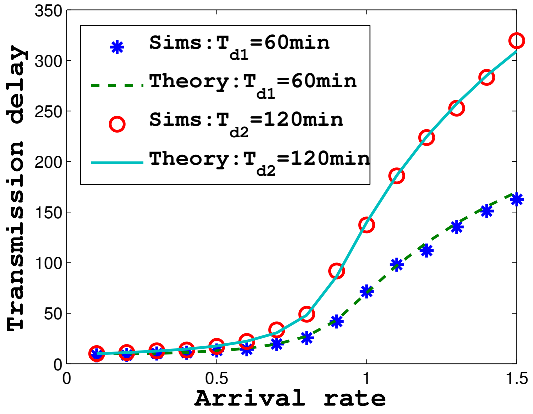

Figure 6.4: The average delay for pedestrian users' scenarios.

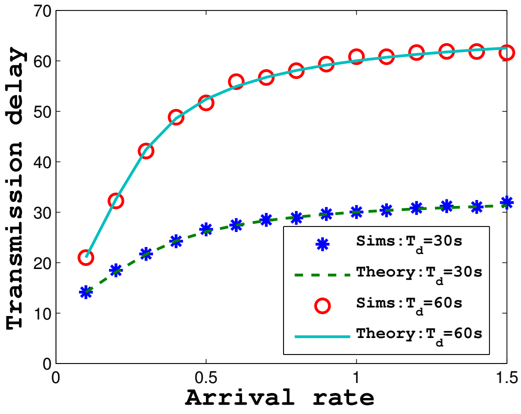

Figure 6.5: The average delay for vehicular users' scenarios.

average delay increases for higher deadlines, since flows with lower deadlines leave the WiFi queue earlier, leading to smaller queueing delays.

Fig. 6.5 further illustrates the average file transmission delay for the vehicular scenario with average deadline times $T_{d1} = 30$s ($\xi_1 = 1/30s^{-1}$) and $T_{d2} = 60$s ($\xi_2 = 1/60s^{-1}$). Despite the differences of the vehicular scenario, similar conclusions can be drawn.

Finally, Table 6.2 depicts the respective probabilities of reneging for the two scenarios. The percentage of flows that abandon the WiFi queue is higher in the vehicular scenario, since the availability ratio of the WiFi network is very small (11%), and deadlines are rather small. These observations agree with [68]. Nevertheless, our theory matches simulation values in all scenarios.

Table 6.2: Probability of reneging for pedestrian and vehicular scenarios.

Scenario	Deadline	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 1.5$
Pedestrian(Theory)	1 hour	0.103	0.109	0.252	0.501
Pedestrian(Simulation)	1 hour	0.1	0.117	0.239	0.508
Vehicular(Theory)	60 s	0.32	0.778	0.889	0.926
Vehicular(Simulation)	60 s	0.32	0.776	0.891	0.925

So far, we have assumed exponential distributions for ON and OFF periods, according to our model. While the actual distributions are subject to the user mobility pattern, a topic of intense research recently, initial measurement studies ([12,68]) suggest these distributions to be "heavy-tailed". To this end, we consider a scenario with "heavy-tailed" ON/OFF distributions (Bounded Pareto). Due to space limitations, we focus on the vehicular scenario only. The shape parameters for the Bounded Pareto ON and OFF periods are $\alpha = 0.59$ and $\alpha = 0.64$, respectively. The average deadline is 100s. Fig. 6.6 compares the average file delay against our theoretical prediction. Interestingly, our theory still offers a reasonable prediction accuracy, despite the considerably higher variability of ON/OFF periods in this scenario. While we cannot claim this to be a generic conclusion for any distribution and values, the results underline the utility of our model in practice.

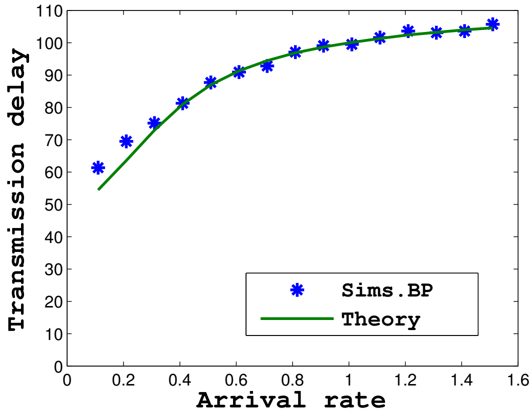

Figure 6.6: The delay for BP ON-OFF periods vs. theory.

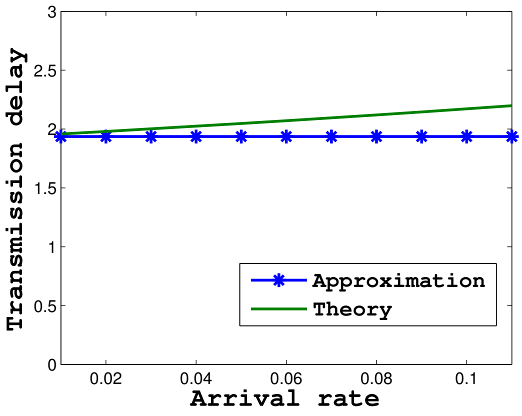

Figure 6.7: Low utilization delay approximation for $AR = 0.75$.

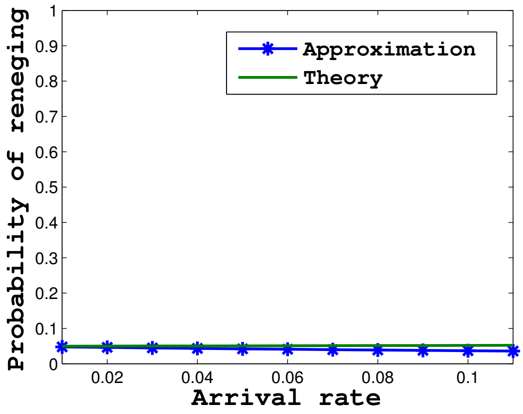

Figure 6.8: Low utilization p_r approx. for $AR = 0.75$.

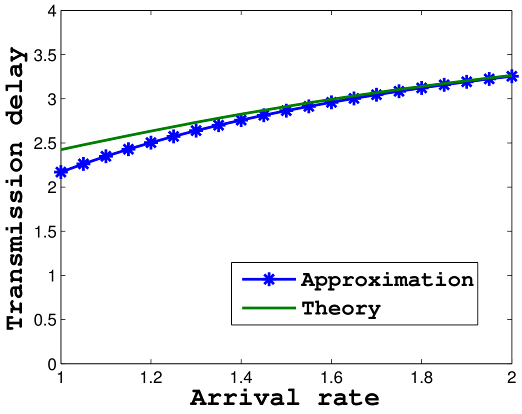

Figure 6.9: High utilization delay approximation for $AR = 0.5$.

6.3.2 Validation of approximations

We next validate the approximations we have proposed in Section 6.2. We start with the low utilization approximation of Section 6.2.2 and consider the availability ratio to be 0.75 (similar accuracy levels have been obtained with other values) and with a deadline of 2 min. Fig. 6.7 shows the packet delay for low arrival rates in the range $0.01 - 0.11$, which correspond to a maximum utilization of up to 0.2. As λ increases, the difference between the approximated result and the actual value increases, since we have considered only the service time for this approximation. The same conclusion holds for the probability of reneging (Fig. 6.8).

Next, we consider the high utilization regime and respective approximation (Eq.(6.49)). We consider utilization values around 0.8. Fig. 6.9 shows the delay for high values of λ, and an availability ratio of 0.5. We can see there that our approximation is very close to the actual delay and should become exact as ρ gets larger.

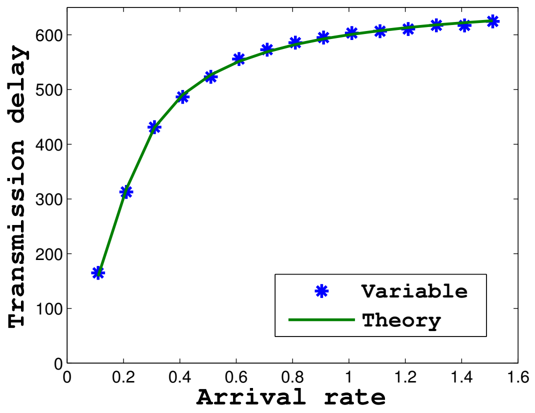

Figure 6.10: Variable WiFi rates with the same average as theory.

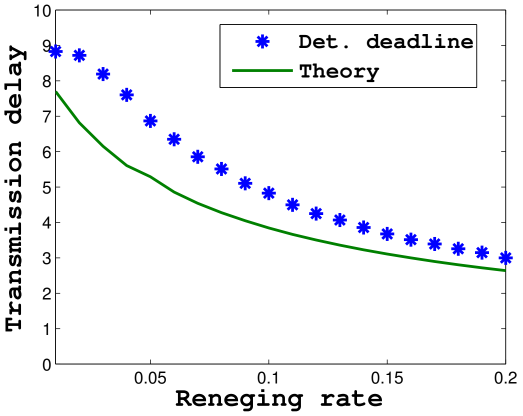

Figure 6.11: The delay for deterministic deadlines vs. theory.

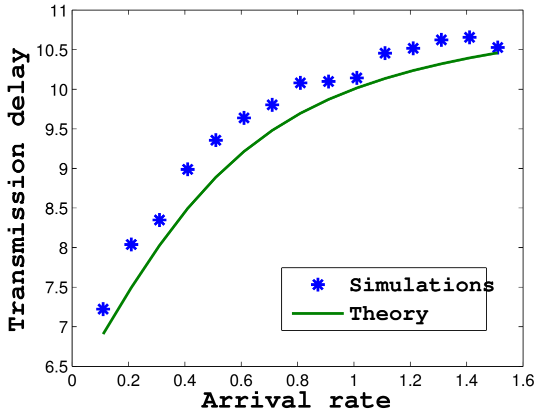

Figure 6.12: Deterministic packets.

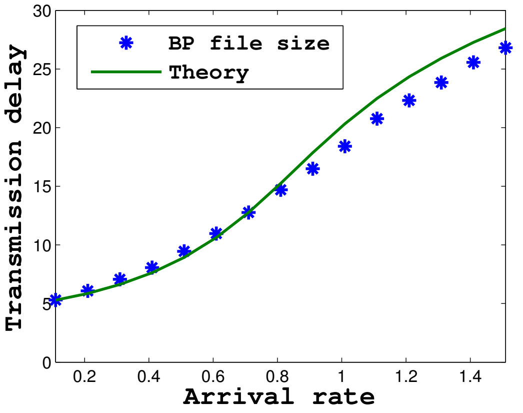

Figure 6.13: The delay for BP packet sizes vs. theory.

6.3.3 Variable WiFi rates and non-exponential parameters

While in our model we consider a fixed transmission rate for all WiFi hotspots, this is not realistic in practice. For this reason, we have also simulated scenarios where the WiFi rate varies uniformly in the range 0.4-1.6 Mbps. Fig. 6.10 shows the delay for for the vehicular scenario with a deadline of 10 minutes. Even in this case, our theory can give solid predictions for the incurred delay.

In all of the above scenarios, we have assumed variable deadlines for each file (drawn from an exponential distribution). In some cases, the user might choose the same deadline for many (or most) flows that can be delayed, which would be a measure of her patience. To this end, we simulate a scenario where the deadline is fixed for an arrival rate of 0.1. The other parameters are the same as for the vehicular scenario. In Fig. 6.11 we compare simulation results for this scenario against our theory (which assumes exponentially distributed deadlines with the same average). It is evident that even in this case, there is a reasonable match with our theory.

To conclude our validation, we finally drop the exponential packet assumption as well, and test our theoretical result vs. generic file size results. Fig. 6.12 compares analytical and simulation results for deterministic (10 s deadline), and Fig. 6.13 does it for Bounded Pareto distributed files sizes (shape parameter $\alpha = 1.2$ and $c_v = 3$), where the deadline is $T_d = 20$s. Mean file size is in both cases 125KB, the availability ratio is 0.5, and the rest of the parameters correspond to the vehicular scenario. Our theoretical prediction remains reasonably close, despite higher size variability.

6.3.4 Delayed offloading gains

In this last part, we will investigate the actual gains from data offloading, in terms of offloading efficiency. Higher offloading efficiency means better performance for both client and operator. We compare the offloading efficiencies for on-the-spot offloading [73] vs. delayed offloading for different deadline times ($T_{d1} = 2min, T_{d2} = 1min$). Fig.6.14 illustrates the offloading efficiency vs. availability ratio for a moderate arrival rate of $\lambda = 0.2$. For comparison purposes we also depict the line $x = y$ (offloading efficiency = availability ratio). First, as expected, we can observe that offloading efficiency increases with AR. However, this increase is not linear. More interestingly, the actual offloading efficiencies are always higher than the respective availability ratios. As expected, the delayed offloading provides higher offloading efficiencies compared to on-the-spot offloading, with higher deadlines leading to higher offloading efficiencies.

6.4 Optimizing Delayed Offloading

The results considered so far allow us to predict the expected system delay when the deadlines are defined externally (e.g. by the user or the application). However, the user (or the device on her behalf) could choose the deadline in order to solve an optimization problem among additional (often conflicting) goals, such as the monetary *cost* for accessing the Internet and the *energy consumption* of the device. For example, the user might want to minimize the delay subject to a maximum (energy or monetary) cost, or to minimize the cost subject to a maximum delay the user can tolerate.

To formulate and solve such optimization problems, we need analytical formulas for the average delay and the incurred cost. We already have such formulas for the delay of files sent over WiFi, where we will use the two approximations of Section 6.2.2 and 6.2.3. Furthermore, we can assume that files transmitted over the cellular network incur a fixed delay Δ, capturing both the service and queueing delays over the cellular interface[6]. To proceed, we need to also assume simple models for energy and cost, in order to get some initial intuition about the tradeoffs involved. We are aware that reality is more complex (for both energy and cost) and may differ based on technology (3G, LTE), provider, etc. We plan to extend our models in future work.

Assume a user has to download or upload a total amount of data equal to L. On average $p_r \cdot L$ data units will be transmitted through the cellular interface. Assume further that D_c and D_w denote the costs per transmitted data unit for a cellular and WiFi network, where $D_w < D_c$ (often $D_w = 0$). Finally, let c_c and c_w denote the transmission rates, and E_c and E_w energy

[6] We could also try to model the cellular queue as an M/M/1 or G/M/1 system, but we are more interested in the dynamics of the WiFi queue, since this is where the reneging decisions take place. To keep things simple, we defer this to future work.

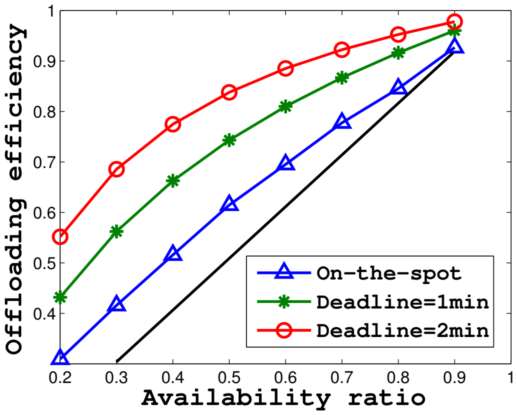

Figure 6.14: Offloading gains for delayed vs. on-the-spot offloading.

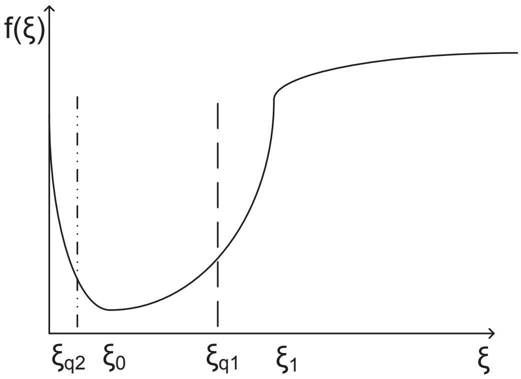

Figure 6.15: The delay function for the optimization problem.

spent per time unit during transmission over the cellular and WiFi network, respectively. It is normally the case that $c_c < c_w$ as well as $E_c \approx E_w$ [100].

It follows then that the total monetary and energy costs, D, and E, could be approximated by

$$D = (D_c - D_w)p_r + D_w \text{ and } E = \left(\frac{E_c}{c_c} - \frac{E_w}{c_w}\right) p_r + \frac{E_w}{c_w}. \tag{6.52}$$

6.4.1 Optimization problems

Eq.(6.52) suggests that both the average power consumption and cost depend linearly on the probability of reneging, p_r, which we have also derived in Section 6.2, and which is a function of the system deadline $\frac{1}{\xi}$. The system delay is also a function of ξ. We can thus formulate optimization problems of the following form, for both the high and low utilization regimes, where ξ is the optimization parameter

$$\begin{aligned} \min_{\xi} \quad & E[T] + p_r\Delta \\ \text{s. t.} \quad & p_r \le P_r^{max}, \end{aligned} \tag{6.53}$$

where $E[T]$ is given by Eq.(6.46), and p_r by Eq.(6.48), for low utilization, and Eq.(6.49) and Eq.(6.51), for high utilization, respectively. Due to the linearity of Eq.(6.52), we can express the constraint directly for p_r, where P_r^{max} depends on whether we consider monetary cost, energy or a weighted sum of both, and the respective parameters. Finally, we can also exchange the optimization function with the constrain, to minimize the cost, subject to a maximum delay. This provides us with a large range of interesting optimization problems we can solve.

If we express the inequality constraint in Eq.(6.53) through ξ, we have the equivalent constraint $\xi \le \frac{\theta_3 P_r^{max}}{\theta_1 - \theta_2 P_r^{max}}$. The probability of reneging from Eq.(6.48) is an increasing function of ξ, since $p_r'(\xi) > 0$. This implies that maximum p_r corresponds to maximum ξ. We denote by $f(\xi)$ the total average delay of Eq.(6.53) (delay function from now on). Hence, we have

$$f(\xi) = \frac{A_1\xi + A_2}{B_1\xi + B_2} + \frac{\theta_1\xi}{\theta_2\xi + \theta_3}\Delta, \tag{6.54}$$

where $A_1 = \gamma$,$A_2 = (\eta+\gamma)^2 + \eta\mu$,$B_1 = (\mu+\eta)(\gamma+\eta)$, $B_2 = \mu\gamma(\gamma+\eta)$. In order to solve the optimization problem given by Eq.(6.53), we need to know the behavior of the delay function. For that purpose, we analyze the monotonicity and convexity of Eq.(6.54). To do that we need the first and second derivatives, which are

$$f^{'}(\xi) = \frac{A_1B_2 - A_2B_1}{(B_1\xi + B_2)^2} + \frac{\theta_1\theta_3\Delta}{(\theta_2\xi+\theta_3)^2}, \text{and}$$

$$f^{''}(\xi) = \frac{2(A_2B_1 - A_1B_2)}{(B_1\xi + B_2)^3} - \frac{\theta_1\theta_2\theta_3\Delta}{(\theta_2\xi+\theta_3)^3}.$$

It is worth noting that $A_1B_2 < A_2B_1$. This prevents delay function being always concave. The delay function is decreasing in the interval for which $f^{'}(\xi) \leq 0$. This happens when

$$\xi \leq \xi_0 = \frac{\theta_3\sqrt{\frac{A_2B_1 - A_1B_2}{\theta_1\theta_3\Delta}} - B_2}{B_1 - \theta_2\sqrt{\frac{A_2B_1 - A_1B_2}{\theta_1\theta_3\Delta}}}.$$

Hence, the delay function is decreasing in the interval $(0, \xi_0)$, and increasing in the rest, with ξ_0 being a minimum. Further, the solution of $f^{''}(\xi) > 0$ gives the interval where the function is convex. This happens when

$$\xi \leq \xi_1 = \frac{\theta_3\sqrt[3]{2\frac{A_2B_1 - A_1B_2}{\theta_1\theta_2\theta_3\Delta}} - B_2}{B_1 - \theta_2\sqrt[3]{2\frac{A_2B_1 - A_1B_2}{\theta_1\theta_2\theta_3\Delta}}}. \tag{6.55}$$

It can be easily proven that $\xi_0 < \xi_1$.

Such constrained-optimization problems are often solved with the Lagrangian method and KKT conditions. However, the optimal solution for our problem can be found more easily. The delay function looks like in Fig. 6.15. The optimal deadline depends on the maximum cost, that is proportional to the probability of reneging. So, we can determine the optimal deadline based on the value of P_r^{max}. If this value of P_r^{max} is quite high, the corresponding reneging rate ξ_{q1} (dashed line in Fig. 6.15) will be higher than the global minimum ξ_0. Consequently, the global minimum of Eq.(6.55) is also the optimal reneging rate. On the other hand, if the maximum cost is quite low (low P_r^{max}), the maximum reneging rate $\xi_{q,2}$ (dotted line in Fig. 6.15) is lower than the global minimum. This implies that the minimum delay will be achieved for the maximum reneging rate of $\xi_{q,2} = \frac{\theta_3 P_r^{max}}{\theta_1 - \theta_2 P_r^{max}}$. In other words, the average deadline time that minimizes the delay for a given maximum cost is

$$T_{d,opt} = \frac{1}{\xi_{opt}} = \frac{1}{\min\left(\xi_0, \frac{\theta_3 P_r^{max}}{\theta_1 - \theta_2 P_r^{max}}\right)}. \tag{6.56}$$

Similar steps can be followed to solve the same optimization problem for high utilization, as well as other problems.

Optimization problem 2: After minimizing the transmission delay subject to a maximum reneging rate (cost, energy), our next goal now is to minimize the reneging probability subject to a maximum transmission delay, which can be for example due to QoS requirements. Hence, the optimization problem in this case would be

$$\begin{aligned} \min_{\xi} \quad & p_r = \frac{\theta_1 \xi}{\theta_2 \xi + \theta_3} \\ \text{s. t.} \quad & E[T] + p_r \Delta \leq T_{max}. \end{aligned} \tag{6.57}$$

Just as in Optimization problem 1, we study the monotonicity and convexity of the delay function, with the only difference that now it is the constrain function. For the probability of reneging, we already now that it is an increasing function. Following a similar procedure as in the previous problem, we get for the optimum value of the deadline (from a quadratic constraint)

$$T_{d,opt} = \frac{1}{\max\left(0, \frac{K_2 - \sqrt{K_2^2 - 4K_1K_3}}{2K_1}\right)}, \tag{6.58}$$

where $K_1 = A_1\theta_2 + \theta_1 \Delta B_1 - T_{max} B_1 \theta_2$, $K_2 = T_{max} B_1 \theta_3 + T_{max} B_2 \theta_2 - A_1 \theta_3 - A_2 \theta_2 - \theta_1 \Delta B_2$, $K_3 = A_2 \theta_3 - T_{max} B_2 \theta_3$, $C_1 = \frac{1}{\lambda}\left(1 + \frac{\gamma}{\eta}\right)(\lambda - \mu \pi_w)$.

Next, we give the solutions to optimization problems for high utilization regime (Optimization problems 3 and 4), where the expressions for $E[T]$ and p_r are given by Eq.(6.49) and Eq.(6.51), respectively.

Optimization problem 3: Having solved the optimization problem for low utilization, we move on to the high utilization regime, and use the approximation for the average file delay to solve two optimization problems. In the first one, our objective function is the transmission delay, and the constrain function is the probability of reneging. So, we have the following problem

$$\begin{aligned} \min_{\xi} \quad & E[T] + p_r \Delta \\ \text{s. t.} \quad & p_r \leq P_r^{max}, \end{aligned} \tag{6.59}$$

Using the same methodology as before, we get the optimal value of the deadline time that minimizes the average delay, given a maximum cost. That value is

$$T_{d,opt} = \frac{1}{\min\left(\sqrt{\frac{C_1}{D_1 \Delta}}, \frac{P_r^{max} - D_2}{D_1}\right)}, \tag{6.60}$$

where $C_1 = \frac{1}{\lambda}\left(1 + \frac{\gamma}{\eta}\right)(\lambda - \mu \pi_w)$, $C_2 = \frac{(\lambda - \mu)\pi_w}{\lambda \eta}$, $D_1 = \frac{\lambda - \mu\left[\pi_w - \pi_{0,w}(1) + \pi_{0,w}^{'}(1)\right]}{\lambda}$, and $D_2 = \frac{\mu}{\lambda}\pi_{0,w}^{'}(1)$.

Optimization problem 4: Finally, in the last optimization problem we want to minimize the probability of reneging subject to a maximum delay a packet should experience in the system. The corresponding optimization problem is

$$\begin{aligned} \min_{\xi} \quad & p_r \\ \text{s. t.} \quad & E[T] + p_r \Delta \leq T_{max}, \end{aligned} \tag{6.61}$$

that gives as a solution

$$T_{d,opt} = \frac{2D_1\Delta}{T_{max} - C_2 - D_2\Delta - \sqrt{(T_{max} - C_2 - D_2\Delta)^2 - 4C_1D_1\Delta}}. \tag{6.62}$$

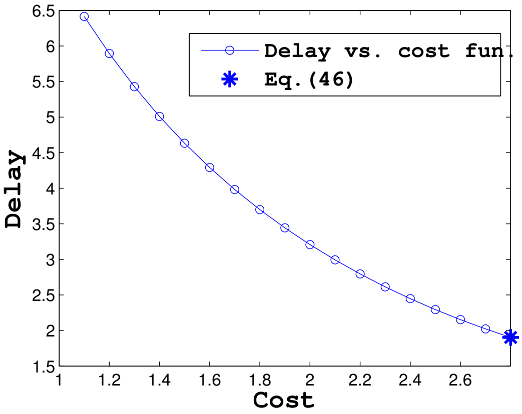

Figure 6.16: The delay vs. cost curve for high cellular rate.

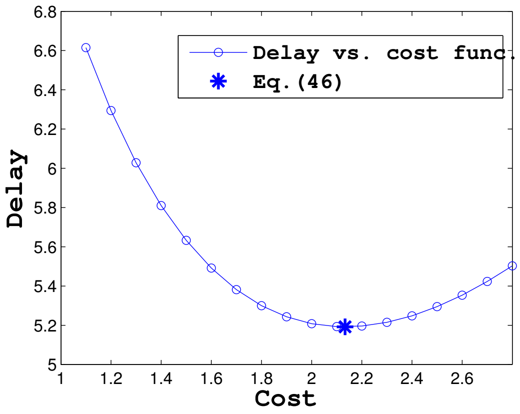

Figure 6.17: The delay vs. cost curve for low cellular rate.

6.4.2 Optimization evaluation

We will now validate the solutions of the previous optimization problem for two different cases. In both of them the arrival rate is 0.1, and the maximum cost per data unit one can afford is 2.8 monetary units. The transmission of a data unit through WiFi costs 1, and through cellular 5 units. The choice of these values is simply for better visualizing the results; different values yield similar conclusions. Fig. 6.16 shows the delay vs. cost curve for cellular rate being 2× lower than WiFi rate. First thing that we can observe is that the minimum delay is achieved for the highest possible cost (2.8). The optimal average deadline is $T_d = 1$s. This is in agreement with the optimal value predicted from Eq.(6.56), and shown with an asterisk in Fig. 6.16. We replace Eq.(6.48) into Eq.(6.52) to get the relationship between the cost and the renege rate. We have shown in Eq.(6.52) that the cost is directly proportional to p_r, and the later one is an increasing function of ξ. This implies that the maximum cost is in fact the maximum ξ (minimum deadline). This practically means that in Eq.(6.53), Δ is small and that the delay in the WiFi queue represents the largest component of the delay. As a consequence of that, it is better to redirect the files through the cellular interface as soon as possible. Hence, in these cases (when cellular rate is comparable to the WiFi), the optimum is to assign the shortest possible deadline constrained by the monetary cost.

Fig. 6.17 corresponds to a scenario with the same parameters as in Fig. 6.16, except that now the cellular rate is much lower (10×). In that case, Δ is high, and $p_r\Delta$ is the largest component of the delay function. As can be seen from Fig. 6.17, it is not the best option to leave as soon as possible from the WiFi queue, i.e. choose the smallest possible deadline. The optimum delay is achieved for $T_d = 5$s. This corresponds to an average cost of $D = 2.1$ which is also very close to the theoretical solution of the problem. This is reasonable since for a large difference between the WiFi and cellular rates it is better to wait and then (possibly) be served with higher rate, than to move to a much slower interface (cellular).

Further, we use the solutions of the four optimization problems for exponentially distributed deadline times to see how accurately our theory can predict the optimal deadline times, but for deterministic deadlines. The optimal policy essentially finds the optimal value for the *average* deadline (assuming these exponential). In practice, the chosen deadline will be assigned to all files, and will be deterministic. We consider four scenarios, one for each optimization problem. The costs are the same as before. The arrival rate for low utilization scenarios is 0.1, while for the high ones, 1.5. In Table 6.3, we show the optimal deadlines by using our model (e.g. Eq.(6.56)), and the optimal deterministic deadlines by using simulations (delay vs. cost plots) with the same parameters as in theory. As can be seen from Table 6.3, the error in determining the optimal deadline decreases for higher arrival rates. The error is in the range 10-20%. This is reasonable, since the simulated scenarios are with deterministic deadlines and in our theory we use exponential deadlines. Another reason is that in optimization problems, we are only using the low and high utilization approximations and not the exact result (Eq.(6.1)).

Table 6.3: Optimal deterministic deadline times vs theory.

Sc.	Constraint	T_d (theory)	T_d (deterministic)	Relative error
1	$D \leq 2.8$	2.71	2.2	18%
2	$T_{max} = 6.6$	1.56	1.22	22%
3	$D \leq 3.8$	1.73	1.55	11%
4	$T_{max} = 15$	7.97	6.82	15%

6.5 Related Work

Some recent influential work in offloading relates to measurements of WiFi availability [12, 68]. Authors in [12] have tracked the behavior of 100 users (most of which were pedestrians) and their measurements reveal that during 75% of the time there is WiFi connectivity. In [68], measurements were conducted on users riding metropolitan area buses. In contrast to the previous study, the WiFi availability reported there is only around 10%. The mean duration of WiFi availability and non-availability periods is also different in the two studies, due to the difference in speeds between vehicular and pedestrian users. The most important difference between the two studies relates to the reported offloading efficiency, with [68] reporting values in the range from 20-33% for different deadlines, and [12] reporting that offloading does not exceed 3%. We believe this is due to the different deadlines assumed together with the different availabilities.

The authors in [88] define a utility function related to delayed offloading to quantitatively describe the trade-offs between the user satisfaction in terms of the price that she has to pay and the experienced delay by waiting for WiFi connectivity. However, their analysis does not consider queueing effects. Such queueing effects may affect the performance significantly, especially in loaded systems (which are of most interest) or with long periods without WiFi. The work in [89] considers the traffic flow characteristics when deciding when to offload some data to the WiFi. However, there is no delay-related performance analysis. In [101], authors consider the problem of efficient utilization of multiple access networks through optimal assignment of traffic flows to different networks. Modeling the cost factors is the focus of [90], which also shows where the offloading APs should be installed. A WiFi offloading system that takes into account a user's throughput-delay tradeoffs and cellular budget constraints is proposed in [102]. However, only heuristic algorithms are proposed, and queueing effects are ignored. Finally, in [87], an integrated architecture has been proposed based on opportunistic networking to switch the data traffic from the cellular to WiFi networks. Summarizing, in contrast to our work, these papers either perform no analysis or use simple models that ignore key system effects such as queueing.

To our best knowledge, the closest work in spirit to ours is [76]. The results in [76] are the extension of the results in [12] containing the analysis for delayed offloading. Authors there also use 2D Markov chains to model the state of the system. However, they use matrix-analytic methods to obtain a numerical solution for the offloading efficiency. Such numerical solutions unfortunately do not provide insights on the dependencies between different key parameters, and cannot be used to formulate and analytically solve optimization problems that include multiple metrics.

As a final note, in [73], we have proposed a queueing analytic model for on-the-spot mobile data offloading, and a closed form solution was derived for the average delay. While the model we propose here shares some similarities (ON/OFF availabilities, 2D Markov chain approach) with the basic model in [73], it is in fact considerably more difficult to solve.

6.6 Conclusion

In this chapter, we have proposed a queueing analytic model for the performance of delayed mobile data offloading, and have validated it against realistic WiFi network availability statistics. We have also considered a number of scenarios where one or more of our model's assumptions do not hold, and have observed acceptable accuracy, in terms of predicting the system delay as a function of the user's patience. Finally, we have also shown how to manipulate the maximum deadlines, in order to solve various optimization problems involving the system delay, monetary costs, and energy costs. In future work, we intend to consider more complex models for both the WiFi and cellular queues, as well as per-flow scheduling and dispatch policies.

Chapter 7

Conclusions and Future Work

The development of intermittent wireless networks has come as an urgent necessity to relieve the current networks after the ubiquitous penetration of latest high tech devices into the technology market on one side, and the bandwidth-exigent applications and services that they can provide on the other side. Such type of networks are *cognitive radio networks*, which are based on the dynamic spectrum access. Another very popular emerging way of intermittent communications is *mobile data offloading* through WiFi AP. Hence, it is very important to be able to predict the performance we could expect in such networks, and understand to what extent the key figure of merits, such as the *delay* and *throughput*, depend on the availability of the secondary resources, traffic intensity, file size. etc.

Based on the things said above, it is very important to perform some analysis. In order to do that, first we must have the appropriate realistic models that are able to capture the behavior of our systems. As a first step in all of the chapters of this thesis was to provide these models, and then to perform the analysis leading to closed-form results, and solve different optimization problems. This then gives a better picture of the performance.

The advantage of our models is that they offer closed-from solutions, which depend only on few network parameters, and still can be used to optimize the overall performance by tuning few of the parameters. The derived expressions can be used to design efficient network protocols and scheduling algorithms.

Specifically, our contributions are summarized as following:

- We commence in Chapter 2 with a queueing analytic model to calculate the average packet delay in an interweave cognitive radio network under generic ON-OFF periods of primary activity. We have shown that variability of licensed user activity is very important, and often much more important than the utilization itself. This is the key for the protocol design. As we have shown, the actual delay is a complex interplay between secondary traffic characteristics (intensity and packet sizes) and channel characteristics (1st and 2nd moments of idle and busy periods of PU).

- Next, in Chapter 3 our objective is to optimize the average throughput. We have analyzed the spectrum scanning process in cognitive radio networks and explored the ways to maximize the average throughput rate. We have introduced the notion of a threshold value as the number of channels we are allowed to lose before the initiation of the scanning procedure. This threshold value provides the best results in terms of average data rate. It is proven that no such threshold value exists for the case of homogeneous (i.i.d.) channels if there is no initial cost to be paid at the beginning of the scanning process. However, this value exists for heterogeneous independent channels and its value depends on the variability of the OFF periods of the channels in use. We have also proposed an adaptive algorithm that determines the moments when to stop transmission depending on which channel was lost, and have shown by simulations that this algorithm provides the highest throughput.

- After having studied separately the models for delay and throughput in Chapters 2 and 3 for interweave mode of spectrum access, in Chapter 4 we have proposed queueing analytic models for the delay analysis of interweave and underlay spectrum access, and we validated them against realistic scenarios. Besides that, we have proposed models relying on renewal-reward theory to determine the average throughput in both modes. We have also provided the bounds on the scanning times for which the interweave access outperforms the underlay. To further improve the performance in terms of both the throughput and delay, we have proposed dynamic policies, and have shown by simulations that the gains can go up to 50%.

In the second part of this thesis, we have proposed models for performance analysis of mobile data offloading for both types, *on-the-spot*, and *delayed* offloading.

- In Chapter 5, we have proposed a queueing analytic model for the performance of on-the-spot mobile data offloading for generic number of access technologies, and we validated it against realistic WiFi network availability statistics. We have provided approximations for different utilization regions and have validated their accuracy compared to simulations and to exact theoretical results. We also showed that our model can be applied to a broader class of distributions for the durations of the periods with and without WiFi availability. Our model can provide insight on the offloading gains by using on-the-spot mobile data offloading in terms of both offloading efficiency and delay. We have shown that the availability ratio of WiFi connectivity, in conjunction with the arrival rate plays a crucial role for the performance of offloading, as experienced by the user.

- Finally, in Chapter 6 we have proposed a queueing analytic model for the performance of delayed mobile data offloading, and have validated it against realistic WiFi network availability statistics. We have also considered a number of scenarios where one or more of our model's assumptions do not hold, and have observed acceptable accuracy, in terms of predicting the system delay as a function of the user's patience. Finally, we have also shown how to tune the maximum deadlines, in order to solve various optimization problems involving the system delay, monetary cost, and energy consumption.

7.1 Future work

While in the area of cognitive radio networks, there has been a lot of research in the past years, the same cannot be said about the research in mobile data offloading, which as a research topic is much more recent. Actually, in terms of performance modeling of mobile data offloading, there are only few proposed models. However, they are either so simple and very unrealistic, or their solutions can be found only numerically. Having proposed the models for both cognitive networks and mobile data offloading, the future research directions we are planning to focus on are enlisted below.

In the area of cognitive radio networks, the future work will turn around the following points:

- We intend to extend our model of Chapter 2 to include multihop cognitive radio networks, and also to use our results in designing better spectrum management, resource allocation and scheduling algorithms.
- As an extension of Chapter 3 we intend to extend our theoretical analysis to capture generic OFF periods, as well as to consider joint scanning and sequencing optimization, which will further enhance the performance of interweave cognitive radio networks.
- For Chapters 3 and 4, we plan to extend our model to capture generic file sizes, and generic ON periods. We are also working on optimizing the power consumption for underlay and interweave spectrum access in CRN.

For mobile data offloading, the future research steps are as following:

- The model we used for on-the-spot and delayed offloading was based on the exponential assumptions for ON and OFF periods. Although in general our models are able to predict the performance of a system for heavy-tailed ON and OFF periods, there are scenarios under which the discrepancy will be quite high. Hence, we are working on more complex models, in which the ON and OFF periods will be subject of heavy-tailed distributions for both on-the-spot and delayed offloading, with file sizes following generic distributions as well.
- We also intend to consider more complex models for both WiFi and cellular queues, as well as per-flow scheduling and dispatch policies.
- The other interesting open problem is to choose deadlines to be proportional to file sizes, and then to find the average file delay in such a system, as well as the percentage of files that will renege.

Chapter 8

Résumé

Les dernières années ont connu une demande croissante de spectre pour les applications sans fil en réponse à la propagation rapide des ordinateurs portables, smartphones et tablettes dans les marchés de la technologie ainsi que les applications qu'ils fournissent étant très gourmandes en bande large. Cela est devenu une préoccupation majeure pour les opérateurs de réseaux mobiles, contraints de travailler près de leurs limites de capacité. En raison des politiques d'attribution du spectre statique, la question de la rareté du spectre est devenue un problème majeur dans l'industrie sans fil actuelle. D'autre part, les mesures de l'utilisation du spectre sans fil sous licence ont révélé que le spectre disponible est plutôt sous-utilisé, présentant une grande variabilité dans l'espace, la fréquence et le temps [1]. Donc, en plus de l'augmentation rapide de la demande, le manque actuel de flexibilité dans l'affectation dynamique du spectre à correspondre aux demandes de temps et d'espace, limite les niveaux de service offerts.

Récemment, différentes solutions ont été proposées pour résoudre ce problème. Une solution possible pour atténuer ces problèmes est l'utilisation de l'accès dynamique du spectre (DSA), avec la radio cognitive (CR) comme la technologie clé [2]. À cette fin, les réseaux radio cognitive (CRN) ont été proposées pour découvrir et exploiter de manière opportuniste les bandes de spectre sous licence (temporairement) inutilisés, dans lequel l'activité des utilisateurs secondaires (SU) est subordonnée aux utilisateurs primaires (PU). Le SU devrait ajuster ses communications, afin qu'il n'y ait aucune perte de valeur à la qualité de service PU. Afin d'atteindre cet objectif, les radios cognitives doivent être équipés avec quelques fonctionnalités supplémentaires, telles que: la détection du spectre, la décision du spectre, le partage du spectre, et la mobilité du spectre.

Basé sur le caractère subordonné des radios cognitives, l'accès au spectre dynamique conduit à un modèle où l'accès de l'utilisateur est intermittente, avec des formes de disponibilités relevant d'un processus aléatoire, affectant les performances de manière non-triviales, et en fonction d'un grand nombre de paramètres.

Un autre moyen important pour faire face à la crise de données, bénéfique à la fois pour les opérateurs cellulaire et les utilisateurs mobiles, est d'utiliser les réseaux hétérogènes (HetNets) comprenant de petites cellules (Femtocells, picocells) et/ou des réseaux WiFi, ainsi que le déchargement agressif. Il y a deux types de déchargement que nous considérons dans cette thèse: *on-the-spot* et *déchargement retardé.*

Les réseaux hétérogènes et le déchargement conduisent également à un accès intermittent (à une ou plusieurs technologies d'accès sans fil), qui est aussi un sujet lié à l'hasard, et peuvent être modélisés d'une manière similaire comme le cas cognitif. Dans cette thèse, nous avons modélisé ce type d'accès avec des processus alternant de renouvellement, et nous avons optimisé la performance à chaque fois que cela était possible.

Les réseaux doivent fournir un certain niveau de QoS à leurs applications. Dans le cas des applications sensible en retard, comme VoIP, streaming audio et vidéo en direct, téléconférence, etc., le délai induit à un paquet transmis par un utilisateur secondaire est d'une importance énorme. Il en va de même pour un utilisateur mobile, qui est en attente de trouver un WiFi AP, et à transmettre/recevoir des données via cette interface. Comme prévu, la performance d'un utilisateur intermittent est déterminée principalement par l'activité des utilisateurs sous licence dans le "voisinage" (pourcentage du temps permettant de communiquer, la durée des périodes de disponibilité, la durée de salves de trafic, etc.). Pour déterminer quantitativement la performances sur un réseau, nous avons besoin de modèles réalistes. Ceci est important pour la conception de protocoles efficaces pour les utilisateurs intermittents, entre autres.

Ainsi, dans cette thèse, nous avons proposé des modèles qui nous permettent d'effectuer une analyse des réseaux de radiocommunication cognitifs et le déchargement des données mobiles. Cela nous permettra de comprendre dans quelle point la performance des utilisateurs secondaires dépendra d'un certain nombre de paramètres dans un réseau de radio cognitive, ainsi que de vérifier comment la performance dans un système de déchargement mobile peut être améliorée.

8.1 Motivation et contributions de la thèse

Les réseaux sans fil doivent fournir un certain niveau de QoS à ses utilisateurs et applications. Quantitativement, cette QoS est exprimé à travers différents paramètres, tels que le délai moyen de paquet, le débit moyen, la gigue, la bande passante, etc. Parmi les figures les plus importantes d'intérêt se trouvent *le délai* et *le débit*. En fonction de la nature de l'application prévue, l'une ou l'autre peut être plus importante. Par exemple, dans le cas d'applications sensibles au retard, comme la VoIP, le streaming audio et vidéo en direct, la téléconférence, etc., le retard induit à un paquet transmis ou reçu est très important. De même, un certain nombre d'applications d'intérêt, tels que: le téléchargement de fichiers, l'échange de fichiers peer-to-peer, le streaming, etc., sont souvent sensibles au débit et pour eux un paramètre très important est le débit moyen.

Comme nous l'avons vu dans le paragraphe précédent, il est extrêmement important de connaître les principaux paramètres que prend une application dans un réseau. Ceci est encore plus souligné dans les réseaux intermittents où les utilisateurs mobiles ne détiennent pas de façon permanente la possibilité de communiquer, et peuvent soit transmettre ou recevoir des données qu'occasionnellement, ou ne peuvent pas faire usage de tout le potentiel qu'ils possèdent durant certaines périodes. Par conséquent, l'évaluation de la performance du réseau sans fil intermittent est une des tâches les plus importantes dans la conception des réseaux mobiles.

Dans ce qui suit, nous présentons les principales raisons pour lesquelles il est important d'effectuer une analyse à la fois des réseaux de radio cognitive et du déchargement des données mobiles.

Les réseaux de radio cognitive. Comme prévu, la performance d'un utilisateur cognitif est déterminée principalement par l'activité des utilisateurs autorisés dans le "voisinage" (pourcentage de temps d'inactivité, la durée des périodes de repos, la durée des pics de trafic, etc.). Dans un CRN, lorsqu'il s'agit d'applications sensibles au retard, il est très important de déterminer quantitativement le retard moyen d'un paquet CR ou d'un fichier dans un tel réseau, et de voir en fonction du résultat si la demande de QoS pour une application CR est satisfaite. La même procédure vaut pour les applications de débit sensible pour lesquelles il est très important de déterminer le débit moyen.

Une autre question importante qui fait surface est de savoir dans quelle mesure le SU devrait réduire sa puissance d'émission dans l'accès underlay au spectre. Pour un fonctionnement normal, il y a un niveau maximum d'interférence qu'un PU peut tolérer [13]. Le SU devrait connaître cette valeur ainsi que la distance à partir du PU. La puissance maximale d'émission autorisée est ensuite calculé comme la somme du budget de l'interférence, la perte en espace libre à partir de la SU jusqu'à la PU et d'autres types de pertes spécifiques aux communications sans fil, tels que fading et l'ombrage [14].

A côté des questions mentionnées ci-dessus, une autre question importante pourrait émerger. À savoir, si un CR peut utiliser à la fois les modes *underlay* et *interweave* d'accès au spectre, lequel devrait-il utiliser de sorte qu'un certain paramètre (délai ou débit) soit optimisé? Or, s'il est possible de passer d'un mode à l'autre, comment et sous quelles conditions la mise en oeuvre de cet hybride impleméntation peut-elle améliorer la performance globale?

Le déchargement des données mobiles. Pour un utilisateur avec la possibilité de commuter entre les interfaces en utilisant soit on-the-spot soit déchargement retardé de données, il est d'une importance cruciale d'évaluer correctement ses paramètres de QoS. Un utilisateur aimerait savoir combien de temps supplémentaire il devrait attendre en utilisant le déchargement retardé. Quels sont les gains, en termes de réduction de retard s'il utilise à l'occasion une interface WiFi avec un débit de données plus élevé? Combien devra-t-il payer en moins s'il utilise un système de déchargement? Quelles sont les économies en termes de consommation d'énergie de la batterie? A quel point l'utilisateur devra-il être tolérant aux retards afin de retirer profit du déchargement?

Des questions d'intérêt similaire peuvent provenir du côté de l'opérateur cellulaire également. Néanmoins, le paramètre le plus important pour l'opérateur est *l'efficacité de déchargement*, qui est le pourcentage de données qu'un utilisateur transmet via l'interface WiFi, à savoir le pourcentage de données dont l'opérateur se débarrasse de son BS.

La première étape dans l'évaluation de la performance des réseaux sans fil est d'avoir *les modèles* appropriés, qui sont réalistes et peuvent capturer dans une grande mesure le comportement des différents facteurs et leur nature d'interaction dans un système donné. Avec les modèles et analyses appropriés, les résultats obtenus peuvent en outre être utilisés pour la conception de planification efficace du spectre, le scanning, et les stratégies de transfert, entre autres.

Dans la section suivante, nous résumons les contributions et nous présentons les grandes lignes de la thèse.

8.2 Contribution et bref aperçu

L'objectif de cette thèse est basé sur l'évaluation analytique de la performance des utilisateurs mobiles avec un accès intermittent aux ressources. Ceci est très important car, en faisant cela, nous sommes en mesure de conclure si les utilisateurs mobiles peuvent atteindre les QoS imposées par les différents types d'applications qu'ils soutiennent. Nous considérons deux types de cette paradigme: *les réseaux de radio cognitive* et *le déchargement des données mobiles.* Dans les réseaux de radio cognitive, la performance de l'utilisateur secondaire dépend fortement du type de l'activité de l'utilisateur principal (i.e. comment la durée de la communication est-elle variable lorsqu'elle a des données à envoyer/recevoir), le PU duty cycle (c'est-à-dire quel

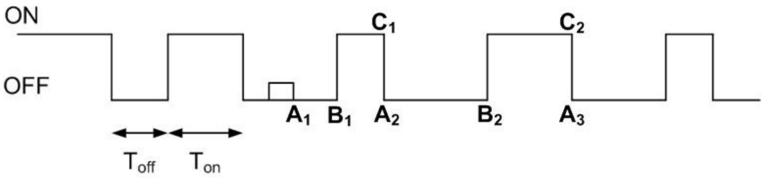

Figure 8.1: Le modèle d'activité d'un utilisateur principal (PU).

pourcentage de temps occupe-t-elle le canal), ainsi que sur les caractéristiques du trafic SU (intensité, taille du paquet). De même, pour un utilisateur mobile prêt à exploiter la fonction de déchargement de données via WiFi AP, la performance dépend de la couverture avec les points d'accès, le nombre d'utilisateurs, l'intensité du trafic, et la taille des fichiers. Par conséquent, notre objectif principal dans cette thèse est de proposer des modèles qui nous permettent de comprendre comment différents facteurs et dans quelle mesure peuvent affecter les performances des utilisateurs mobiles d'intérêt. Ceci nous permettra plus tard de trouver la technique la plus appropriée qu'un utilisateur mobile doit prendre en compte afin d'atteindre des performances optimales, en résolvant différents problèmes d'optimisation.

Un aperçu de la thèse est fourni ci-dessous et les contributions réalisées dans chaque chapitre sont résumées.

Chapitre 2 - Qui m'a interrompu? Analyse de l'effet de l'activité PU sur la performance des utilisateurs cognitifs.

La première étape afin de mieux comprendre la performance des utilisateurs cognitifs dans un CRN est d'avoir les modèles adéquats. Notre objectif dans ce chapitre est d'analyser le retard moyen encouru à un paquet SU. Nous modélisons l'activité PU avec un processus de renouvellement alterné (Fig. 8.1), dans laquelle les périodes ON représentent les intervalles de temps lorsque le PU est actif sur le canal, tandis que les périodes OFF quand SU peuvent transmettre.

Nous utilisons la théorie des files d'attente pour dériver le délai de paquet prévu. Nous avons montré qu'il existe deux types de paquets qui ont des distributions différentes de temps de service. Le premier, connu sous le nom des paquets S', sont ceux qui initient une période chargée et leur temps de service moyen se présente comme suit

$$\begin{aligned} E[S'] &= \Delta + \frac{P}{1-p_1p_2}\left(\Delta \cdot p_2 + \left(e^{\lambda\Delta} - 1\right)\left(\Delta + E[T_1] - \frac{1}{\lambda}\right)\right) \\ &+ \frac{P}{1-p_1p_2}\left(E[T_{ON}] + \left(E[T_2] - \frac{1}{\lambda}\right)(1-p_2)\right). \end{aligned} \tag{8.1}$$

Tous les paramètres apparaissant dans l'équation précédente et les formules à suivre sont expliqués dans le tableau 8.1.

Table 8.1: Les variables et la notation abrégée.

Variable	Definition/Description
Δ	La taille du paquet
λ	Le taux d'arrivée (Processus de Poisson)
T_{ON}	La durée des périodes ON
T_{OFF}	La durée des périodes OFF
$T_{OFF}^{(f)}$	$T_{OFF}\|T_{OFF} < \Delta$: Période OFF d'une durée inférieure à Δ
$T_{OFF}^{(e)}$	OFF period restant
$f_{OFF}^{(e)}(x)$	$\frac{1-F_{OFF}(x)}{E[T_{OFF}]}$
p_1	$\int_0^\infty e^{-\lambda t_{off}} f_{OFF}(t_{off}) dt_{off}$
p_2	$\int_0^\infty e^{-\lambda t_{on}} f_{ON}(t_{on}) dt_{on}$
P	$\int_0^\infty e^{-\lambda t_{off,e}} f_{OFF}^{(e)}(t_{off,e}) dt_{off,e}$
p	$p = P\left[T_{OFF} > \Delta\right]$
N	Nombre de cycles (extra) ON - OFF jusqu'à la réussite de la transmission
T_1	La durée des cycles ON - OFF (pour le paquet arrivant dans la période ON)
T_2	La durée des cycles ON - OFF (pour le paquet arrivant dans la période OFF)

Le deuxième moment du type de paquet S' se présente comme suit

$$\begin{aligned} E[S'^2] &= \Delta^2 + \frac{P}{1-p_1p_2}\left(\Delta^2 p_2 + 2\Delta e^{\lambda\Delta}\left(E[T_1]-\frac{1}{\lambda}\right) + \left(e^{\lambda\Delta}-1\right)\left(\left(E[T_1]-\frac{1}{\lambda}\right)^2 + \frac{1}{\lambda^2} + \Delta^2\right)\right) \\ + & \frac{P}{1-p_1p_2}\left(E[T_{ON}^2] + 2E[T_{ON}]\left(E[T_2]-\frac{1}{\lambda}\right) + \left(\left(E[T_2]-\frac{1}{\lambda}\right)^2 + \frac{1}{\lambda^2}\right)(1-p_2)\right). \end{aligned} \tag{8.2}$$

Les autres type de paquets, S'', sont ceux qui permettent de trouver d'autres paquets déjà en file d'attente à leur arrivée. Le temps de service moyen de paquets S'' est

$$E[S''] = \Delta + \int_0^\Delta x f_{OFF}^{(e)}(x)dx + \frac{1}{p}\left(E[T_{ON}] + E\left[T_{OFF} \mid T_{OFF} < \Delta\right]\right)\int_0^\Delta f_{OFF}^{(e)}(x)dx. \tag{8.3}$$

Le deuxième moment du type de paquet S'' se présente comme suit

$$E[S''^2] = \Omega_1 + \Omega_2, \tag{8.4}$$

où

$$\Omega_1 = \Delta^2 + 2\Delta\int_0^\Delta x f_{OFF}^{(e)}(x)dx + 2\Delta\int_0^\Delta f_{OFF}^{(e)}(x)dx\left(\frac{1}{p}E[T_{ON}] + \left(\frac{1}{p}-1\right)E\left[T_{OFF} \mid T_{OFF} < \Delta\right]\right)$$

et

$$\begin{aligned} \Omega_2 &= \int_0^\Delta x^2 f_{OFF}^{(e)}(x)dx + 2\int_0^\Delta x f_{OFF}^{(e)}(x)dx\left(\frac{1}{p}E[T_{ON}] + \left(\frac{1}{p}-1\right)E[T_{OFF} \mid T_{OFF} < \Delta]\right) \\ + & \int_0^\Delta f_{OFF}^{(e)}(x)dx\left(\frac{2(1-p)}{p^2}E[T_{ON}]^2 + \frac{1}{p}E[T_{ON}^2] + \frac{2}{p}E[T_{ON}]\left(\frac{1}{p}-1\right)E[T_{OFF} \mid T_{OFF} < \Delta]\right) \\ + & \int_0^\Delta f_{OFF}^{(e)}(x)dx\left(\frac{(1-p)(2-p)}{p^2}E[T_{OFF} \mid T_{OFF} < \Delta]^2 + \left(\frac{1}{p}-1\right)Var\left(T_{OFF} \mid T_{OFF} < \Delta\right)\right) \end{aligned}$$

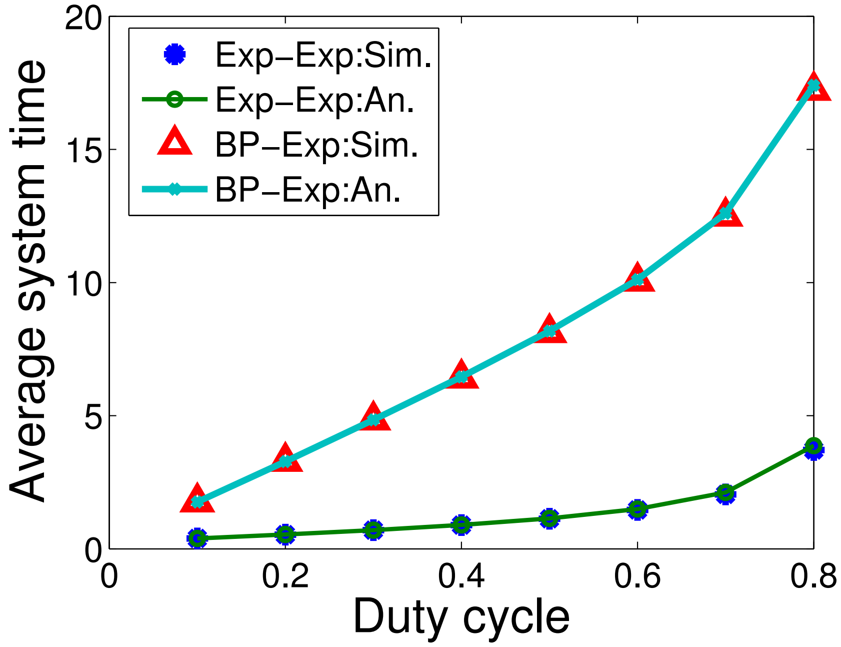

Figure 8.2: Le délai de système pour exp-OFF.

Pour calculer le délai d'attente dans ce système, nous utilisons la théorie de renouvellement-récompense. Le temps d'attente moyen est donné par

$$E[T_Q] = \frac{\lambda E[S''^2]}{2(1-\lambda E[S''])} + \frac{\lambda\left(E[S'^2]-E[S''^2]\right)}{2+2\lambda\left(E[S']-E[S'']\right)}. \tag{8.5}$$

Enfin, pour la durée totale du système, nous obtenons

$$E[T] = \frac{E[S']}{1+\lambda\left(E[S']-E[S'']\right)} + \frac{\lambda E[S''^2]}{2(1-\lambda E[S''])} + \frac{\lambda\left(E[S'^2]-E[S''^2]\right)}{2+2\lambda\left(E[S']-E[S'']\right)}. \tag{8.6}$$

Un avantage de notre modèle est qu'il convient pour des tailles de paquets génériques. Notre modèle conduit finalement à un résultat de forme fermée et montre une interaction complexe entre les paramètres de réseau et du traffic.

Fig. 8.2 montre le delai moyen de paquets dans un réseau cognitive pour deux PUs avec different scenarios d'activité pour des OFF périodes avec une distribution exponentielle. Le taux d'arrivée est $\lambda = 0.1$. Pour les distributions exp-exp, une activité faible de l'utilisateur primaire (cycle de service de 0.2) donne une utilisation de 0.05. Lorsque le cycle de service est 0.8, l'utilisation est 0.29. Pour les périodes ON avec distribution de Pareto délimitée (BP), l'activité inférieure des utilisateurs primaires de 0.2 correspond à une utilisation de 0.09, et une plus grande activité de l'utilisateur autorisé de 0.8 correspond à une utilisation du serveur de 0.41. Différentes ampleurs de cycle de service donnent différents niveaux d'utilisation, puisque le temps moyen de service dépend des valeurs de $E[T_{ON}]$ et $E[T_{OFF}]$, i.e. du cycle de service. La première chose à observer est une bonne accord entre la théorie et les simulations.

En outre, nous pouvons également constater qu'un cycle de service supérieur implique des délais plus élevés, ce qui est cohérent puisque plus le cycle de service est élevé, plus l'utilisateur

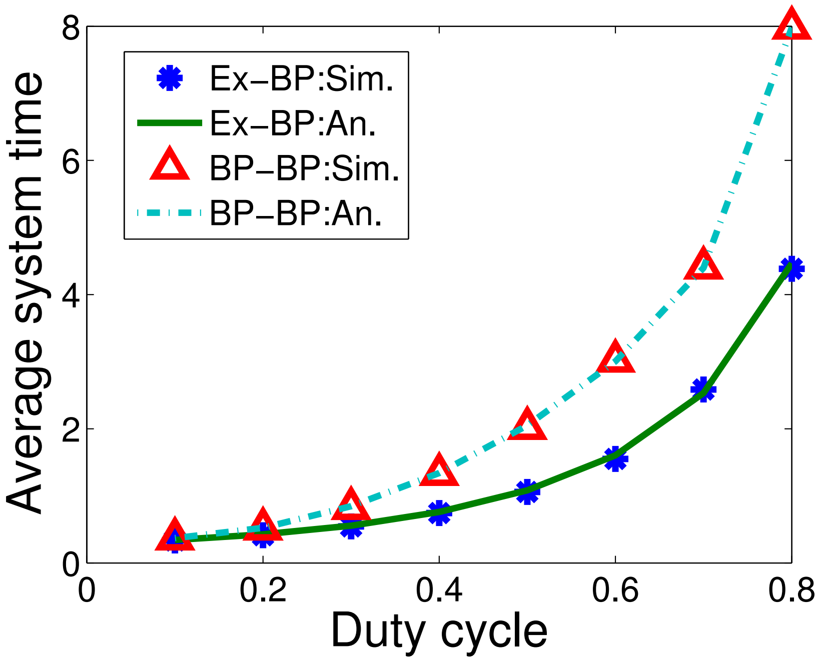

Figure 8.3: Le délai de système pour BP-OFF.

primaire est actif, et moins il y a de temps pour l'utilisateur cognitif de fonctionner. Nous pouvons également voir que, pour les mêmes moyennes de durées ON et OFF (le même cycle de service), les délais sont plus élevés lorsque l'utilisateur primaire a des périodes d'occupations avec une plus grande variabilité. Ceci est la première conclusion intéressante qui émerge de notre modèle. Malgré le fait que les deux canaux se ressemble, du point de vue de l'activité moyenne PU, la variabilité peut affecter davantage les délais. Cela rappelle le paradoxe d'inspection [20], bien que la dynamique des équations Eq.(8.1) et Eq.(8.3) sont en fait plus complexe.

La figure 8.3 montre les délais de paquets pour des périodes OFF qui sont distribués comme BP. Le taux d'arrivée est faible (0.01). Ce taux d'arrivée correspond à la circulation clairsemée. Comme nous pouvons observer sur la figure 8.3, il y a également une bonne relation entre la théorie et les simulations pour les périodes OFF distribuées génériquement.

Dans ce chapitre, on montre également que la variabilité de l'activité de l'utilisateur primaire peut être beaucoup plus important que le cycle de service lui-même, et ce en fonction du coefficient de variation des périodes occupé de PU, il est sous-entendu que les canaux, même avec un très faible cycle de service et une forte variabilité peuvent céder à des délais supérieurs à ceux des canaux avec cycle de service très élevé, mais avec une très faible variabilité. Ce résultat est une conséquence très intéressant de notre modèle qui peut être utilisé lors de la conception des algorithmes d'ordonnancement. Dans ce sens, nous comparons deux canaux. Nous considérons un canal A avec une moyenne haute activité de PU (cycle de service de 0.6) et avec durées des ON périodes exponentiellement distribué de variance égale à 1. Nous l'avons mis contre un canal B avec une activité beaucoup plus faible de PU (cycle de service de 0.3), mais une distribution logarithmique normale de périodes ON (qui ont une queue plus lourde qu'exponentielle). En gardent la moyenne des périodes ON inchangée et en augmentant la variance de la distribution log-normale nous donne un aperçu intéressant de l'effet à la fois de l'activité moyenne PU et de la variabilité de la performance de l'utilisateur cognitive.

Le tableau 8.2 montre le rapport de délai SU sur le canal B vs. le délai SU sur le canal A.

Table 8.2: Le rapport des délais pour deux canaux différents.

Variance pour CH B	1	10	20	30	40	50	60
Le rapport des délais	0.5	1.2	1.8	2.3	2.6	3	3.2

Nous pouvons observer à partir du tableau 8.2 que pour une variance similaire, le canal B qui est moins "occupé" est préférable. Cependant, en augmentant la variance, le ratio ne cesse de croître et dépasse 3 pour un variance à 50. En fait, en théorie, cette différence peut devenir arbitrairement élevé (par exemple pour de véritables distributions à queue lourde, comme Pareto avec le paramètre < 2). Nous pouvons donc conclure que, contrairement à la croyance commune qui tend à considérer une partie du spectre sous-utilisé comme "bon spectre" pour les utilisateurs cognitifs, l'impact de la variabilité de l'activité de l'utilisateur primaire est beaucoup plus important que le cycle de service lui-même. Cela pourrait être très important en ce qui concerne par exemple les applications sensibles au retard mais à faible débit, comme M2M.

Comme une implication de notre modèle, un schéma de planification simple peut être proposé. Notamment, nous avons prouvé par des simulations que l'allocation de canal fait individuellement pour chaque flux basé sur notre formule analytique utilisée pour prédire le délai sur un canal donné, fournit de meilleurs résultats que l'attribution des canaux de l'ensemble possible en les classant en fonction de leurs probabilités d'inactivité ou en fonction de la moyenne des durées OFF excessives.

Les articles liés à ce chapitre sont:

- *F. Mehmeti, T. Spyropoulos, "Who Interrupted Me? Analyzing the Effect of PU Activity on Cognitive User Performance", in Proc. of IEEE International Conference on Communications (IEEE ICC 2013), Budapest, Hungary, 2013.*

- *F. Mehmeti, T. Spyropoulos, "Analysis of Cognitive User Performance Under Generic Primary User Activity", Tech. Report, RR-12-274, Eurecom, 2012.*

Chapitre 3 - Scanner ou ne pas scanner: L' Effet de l'hétérogénéité du canal sur les règles de scannage optimal.

Alors que dans le Chapitre 2 nous nous sommes intéressés à la performance de SU en termes de délai, dans le Chapitre 3 notre principal intérêt réside sur l'analyse de débit moyen et de son optimisation. Nous considérons le problème du scannage dans les scénarios lorsque plusieurs canaux sont regroupés. Plus important, à la place d'optimiser le scannage de la séquence lorsque le scannage est déclenche, nous étudions le problème complémentaire d'optimisation de quand déclencher cette fonction de scannage afin d'optimiser le débit qui peut être maintenu par le SU. Nous introduisons la notion d'une valeur de seuil comme le nombre de canaux que nous sommes autorisés à perdre, avant l'initialisation de la procédure de scannage. Cette valeur seuil donne les meilleurs résultats en termes de débit. Il est prouvé qu'il n'existe pas de telles valeur de seuil pour le cas de canaux homogènes (IID) s'il n'y a aucun coût initial devant être payé au début du processus de scannage. La valeur initiale de scannage (T_0) pour laquelle il est préférable de

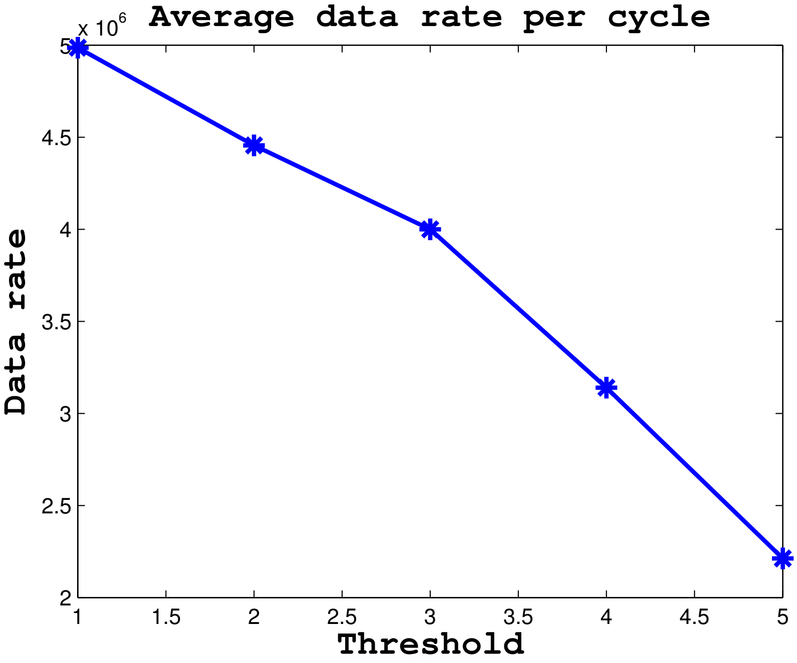

Figure 8.4: Canaux homogènes exponentiels.

poursuivre la transmission doit satisfaire la condition suivante

$$T_0 \geq \frac{1}{N(N-1)\lambda_{off}}. \tag{8.7}$$

Dans la dernière expression, N représente le nombre de canaux utilisés, et λ_{off} est l'inverse de la durée moyenne de la période disponible. La Fig. 8.4 montre le débit moyen pour une radio qui utilise $N = 5$ canaux pour transmettre ses données. Les périodes ON pour les activités des utilisateurs primaires dans tous les canaux sont identiques et choisis à partir d'une distribution uniforme dans la gamme entre 10 et 20 s^{-1}. Les périodes OFF sont également identiques pour tous les canaux avec la durée moyenne de 1 s. Le nombre de canaux de secours est assez grand. L'axe des x indique la valeur de seuil utilisée dans chaque cas. À partir de la Fig. 8.4 nous pouvons observer que le meilleur résultat est obtenu si nous commençons le scannage immédiatement après avoir perdu le premier canal et en augmentant la valeur de seuil, le moyen débit baisse.

Fig. 8.5 illustre la dépendance de débit moyen par cycle sur la valeur de seuil pour distribution exponentielle (IID) des périodes OFF où $\lambda_{OFF} = 1s^{-1}$. Nous montrons l'effet d'un plus grand nombre de canaux utilisés ($N = 10$). Les autres paramètres sont identiques comme pour le scénario de la Fig. 8.4. Ici aussi, la courbe prouve notre affirmation selon laquelle il n'y a pas de valeur seuil qui donne de meilleurs résultats en termes de débit pour canaux i.i.d. exponentiels.

Néanmoins, cette valeur existe pour les canaux indépendants hétérogènes et sa valeur dépend de la variabilité des périodes OFF des canaux d'utilisation. La condition pour cela se trouve dans le résultat suivant:

Result 17. *Pour les périodes OFF distribuées de façon exponentielle hétérogènes, il n'est pas*

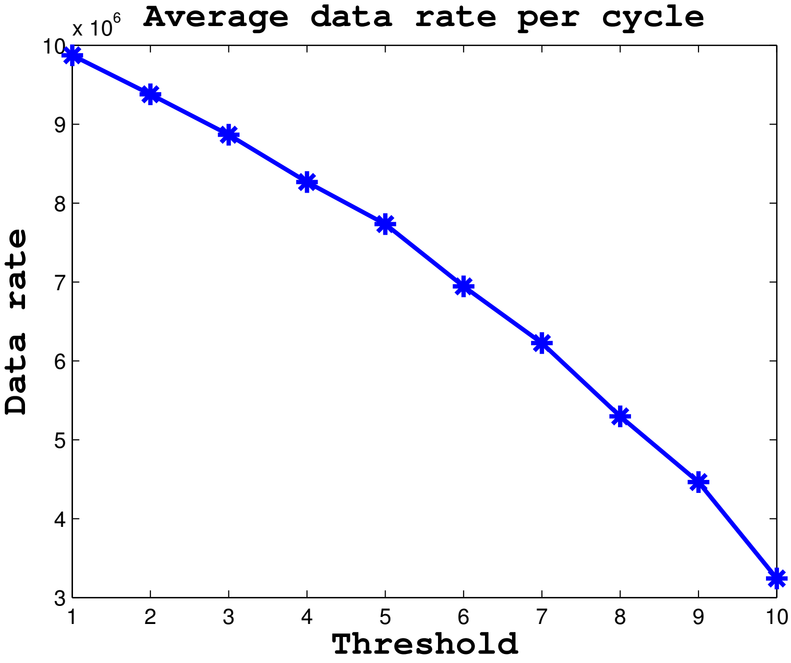

Figure 8.5: Canaux homogènes exponentiels.

toujours optimal de scanner immédiatement. Au lieu de cela, la transmission doit procéder avec des canaux restants, si elles satisfont la relation suivante:

$$c_\lambda^2 \geq 1, \tag{8.8}$$

où c_λ est le coefficient de variation des OFF périodes des canaux restant disponibles.

Il y a deux choses intéressantes à noter à propos du résultat ci-dessus. Tout d'abord, contrairement au cas du canal homogène, le scannage est immédiatement sous-optimale et un meilleur seuil que $L = 1$ peut être trouvé. Deuxièmement, ce seuil augmente lorsque la variabilité de l'amas de canaux disponibles pour la SU augmente.

Nous proposons également un algorithme adaptatif qui détermine les instants où arrêter la transmission en fonction du canal perdu. À savoir, si un canal qui est préférable (en termes de la durée moyenne plus grande des périodes OFF) que la moyenne de l'amas de canal est perdu, nous montrons qu'il est préférable d'arrêter le processus de transmission et d'engager la procédure de scannage, et vice versa. Nous montrons par des simulations que cet algorithme fournit le plus haut débit. Le SU devrait continuer avec la transmission si le canal perdu satisfait la relation suivante

$$E[\lambda_{lost}] \geq \frac{1 + T_s \sum_i \lambda_i}{NT_s}, \tag{8.9}$$

où T_s est le temps moyen de scannage. Un cas particulier serait si la relation existe $T_s \sum_i \lambda_i >> 1$. Alors, Eq.(8.9) se réduit à

$$E[\lambda_{lost}] \geq \frac{\sum_i \lambda_i}{N}. \tag{8.10}$$

Ensuite, si le canal perdu est plus mauvais (avec λ_{off} plus grande) que la moyenne des canaux utilisés, alors il est préférable de reprendre la transmission. Ceci est plutôt intuitif, car en se débarrassant des canaux à faible disponibilité, nous sommes laissés seulement avec ceux à la haute disponibilité, et si nous décidons de scanner à la place, le nouveau canal devrait seulement faire le debit moyen de l'amas de canal plus mauvais. D'autre part, si nous perdons un bon canal alors nous devrions interrompre, puisque le reste des canaux sont probablement plus mauvais que la moyenne et le scannage pourrait améliorer cette situation.

Fig. 8.6 illustre les avantages de l'utilisation du règle (*policy-politique*) d'adaptation (en ligne). Le nombre de canaux utilisés est 15. Les taux des périodes OFF sont tirées de la distribution uniforme dans l'intervalle de 0.1 à 100 s^{-1}. La Fig. 8.6 montre les débits pour trois cas:

1. Seuil $L = 1$,
2. Le seuil idéal pour l'algorithme *offline*, and
3. Algorithme *online* (adaptif) avec le seuil variable.

Le (offline) seuil idéal qui fournit des taux de données maximum (cas 2) est $L = 5$. À partir de la Fig. 8.6 nous pouvons observer que notre algorithme en ligne (online) proposée prévoit le débit de données plus élevé par cycle. Ceci est une conséquence du fait que nous prenons la décision de quand commencer le scannage à la volée, en fonction du canal que nous avons perdu. Pour l'algorithme hors ligne cependant, nous devons prendre la décision à l'avance, sur la base des caractéristiques moyennes de l'amas de canaux et de la qualité attendue des canaux perdus. Il peut arriver qu'un canal avec une longue durée moyenne soit perdu avant un mauvais canal, ce qui donne lieu à cette différence entre l'algorithme hors ligne et l'algorithme en ligne. Nous pouvons également observer que le débit est plus bas si nous déclenchons immédiatement le scannage.

L'article lié à ce chapitre est:

- *F. Mehmeti, T. Spyropoulos, "To Scan or Not To Scan: The Effect of Channel Heterogeneity on Optimal Scanning Policies", in Proc. of IEEE International Conference on Sensing, Communications, and Networking (IEEE SECON 2013), New Orleans, USA, 2013.*

Chapter 4 - Rester ou commuter? Analyse et comparaison de l'accès au spectre *Interweave* et *Underlay* dans les réseaux de radio cognitive.

Alors que dans le Chapitre 2, nous proposons un modèle pour trouver le délai moyen qui est générique, dans le Chapitre 4, en utilisant la théorie des files d'attente, basée sur les chaînes à deux dimensions de Markov, nous dérivons des expressions de forme fermée pour le délai prévu à l'accès *underlay* et *interweave* au spectre comme une fonction de paramètres clés du réseau (moyenne de temps d'inactivité de PU, les taux de transmission, les statistiques de temps de scannage) et les statistiques de trafic des utilisateurs (intensité du trafic, taille du fichier).

Pour le mode d'accès *interweave* au spectre avec distribution exponentielle du temps de scannage, nous utilisons la chaîne de Markov 2D de Fig. 8.7.

Le temps moyen du système dans ce cas est donné comme

$$E[T_{exp}] = \frac{\eta_H(\eta_H + \mu_H)(E[T_s])^2 + 2\eta_H E[T_s] + 1}{(1 + \eta_H E[T_s])(\mu_H - \lambda - \lambda\eta_H E[T_s])}. \tag{8.11}$$

Les variables utilisées dans l'expression précédente et dans le reste des équations de ce chapitre sont expliquées dans le tableau 8.3.

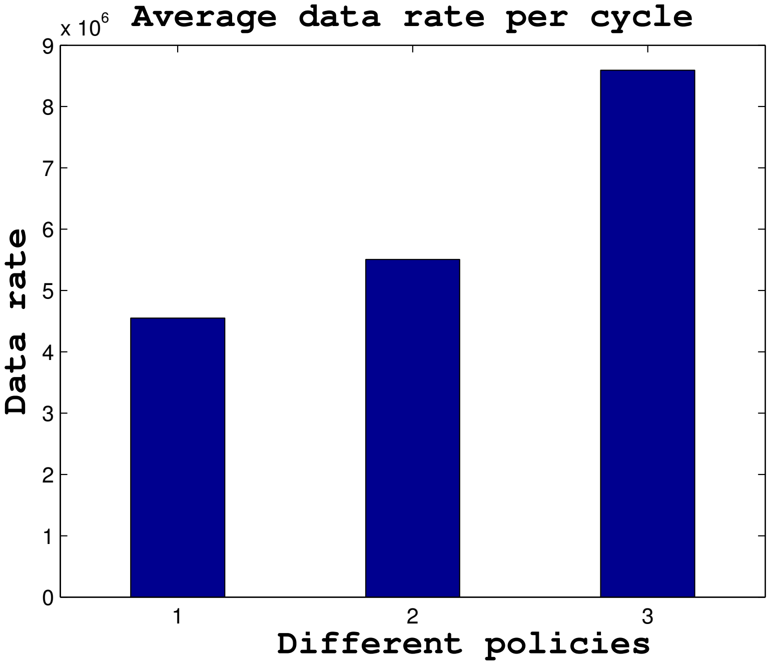

Figure 8.6: Le débit pour les différentes politiques (règles).

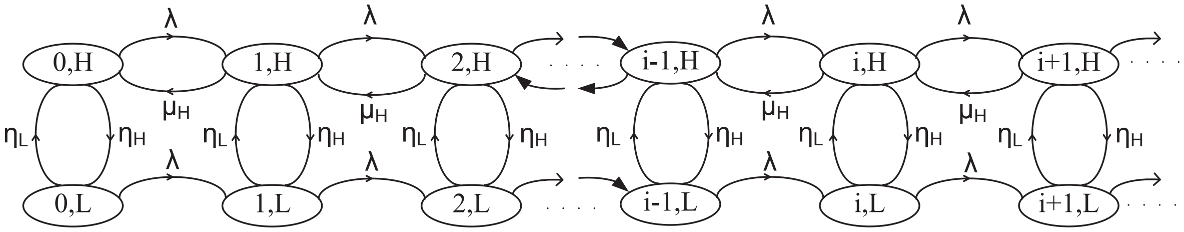

Figure 8.7: La chaîne de Markov 2D pour le temps de scannage avec distribution exponentielle.

Dans le cas où le temps de scannage est soumis à une distribution de faible variabilité, nous le modélisons avec une distribution Erlang avec k-stage, et la chaîne de Markov pour ce cas ci ressemble à Fig.8.8.

Dans ce cas, le délai moyen du système est donnée par

$$E[T_{erl}] = \frac{\eta_H \left[\eta_H + \frac{(k+1)}{2k}\mu_H\right] (E[T_s])^2 + 2\eta_H E[T_s] + 1}{(1 + \eta_H E[T_s])(\mu_H - \lambda - \lambda\eta_H E[T_s)]}. \tag{8.12}$$

Si le temps de scannage a une forte variabilité, alors nous le modélisons avec une distribution hyper-exponentielle. La chaîne 2D de Markov correspondante est représenté sur la Fig. 8.9.

Le délai moyen du fichier dans ce cas est donné comme

$$E[T_{hyp}] = \frac{(\eta_H E[T_s])^2 + \eta_H \mu_H \left(\frac{p}{\eta_L^2} + \frac{1-p}{\eta_V^2}\right) + 2\eta_H E[T_s] + 1}{(1 + \eta_H E[T_s])(\mu_H - \lambda - \lambda\eta_H E[T_s])}. \tag{8.13}$$

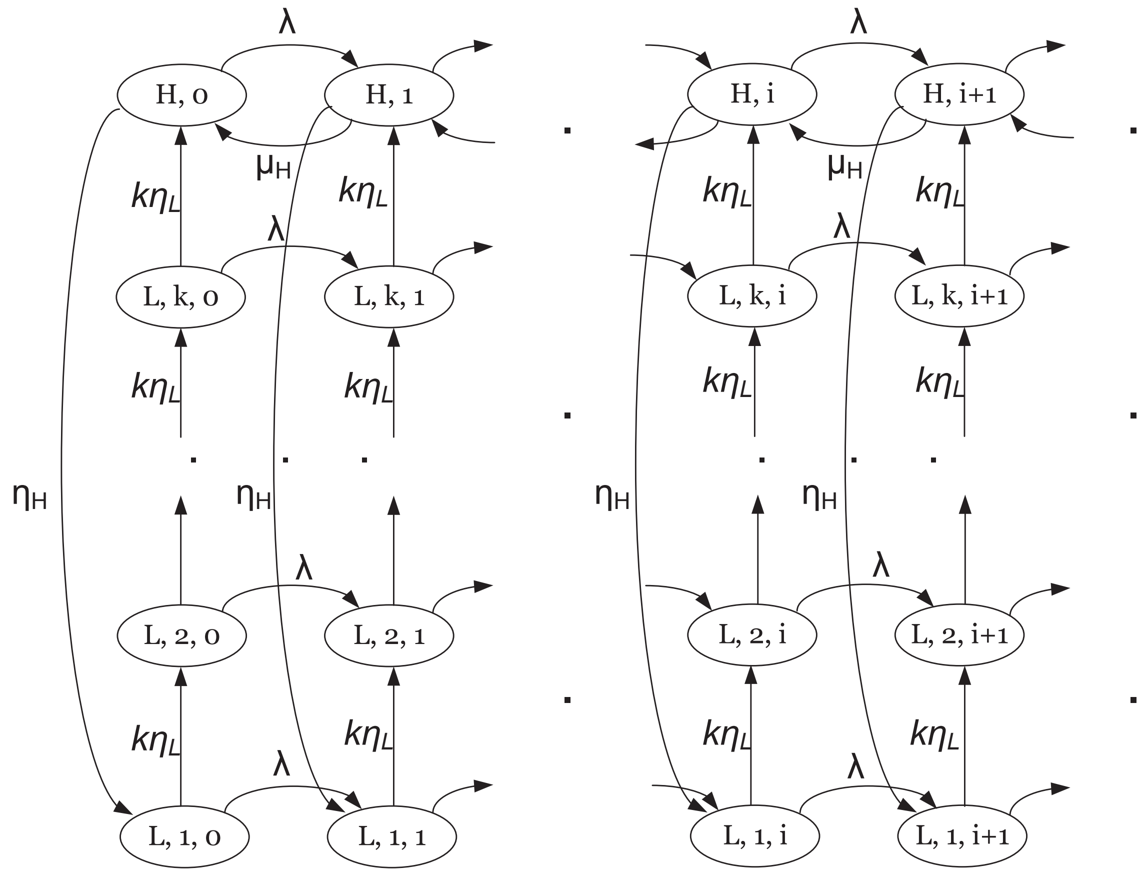

Figure 8.8: La chaîne de Markov 2D pour le temps de scannage Erlang distribué.

Table 8.3: Les variables et la notation abrégée.

Variable	Definition/Description
T_{ON}	Durée des périodes d'inactivité de PU
T_{OFF}	Durée de PU périodes occupés ou des temps de scannage
λ	Taux moyen d'arrivée de fichier à l'utilisateur mobile
$\pi_{i,L}$	Probabilité de trouver i fichiers dans le OFF (bas) période
$\pi_{i,H}$	Probabilité de trouver i fichiers dans le ON (haute) période
$\pi_{i,V}$	Probabilité de trouver i fichiers dans le OFF (état-V) période
$\eta_H(\eta_L)$	Le taux de quitter l'état ON (OFF)
η_V	Le taux de quitter l'état-V
$\mu_H(\mu_L)$	Le débit en une haute (bas) période
Δ	La taille moyenne de fichier
T_s	La durée de scannage

D'autre part, nous modélisons le mode *underlay* d'accès au spectre avec la chaîne 2D de Markov de Fig. 8.10.

Après quelques analyses, on obtient l'expression suivante pour le délai moyen du fichier

$$E[T_u] = \frac{\eta_H + \eta_L + \mu_H(1-\pi_{0,H}) + \mu_L(1-\pi_{0,L}) - \lambda + \frac{\mu_L\mu_H}{\lambda}(\pi_{0,L}+\pi_{0,H}-1)}{\mu_H\eta_L + \mu_L\eta_H - \lambda(\eta_H+\eta_L)}. \tag{8.14}$$

Après avoir présenté les formules pour le délai moyen des modes interweave et underlay, nous sommes en mesure de comparer les retards encourus dans chacun d'eux. Comme nous avons pu le remarquer, le délai dépend des statistiques de l'activité PU, du débit, de l'intensité du trafic, et du temps de scannage. Dans un premier scénario, nous supposons que le SU doit, au départ, décider lequel des modes d'accès à utiliser: *underlay* (c'est à dire toujours rester sur un même canal et transmettre avec la puissance autorisée), ou *interweave* (c'est à dire devenir silencieux chaque fois qu'un PU arrive sur le canal et scanner pour un nouveau). Pour ce cas, nous allons simplement nous référer à "la politique statique".

En général, pour l'accès interweave de surperformer l'accès underlay, le temps de scannage attendue $E[T_s]$ devrait être assez court pour assurer que le coût d'opportunité de la non transmission/recevabilité des données pendant un certain temps (qui est autorisé dans underlay) soit amortie par la découverte rapide d'un nouvel espace blanc. Dans le tableau 8.4, nous fournissons une expression analytique pour les valeurs maximale $E[T_s]$ pour lequelles l'accès interweave possède des délais inférieurs. Comme on peut le voir à partir du tableau 8.4, il existe une dépendance complexe sur les divers paramètres du système. Qui plus est, ce point "limite" dépend en outre de la variabilité du temps de scannage.

Enfin, nous proposons une politique «hybride» qui peut commuter entre les deux modes de manière dynamique afin d'améliorer encore les performances de délai. Nous effectuons les mêmes étapes pour le cas de débit (à long terme) pour les deux modes en utilisant la théorie de renouvellement-récompense, à savoir fournir des expressions en forme fermées pour la performance individuelle, les comparer analytiquement, et enfin les optimiser. Avec l'utilisation d'un large éventail de scénarios de simulation réalistes, nous validons largement nos prédictions analytiques, et nous explorons les conditions dans lesquelles la politique underlay ou interweave fonctionne mieux.

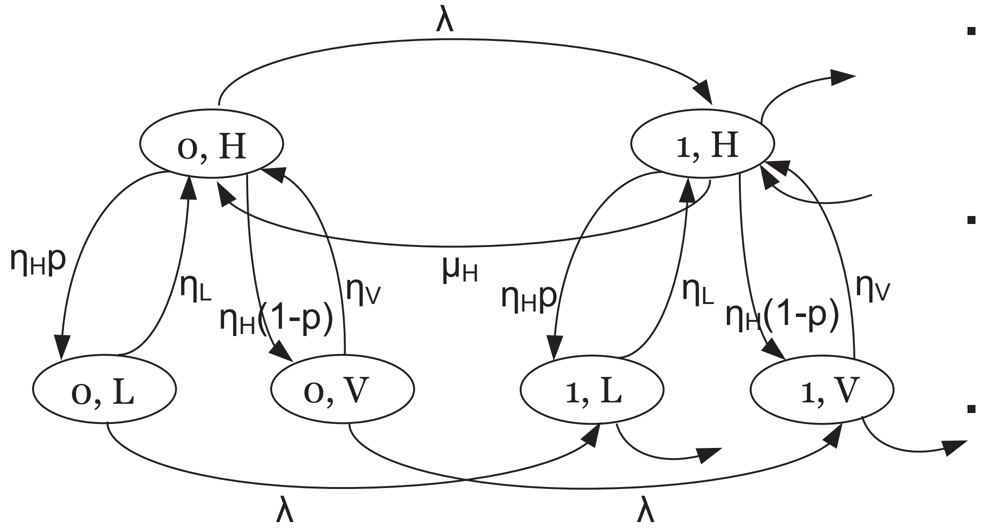

Figure 8.9: La chaîne de Markov 2D pour le temps de scannage hyper-exponentielle.

Table 8.4: La comparaison analytique des modes underlay et interweave.

T_s	Condition	Notation
Erlang	$E[T_s] < \frac{-B_2+\sqrt{B_2^2-4B_1B_3}}{2B_1}$	$B_1 = 2\eta_H^2 k + \eta_H(k+1)\mu_H + 2kE[T_u]\lambda\eta_H^2$ $B_2 = \eta_H(4k - 2k\mu_H E[T_u] + 4kE[T_u]\lambda)$ $B_3 = 2k(1-(\mu_H-\lambda)E[T_u])$
Exponentiel	$E[T_s] < \frac{-A_2+\sqrt{A_2^2-4A_1A_3}}{2A_1}$	$A_1 = \eta_H(\mu_H+\eta_H) + \lambda\eta_H^2 E[T_u] > 0$ $A_2 = 2\eta_H - \eta_H(\mu_H-2\lambda)E[T_u]$ $A_3 = 1-(\mu_H-\lambda)E[T_u]$
Hyper exponentielle	$E[T_s] < \frac{-C_2+\sqrt{C_2^2-4C_1C_3}}{2C_1}$	$C_1 = \eta_L\eta_V\eta_H^2\,(1+\lambda E[T_u])$ $C_2 = \eta_H\left[\mu_H(\eta_V+\eta_L) + \eta_L\eta_V\,(2+2\lambda E[T_u] - \mu_H E[T_u])\right]$ $C_3 = \eta_L\eta_V - \eta_H\mu_H - \eta_L\eta_V\,(\mu_H-\lambda)\,E[T_u]$

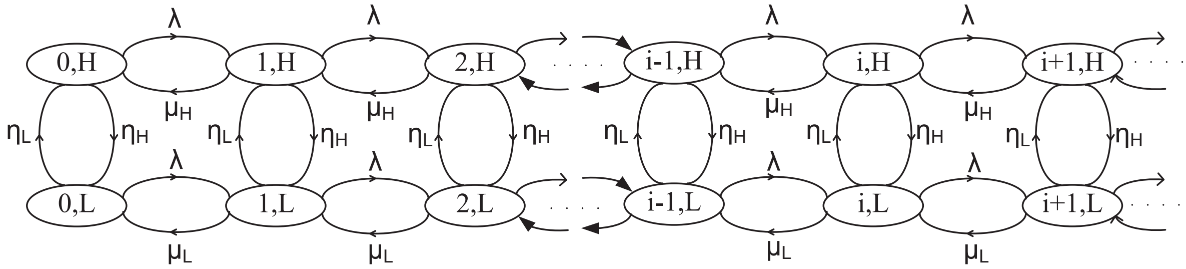

Figure 8.10: La chaîne de Markov 2D pour le modèl underlay.

Notre politique dynamique surpasse considérablement les deux modes de spectre. Il est défini, à la fois pour la minimisation du délai et la maximisation du débit, comme:

Definition 5. La politique dynamique.

- *Le SU résidera sur le canal actuel s'il est disponible (aucune activité PU) et continuera son activité ici.*
- *Si un PU est détecté, le SU continuera à transmettre avec une puissance inférieure, jusqu'à un instant t, appelé le «point tournant».*
- *Si le PU ne libère pas le canal à l'instant t, le SU cesse la transmission et lance le scannage pour un nouveau canal disponible.*
- *Si le PU quitte le canal avant l'instant t, alors le SU reprend la transmission à une puissance supérieure, et réinitialise le point tournant à t d'unités à venir.*

Le tableau 8.5 résume tous les scénarios possibles en termes de la nature de la distribution des périodes occupés par un PU.

Table 8.5: Résumé des politiques dynamiques.

Scenario	Politique dynamique optimale
Static interweave mieux	Static interweave ($t_{opt} = 0$)
Static underlay mieux + IFR OFF	Static underlay ($t_{opt} = \infty$)
Static underlay mieux + Exp. OFF	Static underlay ($t_{opt} = \infty$)
Static underlay mieux + DFR OFF	Politique dynamique avec $t_{opt} \in (0, \infty)$

Comme l'utilisation des canaux dans les deux modes underlay et interweave peuvent être modélisés avec un processus alternant de renouvellement (ON-OFF), nous pouvons utiliser l'approche du renouvellement-récompense pour calculer le débit à long terme de chaque mode, et de les comparer. Dans les deux cas, un cycle est constitué d'une période ON et d'une période OFF, et la récompense correspond à la quantité de données envoyées pendant un cycle. Les récompenses respectives sont présentées dans la Fig. 8.11 (underlay) et Fig. 8.12 (interweave). Deux différences peuvent être vues là: (i) il y a une non-nulle récompense pendant les périodes OFF en mode underlay; (ii) les périodes OFF dans le cas de interweave dépendent des statistiques de temps de scannage, plutôt que du temps de l'activité PU.

En comparant le débit pour les deux modes, nous obtenons le résultat suivant:

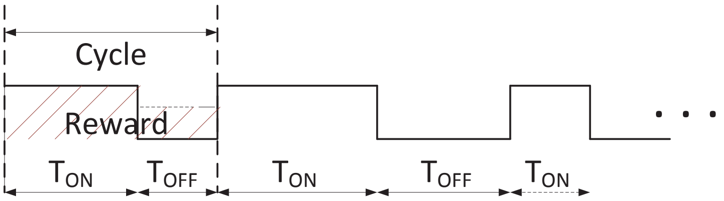

Figure 8.11: Les renouvellements pour le cas underlay.

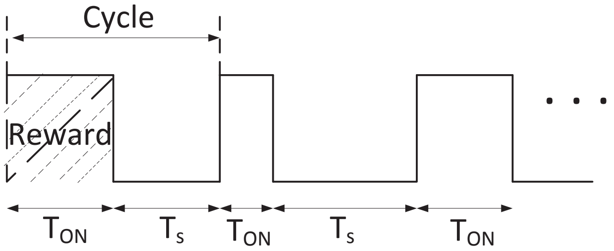

Figure 8.12: Les renouvellements pour le cas interweave.

Result 18. *La technique de l'accès interweave au spectre surpasse le mode underlay d'accès au CRN si la condition suivante est remplie*

$$\frac{E[T_s]}{E[T_{OFF}]} < \frac{1-\frac{\mu_L}{\mu_H}}{1+\frac{\mu_L}{\mu_H}\frac{E[T_{OFF}]}{E[T_{ON}]}}. \tag{8.15}$$

Aussi bien dans le cas de maximisation de débit (comme dans le cas de minimisation de délai), si le mode "statique" interweave d'accès est mieux en moyenne (dans le sens du résultat précédent), alors il est facile de voir qu'il n'est plus nécessaire d'employer le mode underlay. Cependant, si le mode underlay est mieux en moyenne, on peut encore améliorer les performances en décidant de scanner et de passer à un nouveau canal (inactif) dans certains cas. Nous présentons une telle politique dynamique qui tente de prédire les gains potentiels "sur la volée" et peuvent choisir de transmettre à faible puissance ou de scanner à tout moment. De même que dans la politique du délai, nous déterminons le point tournant optimal t_{opt} lorsque le SU devrait passer au mode interweave, de sorte que le débit maximum est atteint.

Par conséquent, ceci est pareil que la politique dynamique de délai. La seule différence réside dans le choix du point tournant optimal t_{opt}, que nous choisissons maintenant afin de maximiser le débit à la place. Nous pouvons à nouveau utiliser l'approche de renouvellement-récompense, avec la différence que maintenant les périodes basses peuvent être soit des périodes basses comme dans le scénario de underlay, soit une combinaison d'une période basse et le temps de scannage.

Comme dans le cas de la politique dynamique de délai, la clé dans la décision de commutation est dans l'incertitude quant à combien de temps l'activité PU (i.e. le cycle OFF) va durer. Même si nous connaissons la valeur attendue (qui est la base pour le choix de la politique statique), la

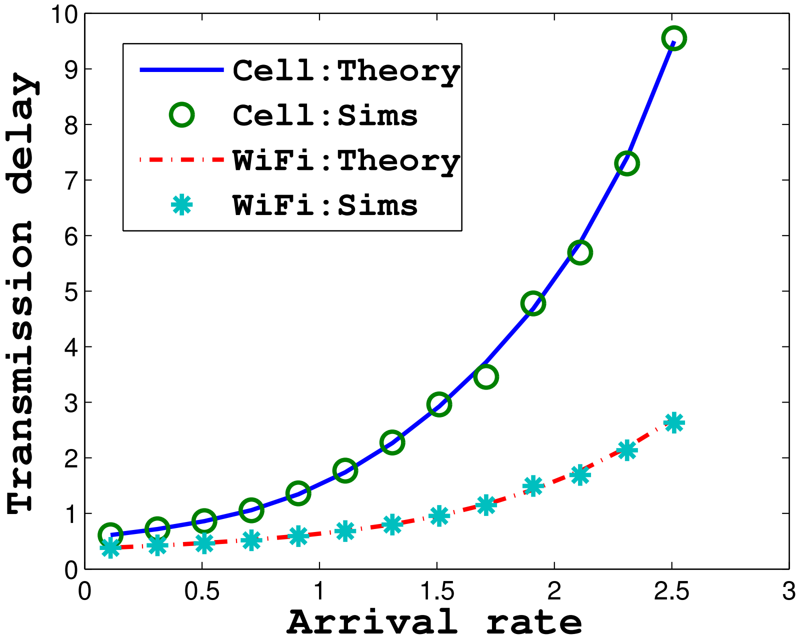

Figure 8.13: Le délai pour l'accès underlay au spectre.

variabilité au-delà de cette valeur moyenne joue un rôle important. Il est facile de voir que les mêmes arguments, comme dans le cas de minimisation de délai, peuvent être fait pour montrer que toute amélioration au-delà des politiques statiques est seulement possible si underlay est mieux en moyenne, et que les périodes OFF ont un taux de défaillance décroissant. Ainsi, le résultat suivant dérive le point tournant optimale t_{opt} pour les périodes OFF qui sont Pareto avec des paramètres L, α.

Result 19. *La valeur optimale du point tournant qui maximise le débit pour Pareto OFF et ON exponentiel est la solution de*

$$\frac{\mu_H - \mu_L}{\eta_H} t^{\alpha} - \alpha E[T_s] \left(\frac{\mu_H}{\eta_H} - \mu_L L \frac{\alpha}{1-\alpha} \right) t^{(\alpha-1)} - \frac{\mu_L E[T_s] L^{\alpha}}{1-\alpha} = 0. \tag{8.16}$$

La Fig. 8.13 compare les résultats de simulation à nos modèles analytiques de prédiction pour le délai moyen de fichiers SU tandis que le taux d'arrivée des fichiers augmentent. Comme on peut le voir, les résultats théoriques correspondent aux résultats obtenus à partir de simulations. Comme cela est prévu dans les systèmes de files d'attente, le retard augmente lorsque le taux d'arrivée (et donc l'utilisation du système) augmente. Dans la même figure, le délai basé sur des statistiques du réseau WiFi [19] est représenté ainsi. Encore une fois, il y a un accord solide entre la théorie et les simulations. Le délai pris dans le scénario cellulaire est plus grand. Ceci est dû car les débits de données dans le scénario de connexion sont considérés comme étant plus haut et le PU est moins actif dans le cas présent.

Ensuite, nous continuons avec la validation pour l'accès interweave. Pour les deux scénarios, le temps de scannage moyen est de 0.5 s et est Pareto délimité. La Fig. 8.14 montre le débit moyen contre la disponibilité du canal pour les deux réseaux. Par opposition à la validation

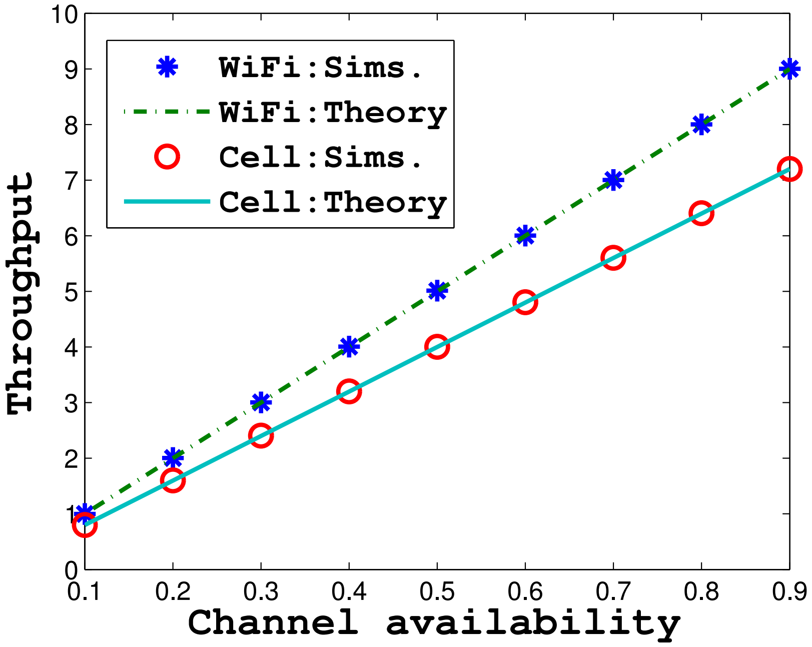

Figure 8.14: Le débit pour l'accès interweave au spectre.

du modèle underlay, on utilise la disponibilité des canaux sur l'axe des x. Il désigne le % de temps que le SU peut transmettre, et est égal à $\frac{E[T_{ON}]}{E[T_{ON}]+E[T_s]}$. Maintenant, le cycle de service ne signifie pas grand-chose, puisque le SU ne réside pas dans un seul canal et différents canaux ont différents cycles de service.

Comme autre scénario intéressant, nous considérons l'accès underlay avec des paramètres: $\eta_H = 0.1, \eta_L = 1, \mu_H = 10, \mu_L = 0.5$. La Fig. 8.15 montre le délai moyen de fichier (notée I) contre les différents délais moyens de scannage (exp. distribué), pour trois intensités de circulation différentes (faible, moyen, élevé). Sur la même figure, pour chaque intensité du trafic le délai correspondant de underlay (notée U) est représenté également. Enfin, les valeurs maximales théoriques pour les temps de scannage attendus (tableau 8.4), pour lesquels le interweave mode d'accès surpasse l'accès underlay sont représentés avec des petits cercles. Pour le cas de la circulation clairsemée, le mode interweave commence à devenir meilleur pour le temps de scannage inférieures à 0.8 s. La première chose à observer est que la valeur maximale prévue pour le temps de scannage (à savoir le point de passage) est correcte. Le deuxième résultat important est que cette limite est plus élevée lorsque la charge augmente. Cela est dû au fait que pour des charges plus élevées, le délai de la file d'attente est la plus grand composant de délai. Par conséquent, il vaut la peine d'attendre un certain temps, trouver un canal libre et ensuite se débarrasser des données en file d'attente à un taux supérieur. Nous avons également remarqué qu'augmenter la charge conduit en outre à des incréments de plus en plus petits de ce point de passage.

La minimisation de délai. Comme indiqué précédemment, la politique dynamique peut offrir des avantages supplémentaires lorsque les périodes d'activité PU (à savoir les périodes OFF dans le mode underlay) sont soumis à une distribution de probabilité avec un taux de défaillance décroissant (i.e. avec une très grande variabilité). Nous considérons un scénario où les périodes OFF ont des distributions Pareto délimitée, avec des paramètres $L = 0.2, H = 100, \alpha = 1.2$. Le

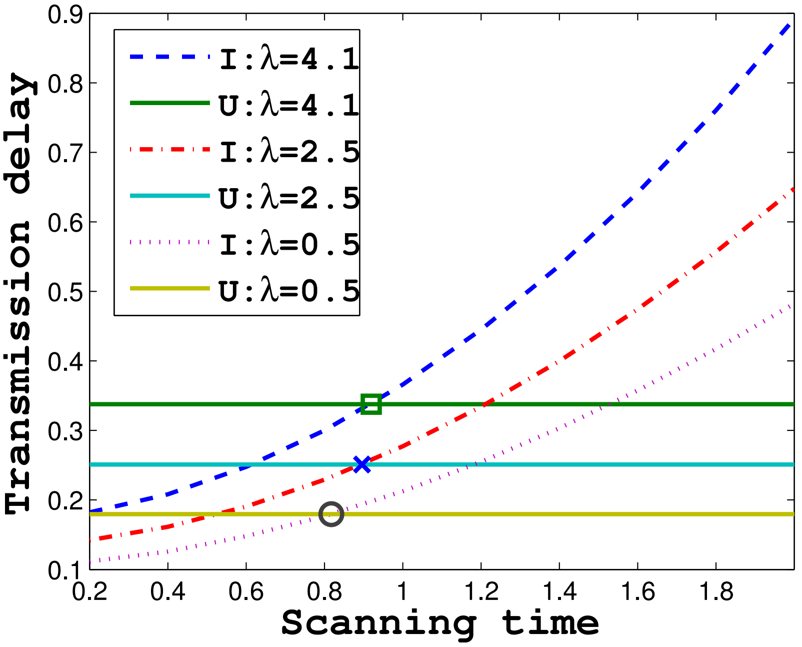

Figure 8.15: La politique statique de délai pour λ différentes et temps exponentiel de scannage.

temps de scannage moyen est $E[T_s] = 1$s. La Fig.8.16 montre le délai moyen par rapport au taux d'arrivée. Selon la politique statique, le mode underlay est mieux que le mode interweave. Cependant, le meilleur résultat est obtenu avec la politique dynamique, qui offre une réduction de délai supplémentaire de 20 à 50%.

La maximisation de débit. Ensuite, nous considérons le scénario avec des périodes OFF de Pareto (les paramètres sont identiques à ceux de la Fig. 8.16). Pour la durée moyenne des périodes ON et OFF, ainsi que pour la durée de scannage il en est de même. La Fig. 8.17 montre le débit moyen des politiques différentes. La politique dynamique augmente encore le débit de 20% supplémentaires. Ceci est cohérent avec nos revendications précédentes, que le débit peut être améliorée davantage lorsque le mode underlay est meilleure. Depuis que les périodes OFF sont distribués par distribution de Pareto, certains d'entre eux seront très long et auront un impact décisif sur le débit. Par conséquent, trouver un autre canal libre permettra d'améliorer la performance.

Les articles liés à ce chapitre sont:

- *F. Mehmeti, T. Spyropoulos, "Stay or Switch? Analysis and Comparison of Interweave and Underlay Spectrum Access in Cognitive Radio Networks", submitted to IEEE Transactions on Mobile Computing, November 2014.*

- *F. Mehmeti, T. Spyropoulos, "Underlay vs. Interweave: Which one is better?", Tech. Report, RR-14-296, Eurecom, 2014.*

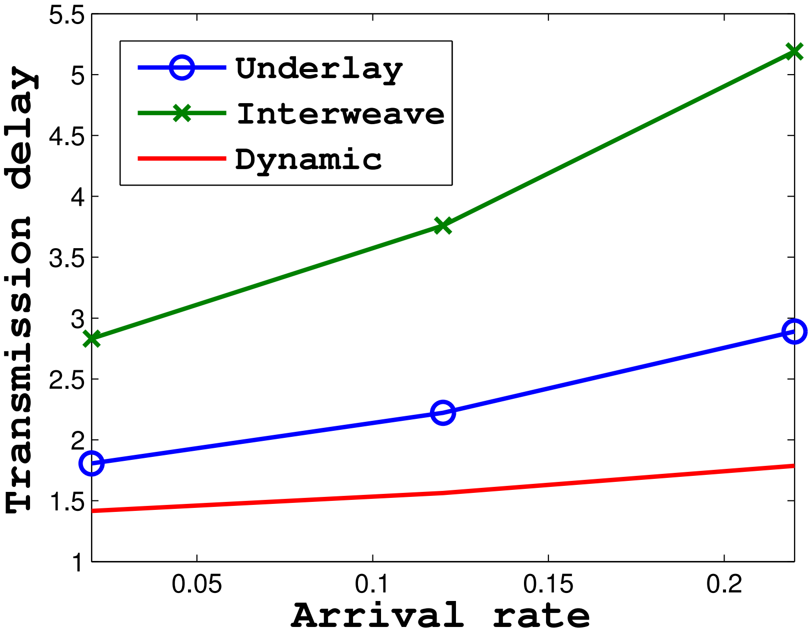

Figure 8.16: La politique dynamique de délai pour des périodes OFF de Pareto.

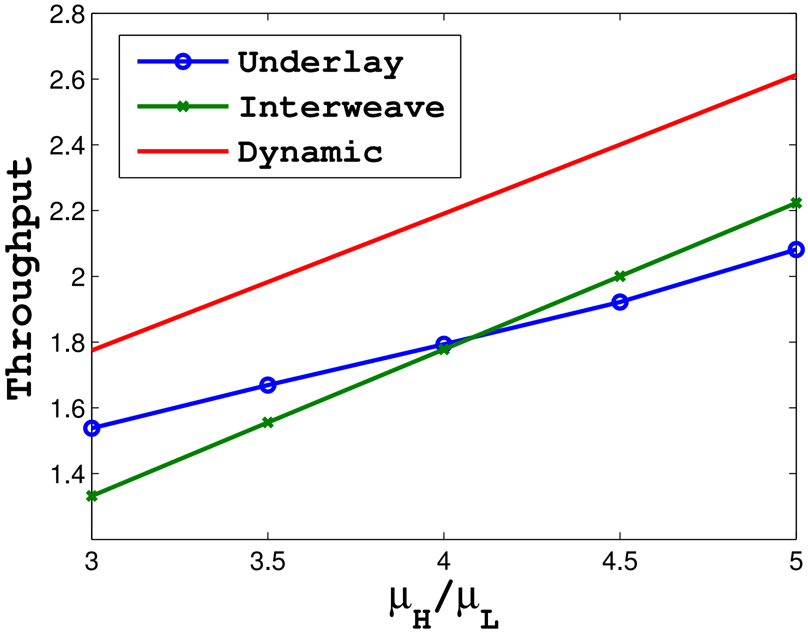

Figure 8.17: La politique dynamique de débit pour des périodes OFF de Pareto.

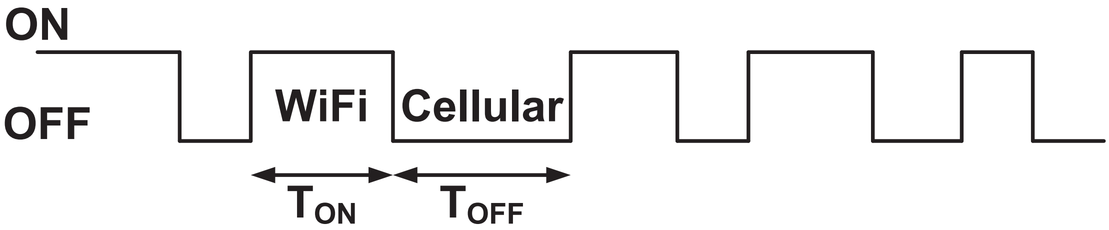

Figure 8.18: Le modèle de la disponibilité du réseau WiFi.

Chapitre 5 - Analyse de la performance de déchargement on-the-spot des données mobiles.

Dans le Chapitre 5, nous avons proposé un modèle analytique de file d'attente pour l'analyse de la performance de déchargement on-the-spot des données mobiles. Le modèle de ce chapitre est semblable au modèle sous-jacent du Chapitre 4. Nous considérons un scénario simple où l'utilisateur peut choisir entre le WiFi et une technologie cellulaire unique. La disponibilité de WiFi est modélisée avec un processus de renouvellement alterné (Fig. 8.18).

Avec l'utilisation des fonctions de génération de probabilité (PGF) nous avons résolu la chaîne 2D de Markov correspondante et nous avons obtenu l'expression suivante pour le délai moyen de transmission (les paramètres ont une signification similaire à ceux de Tab. 8.3, où ceux avec index "c" correspondent à la période cellulaire, et ceux avec "w" aux périodes WiFi)

$$E[T] = \frac{1}{\mu - \lambda} + \frac{\mu_c(\mu_w - \lambda)\pi_{0,c} + (\mu_c - \lambda)(\mu_w(\pi_{0,w} - 1) + \lambda)}{\lambda(\eta_c + \eta_w)(\mu - \lambda)}. \tag{8.17}$$

L'efficacité de déchargement (le pourcentage de données transmisent à travers l'interface du réseau WiFi) est donnée par

$$OE = \frac{\mu_w}{\mu_w + \mu_c \frac{G_c(1) - G_c(0)}{G_w(1) - G_w(0)}}. \tag{8.18}$$

Ici, $G_c(1) - G_c(0)$ représente le pourcentage du temps avec au moins un fichier tout en étant dans une période cellulaire. De même, $G_w(1) - G_w(0)$ est le pourcentage de temps avec au moins un fichier dans une période de WiFi.

Nous généralisons notre analyse au cas où de multiples technologies cellulaires (et les taux respectifs), tels que: 3G, 4G, etc., sont disponibles à un utilisateur, par exemple, en fonction de son emplacement, et/ou des taux différents sont offerts par la même technologie (par exemple, adaptation de débit, intérieur/extérieur, etc.). Nous modélisons ce cas plus complexe avec une chaîne de Markov plus complexe qui a M niveaux, où chaque niveau correspond à une certaine technologie. Ceci est représenté sur la Fig. 8.19. Après avoir résolu cette chaîne (le résultat est très complexe), on obtient le délai moyen de fichier.

La Fig. 8.20 montre le délai moyen de transmission du fichier (i.e. file d'attente + transmission) pour un scénario de piétons, pour différents taux d'arrivée. La gamme des taux d'arrivée indiqué correspond à une utilisation du serveur de 0 à 0.9. Nous pouvons observer, dans la Fig. 8.20, qu'il y a un bon accord entre la théorie et les simulations. En outre, le délai moyen de transmission de fichier est augmenté avec le taux d'arrivée, comme prévu, en raison d'effets

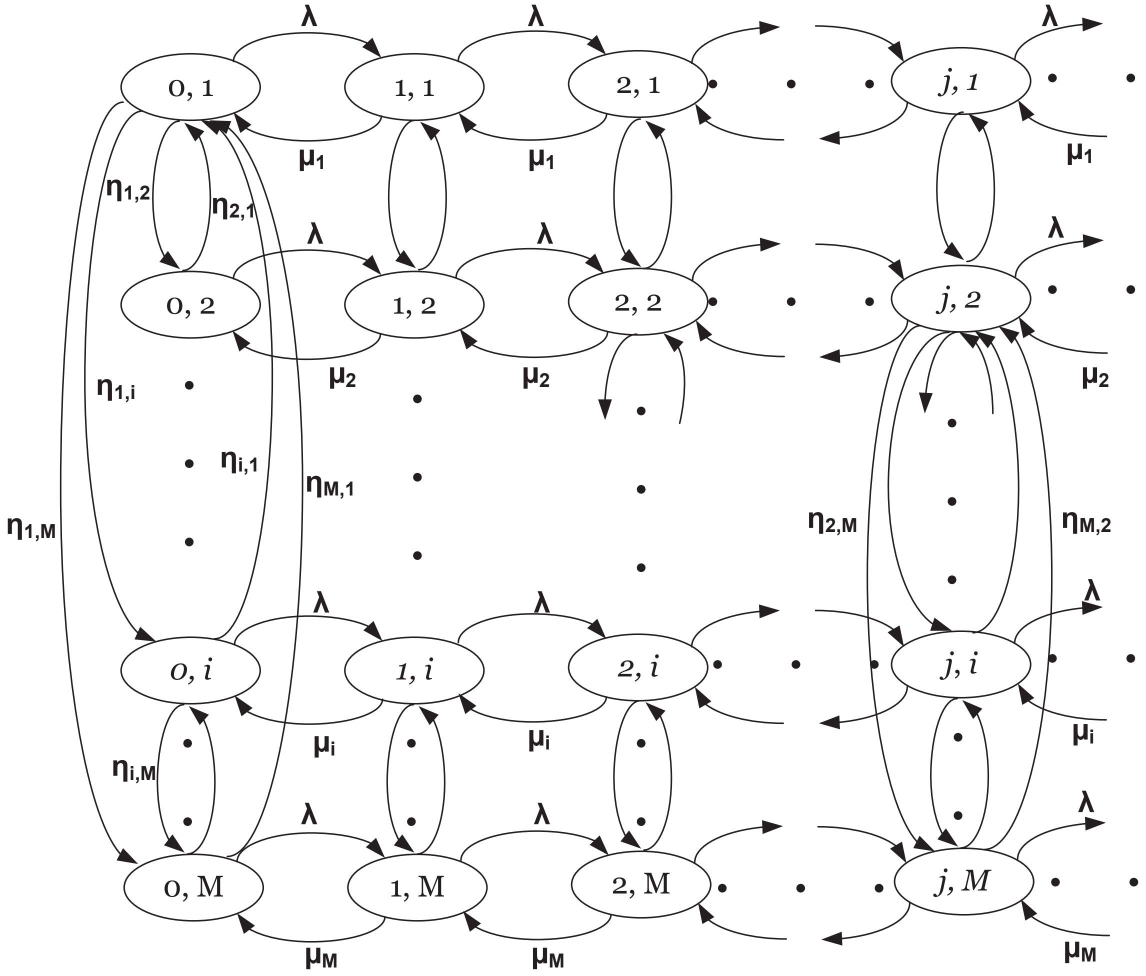

Figure 8.19: La chaîne Markov 2D pour le modèle de déchargement on-the-spot des données mobiles avec plusieurs niveaux.

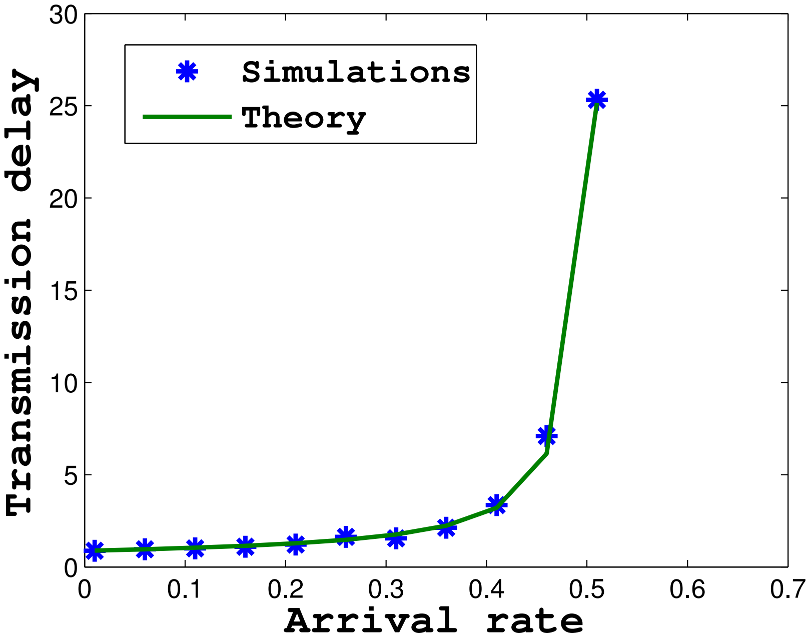

Figure 8.20: L'utilisateur piétonne.

Table 8.6: Les paramètres pour différentes technologies d'accès [3, 4].

Technologie	Débit	Durée moyenne
WiFi	2 Mbps	10 s
3G	1 Mbps	3 s
HSPA	1.5 Mbps	5 s
LTE	10 Mbps	5 s
Pas de réseaux	0	2 s

de file d'attente. La Fig. 8.21 illustre en outre le délai moyen de transmission de fichier pour un scénario des véhicules. On peut observer là que le temps de transmission moyen est supérieur à la Fig. 8.20. Cela est raisonnable, en raison de la faible disponibilité de WiFi, résultant en plus de la circulation étant transmise à travers l'interface de réseau cellulaire plus lente. Une fois de plus, nous pouvons observer une bonne accord entre la théorie et les simulations.

Ensuite, nous considérons les scénarios avec de multiples technologies d'accès (WiFi, 3G, HSPA, LTE) ou même sans couverture de réseau du tout. À savoir, il existe des opérateurs qui pourraient offrir une couverture 4G seulement dans certaines régions, alors que dans les autres, ils ne proposent que la 3G. Il pourrait également exister des régions avec peu ou pas de couverture du tout (dans les zones peu peuplées). Dans ce qui suit, nous verrons comment notre théorie à plusieurs niveaux de la Section 5.3 va faire face à des scénarios (simulées) réels. Sauf indication contraire, les taux de données et durées moyennes sont donnés dans le tableau 8.6.

Tout d'abord, nous nous concentrons sur le scénario selon lequel il y a 3 choix possibles du réseau: WiFi, 3G et LTE. La politique ici est que le WiFi est le réseau qui détient une priorité absolue. Quand il n'y a pas de couverture WiFi, LTE a la priorité sur la 3G. Nous supposons

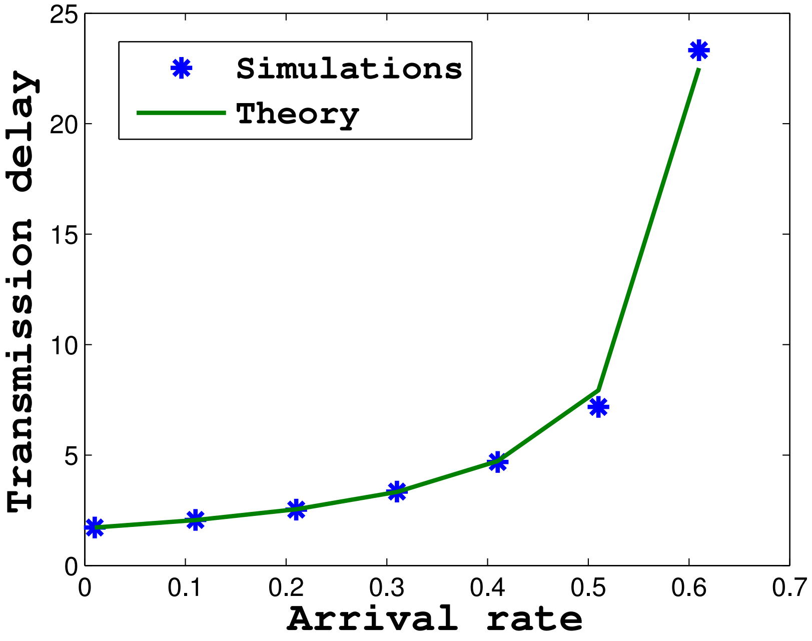

Figure 8.21: L'utilisateur en véhicule.

qu'il ya toujours disponibilité du réseau 3G. Il existe une probabilité égale à passer à une autre technologie d'accès après avoir quitté le réseau courant. Puisqu'il n'y a que 3 niveaux possibles, cette probabilité est égale à 0.5. Les flux sont distribués exponentiellement avec la taille moyenne de 125 kB, et le processus d'arrivée est Poisson. Le taux de disponibilité du réseau WiFi est 50%.

La Fig. 8.22 montre le délai moyen de fichier vs. le taux d'arrivée pour ce système. Comme on peut le voir, notre théorie est en adéquation avec les simulations. Comme prévu, le délai augmente avec l'augmentation du taux d'arrivée de trafic, en raison de l'effet de files d'attente.

Le deuxième scénario représenté à la Fig. 8.22 correspond à 4 niveaux possibles: WiFi, 3G, HSPA et LTE. Les paramètres sont indiqués dans le Tableau 8.6. La probabilité de passer à un certain niveau est maintenant 1/3. Le taux de disponibilité se trouve maintenant d'être 43.5%. Il est à nouveau un bon accord avec la théorie. Le délai moyen est un peu plus élevé par rapport à l'exemple précédent, puisque le taux de données HSPA est inférieur par rapport au WiFi, et que les autres caractéristiques du réseau sont les mêmes.

Nous considérons également la possibilité de ne pas avoir de couverture du réseau du tout. Maintenant, nous devons tenir compte de 5 niveaux (Fig. 8.22). Étant donné que les 4 autres niveaux ont les mêmes paramètres que ci-dessus, pour l'indisponibilité du réseau nous choisissons la durée moyenne de 2 s. La probabilité de rencontrer un niveau spécifique après avoir quitté celui utilisé est de 0.25. Le taux de disponibilité est 40%. Encore une fois, il y a une adéquation entre la théorie et les simulations qui montre que notre théorie est correcte. Comme prévu, le délai est plus grand, car il y a des périodes où il n'y a aucune connectivité.

Les articles liés à ce chapitre sont:

- *F. Mehmeti, T. Spyropoulos, "Performance Analysis of On-the-Spot Mobile Data Offload-*

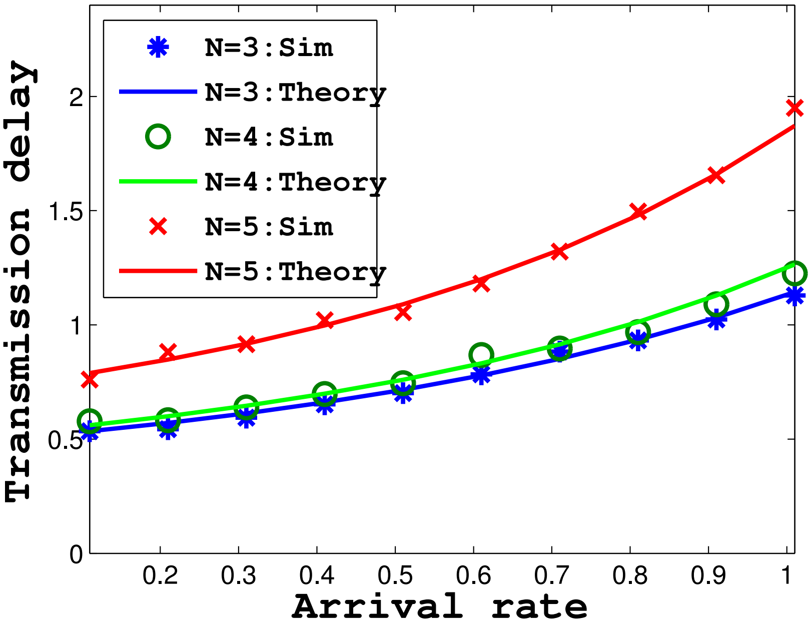

Figure 8.22: Le délai de transmission.

ing", in Proc. of IEEE Global Telecommunications Conferenece (IEEE GLOBECOM), Atlanta, USA, 2013.

- *F. Mehmeti, T. Spyropoulos, "Performance Analysis of Mobile Data Offloading in Heterogeneous Networks", submitted to IEEE Transactions on Mobile Computing, September, 2014.*

Chapitre 6 - Est-ce la peine d'être patient? Analyse et optimisation de déchargement retardé (différé) des données mobiles.

Les principales contributions du Chapitre 6 peuvent être résumées comme suit. Nous proposons un modèle analytique de file d'attente pour le problème du déchargement différé, sur la base de chaînes de Markov à deux dimensions, et d'en tirer des expressions pour le délai moyen et d'autres indicateurs de performance en fonction des échéances, et les paramètres clés du système. Nous donnons également des approximations forme fermées pour les différents régimes d'intérêt. Nous fournissons quelques idées à propos du processus de file d'attente cellulaire et de comment celui-ci peut être modélisé, et quel genre d'approximations peuvent y être utilisés, pour fournir un système plus fiable en termes de mieux décrire ce qui se passe réellement dans un système réel. Le modèle de chaîne de Markov que nous utilisons dans le Chapitre 6 est légèrement différent de celui du Chapitre 5. Alors que dans la chaîne de Markov du Chapitre 5 tous les taux de transition sont indépendants de l'état dans lequel se trouve le système, dans une partie de la chaîne de Markov de notre modèle du Chapitre 6, le taux de transition dépend de l'état du système.

Nous validons largement nos résultats, en utilisant également des scénarios et des paramètres observés dans les traces de mesure réels qui dérogent aux hypothèses formulées dans notre modèle. Nous formulons et résolvons des problèmes de base d'optimisation coût-performances, et en tirons les régions de compromis réalisables en fonction des paramètres du réseau (disponibilité WiFi, la charge de l'utilisateur, etc.). Enfin, nous montrons que nos résultats sont valables pour d'autres disciplines de type de service, tels que le partage du processeur (PS), Dernier arrivé premier servi (LCFS), etc.

Les articles liés à ce chapitre sont:

- *F. Mehmeti, T. Spyropoulos, "Is it Worth to be Patient? Analysis and Optimization of Delayed Mobile Data Offloading", in Proc. of IEEE International Conference on Computer Communications (IEEE Infocom 2014), Toronto, Canada, 2014.*
- *F. Mehmeti, T. Spyropoulos, "Performance Modeling, Analysis and Optimization of Delayed Mobile Data Offloading under Different Service Disciplines", to be submitted to IEEE/ACM Transactions on Networking, March, 2015.*
- *F. Mehmeti, T. Spyropoulos, "Optimization of Delayed Mobile Data Offloading", Tech. Report, RR-13-286, Eurecom, 2013.*

Bibliography

[1] B. Wang and K. Liu, “Advances in cognitive radio networks: A survey,” *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 1, 2011.

[2] J. Mitola, “Cognitive radio: An integrated agent architecture for software defined radio,” Ph.D. dissertation, Royal Institute of Technology, 2000.

[3] R. Research, “Beyond LTE: Enabling the mobile broadband explosion,” 2014.

[4] Http://www.swisscom.ch/en/residential/mobile/mobile-network.html.

[5] I. F. Akiyildiz, W. Y. Lee, M. C. Vuran, and S. Mohanty, “A survey on spectrum management in cognitive radio networks,” *IEEE Comm. Mag.*, vol. 46, no. 4, apr. 2008.

[6] W. Y. Lee and F. Akyildiz, “A spectrum decision framework for cognitive radio networks,” *IEEE Tran. Mob. Computing*, vol. 10, no. 2, 2011.

[7] “Cisco visual networking index: Global mobile data traffic forecast update, 2012-2017,” Feb. 2013, http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537 /ns705/ns827/ white paper c11-520862.pdf.

[8] “Mobile data offloading through WiFi,” 2010, Proxim Wireless.

[9] “Growing data demands are trouble for Verizon, LTE capacity nearing limits,” http://www.talkandroid.com/97125-growing-data-demands-are-trouble-for-verizon-lte-capacity-nearing-limits/, 2012.

[10] T. Kaneshige, “iPhone users irate at idea of usage-based pricing,” Dec. 2009, http://www.pcworld.com/article/184589/ATT IPhone Users Irate at Idea of Usage Based Pricing.html.

[11] Http://www.3gpp1.eu/ftp/Specs/archive/23 series/23.829/.

[12] K. Lee, I. Rhee, J. Lee, S. Chong, and Y. Yi, “Mobile data offloading: How much can WiFi deliver,” in *Proc. of ACM CoNEXT*, 2010.

[13] K. Tourki, K. Qaraqe, and M. Alouini, “Outage analysis for underlay relay-assisted cognitive networks,” in *Proc. of IEEE GLOBECOM*, 2012.

[14] A. Molisch, *Wireless Communications*, 2nd ed. John Wiley & Sons, 2011.

[15] “AT&T: Improving 3G network,” http://gigaom.com/2008/06/08/3g-network-iphone/.

[16] P. Bahl, R. Chandra, T. Moscibroda, R. Murty, and M. Welsh, "White space networking with Wi-Fi like connectivity," in *Proc. of ACM SIGCOMM*, 2009.

[17] R. Lu, X. Li, X. Liang, X. Shen, and X. Lin, "GRS: The green, reliability, and security of emerging machine to machine communications," *IEEE Comm. Mag*, vol. 49, no. 4, pp. 28–35, April 2011.

[18] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary user behavior in cellular networks and implications for dynamic spectrum access," *IEEE Comm. Mag.*, vol. 47, no. 3, mar 2009.

[19] S. Geirhofer and L. Tong, "Dynamic spectrum access in the time domain: Modeling and exploiting white space," *IEEE Comm. Mag.*, vol. 45, 2007.

[20] S. M. Ross, *Stochastic Processes.* John Wiley & Sons, 1996.

[21] X. W. C. Zhang and J. Li, "Cooperative cognitive radio with priority queueing analysis," in *Proc. IEEE ICC*, 2009.

[22] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, 2007.

[23] H. Kim and K. Shin, "Fast discovery of spectrum opportunities in cognitive radio networks," in *Proc. of IEEE DySPAN*, 2008.

[24] L. Kleinrock, *Queueing theory, Volume I: Theory.* John Wiley & Sons, 1975.

[25] H. Su and X. Zhang, "Cross-layer based opportunistic MAC protocols for QoS provisionings over cognitive radio wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 118–129, Jan. 2008.

[26] L. T. S. Geirhofer and B. Sadler, "Cognitive medium access: Constraining interference based on experimental models," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, Jan. 2008.

[27] J. B. D. Willkomm, S. Machiraju and A. Wolisz, "Primary users in cellular networks: A large-scale measurement study," in *Proc. of IEEE DySPAN*, 2008.

[28] A. Lertsinsrubtavee, N. Malouch, and S. Fdida, "Controlling spectrum handoff with a delay requirement in cognitive radio networks," in *Proc. of ICCCN*, Aug. 2012.

[29] I. Suliman and J. Lehtomaki, "Queueing analysis of opportunistic access in cognitive radios," in *Proc. of CogART*, 2009.

[30] T. Q. D. T. Hung and H.-J. Zepernick, "Average waiting time of packets with different priorities in cognitive radio networks," in *Proc. of IEEE ISWPC*, 2009.

[31] L. Kleinrock, *Queueing theory, Volume II: Computer Applications.* John Wiley & and Sons, 1976.

[32] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Comm.*, vol. 23, no. 2, feb. 2005.

[33] G. Ganesan and Y. Li, "Cooperative spectrum sensing in cognitive radio networks," in *Proc. of IEEE DySPAN*, 2005.

[34] S. Zheng, X. Yang, S. Chen, and C. Lou, "Target channel sequence selection scheme for proactive-decision spectrum handoff," *IEEE Comm. Letters*, vol. 15, no. 12, dec. 2011.

[35] L. Wang, C. Wang, and C. Chang, "Modeling and analysis for spectrum handoffs in cognitive radio networks," *IEEE Tran. Mob. Computing*, vol. 11, no. 9, sep. 2012.

[36] C. Wang, L. Wang, and F. Adachi, "Modeling and analysis for reactive-decision spectrum handoff in cognitive radio networks," in *Proc. of IEEE Globecom*, 2010.

[37] H. Kim and K. Shin, "Efficient discovery of spectrum opportunities with MAC-layer sensing in cognitive radio networks," *IEEE Tran. Mob. Computing*, vol. 7, no. 5, may. 2008.

[38] H. Jiang, L. Lai, R. Fan, and H. V. Poor, "Optimal selection of channel sensing order in cognitive radio," *IEEE Tran. Wireless Comm.*, vol. 8, no. 1, jan. 2009.

[39] J. Zhao and X. Wang, "Channel sensing order in multi-user cognitive radio networks," in *Proc. of IEEE DySPAN*, 2012.

[40] R. Fan and H. Jiang, "Channel sensing-order setting in cognitive radio networks: A two-user case," *IEEE Tran. Vehicular Technology*, vol. 58, no. 9, nov. 2009.

[41] H. T. Cheng and W. Zhuang, "Simple channel sensing order in cognitive radio networks," *IEEE J. Sel. Areas in Comm.*, vol. 29, no. 4, apr. 2011.

[42] F. Mehmeti and T. Spyropoulos, "Analysis of cognitive user performance under generic primary user activity," EURECOM, Tech. Rep., 2012, http://www.eurecom.fr/∼spyropou/papers/SU-performance-techreport.pdf.

[43] M. Kartheek and V. Sharma, "Providing QoS in a cognitive radio network," in *Proc. of COMSNETS*, 2012.

[44] L. Xiukui and S. Zekavat, "Traffic pattern prediction and performance investigation for cognitive radio systems," in *Proc. of IEEE WCNC*, 2008.

[45] J. Stewart, *Calculus.* Brooks/Cole, 2011.

[46] http://share.cisco.com/internet-of-things.html.

[47] N. Mahmood, F. Yilmaz, G. Oien, and M. Alouini, "On hybrid cooperation in underlay cognitive radio networks," *IEEE Tran. Wireless Comunications*, vol. 12, no. 9, 2013.

[48] Z. Yang, X. Xie, and Y. Zheng, "A new two-user cognitive radio channel model and its capacity analysis," in *Proc. of ISCIT*, 2009.

[49] M. F. Neuts, *Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach.* John Hopkins University Press, 1981.

[50] F. Mehmeti and T. Spyropoulos, "To scan or not to scan: The effect of channel heterogeneity on optimal scanning policies," in *Proc. of IEEE SECON*, 2013.

[51] D. Gozupek, S. Buhari, and F. Alagoze, “A spectrum switching delay-aware scheduling algorithm for centralized cognitive radio networks,” *IEEE Tran. Mobile Computing*, vol. 12, no. 7, 2013.

[52] U. Yechiali and P. Naor, “Queueing problems with heterogeneous arrivals and service,” *Operations Research*, 1971.

[53] H. Hakim, H. Boujemaa, and W. Ajib, “Performance comparison between adaptive and fixed transmit power in underlay cognitive radio networks,” *IEEE Tran. Comunications*, vol. 61, no. 12, 2013.

[54] M. Seyfi, S. Muhaidat, and J. Liang, “Relay selection in underlay cognitive radio networks,” in *Proc. of IEEE WCNC*, 2012.

[55] M. Ki, H. Lee, and J. Song, “Performance analysis of distributed cooperative spectrum sensing for underlay cognitive radio,” in *Proc. of ICACT*, 2009.

[56] B. Wang and D. Zhao, “Performance analysis in CDMA-based cognitive wireless networks with spectrum underlay,” in *Proc. of IEEE GLOBECOM*, 2008.

[57] H. Chamkhia, M. Hasna, R. Hamila, and S. Hussain, “Performance analysis of relay selection schemes in underlay cognitive networks with decode and forward relaying,” in *Proc. of IEEE PIMRC*, 2012.

[58] P. Wang, D. Niyato, and H. Jiang, “Voice-service capacity analysis for cognitive radio networks,” *IEEE Tran. Vehicular Technology*, vol. 54, no. 4, 2010.

[59] S. Gunawardena and Z. Weihua, “Capacity analysis and call admission control in distributed cognitive radio networks,” *IEEE Tran. Wireless Communications*, vol. 10, no. 9, 2011.

[60] T. Chu, H. Phan, and H. Zepernick, “On the performance of underlay cognitive radio networks using M/G/1/K queueing model,” *IEEE Comm. Letter*, vol. 17, no. 5, 2013.

[61] L. Sibomana, H. Zepernick, H. Tran, and C. Kabiri, “Packet transmission time for cognitive radio networks considering interference from primary user,” in *Proc. of IWCMC*, 2013.

[62] S. Ross, *Introduction to probability models.* Academic Press, 2006.

[63] F. Mehmeti and T. Spyropoulos, “Who interrupted me? Analyzing the effect of PU activity on cognitive user performance,” in *Proc. of IEEE ICC*, 2013.

[64] H. Song, J. P. Hong, and W. Choi, “On the optimal switching probability for a hybrid cognitive radio system,” *IEEE Tran. Wireless Comm.*, vol. 12, no. 4, 2013.

[65] A. Giorgetti, M. Varrella, and M. Chiani, “Analysis and performance comparison of different cognitive radio algorithms,” in *Proc. of CogART*, 2009.

[66] M. Qutqut, F.M.Al-Turjman, and H. Hassanein, “MWM: Mobile femtocells utilizing WiFi (a data offloading framework for cellular networks using mobile femtocells),” in *Proc. of IEEE ICC*, 2013.

[67] C. B. Sankaran, “Data offloading techniques in 3GPP Rel-10 networks: A tutorial,” *IEEE Communications Magazine*, June 2012.

[68] A. Balasubramanian, R. Mahajan, and A. Venkataramani, “Augmenting mobile 3G using WiFi,” in *Proc. of ACM MobiSys*, 2010.

[69] J. Eriksson, H. Balakrishnan, and S. Madden, “Cabernet: Vehicular Content Delivery Using WiFi,” in *14th ACM MOBICOM*, San Francisco, CA, September 2008.

[70] Y. Y. X. Meng, S. Wong and S. Lu, “Characterizing flows in large wireless data networks,” in *Proceedings of ACM MOBICOM*, 2014.

[71] T. Osogami and M. Harchol-Balter, “Closed form solutions for mapping general distributions to minimal PH distributions,” *Performance Evaluation*, 2003.

[72] J. L. Snell and C. Grinsted, *Introduction to Probability.* American Mathematical Society, 2006.

[73] F. Mehmeti and T. Spyropoulos, “Performance analysis of on-the-spot mobile data offloading,” in *Proc. of IEEE Globecom*, 2013.

[74] A. Abhari and M. Soraya, “Workload generation for YouTube,” *Multimedia Tools and Applications*, 2010.

[75] I. Mitrany and B.Avi-Itzhak, “A many-server queue with service interruptions,” *Operations Research*, vol. 16, no. 3, 1968.

[76] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, “Mobile data offloading: How much can WiFi deliver,” *IEEE/ACM Trans. Netw.*, 2013.

[77] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, “Energy consumption in mobile phones: A measurement study and implications for network applications,” in *Proc. of ACM IMC*, 2009.

[78] B. Han, P. Hui, A. Kumar, M. V. Marathe, and J. S. A. Srinivasan, “Mobile data offloading through opportunistic communications and social participation,” *IEEE Tran. Mob. Computing*, vol. 11, no. 5, 2012.

[79] Y. Li, M. Qian, D. Jin, P. Hui, Z. Wang, and S. Chen, “Multiple mobile data offloading through delay tolerant networks,” in *Proc. of ACM CHANTS*, 2011.

[80] Y. Li, D. Jin, Z. Wang, L. Zeng, and S. Chen, “Coding or not: Optimal mobile data offloading in opportunistic vehicular networks,” *IEEE Tran. Intelligent Transportation Systems*, 2013.

[81] A. Pyattaev, K. Johnsson, S. Andreev, and Y. Koucheryavy, “3GPP LTE traffic offloading onto WiFi direct,” in *Proc. of IEEE WCNC*, 2013.

[82] S. Singh, H. Dhillon, and J. Andrews, “Offloading in heterogeneous networks: Modeling, analysis, and design insights,” *IEEE Tran. Wireless Communications*, vol. 12, no. 5, 2013.

[83] J. Lee, Y. Yi, S. Chong, and Y. Jin, "Economics of WiFi offloading: Trading delay for cellular capacity," in *Proc. of IEEE Infocom Workshop SDP*, 2013.

[84] D. Kim, Y. Noishiki, Y. Kitatsuji, and H. Yokota, "Efficient ANDSF-assisted Wi-Fi control for mobile data offloading," in *Proc. of IWCMC*, 2013.

[85] A. Y. Ding, B. Han, Y. Xiao, P. Hui, A. Srinivasan, M. Kojo, and S. Tarkoma, "Enabling energy-aware collaborative mobile data offloading for smartphones," in *Proc. of IEEE SECON*, 2013.

[86] J. Kim and N. Song, "Placement of WiFi access points for efficient WiFi offloading in an overlay network," in *Proc. of IEEE PIMRC*, 2013.

[87] S. Dimatteo, P. Hui, B. Han, and V. Li, "Cellular traffic offloading through WiFi networks," in *Proc. of IEEE MASS*, 2011.

[88] D. Zhang and C. K. Yeo, "Optimal handing-back point in mobile data offloading," in *Proc. of IEEE VNC*, 2012.

[89] S. Wietholter, M. Emmelmann, R. Andersson, and A. Wolisz, "Performance evaluation of selection schemes for offloading traffic to IEEE 802.11 hotspots," in *Proc. of IEEE ICC*, 2012.

[90] K. Berg and M. Katsigiannis, "Optimal cost-based strategies in mobile network offloading," in *Proc. of ICST CROWNCOM*, 2012.

[91] U. Yechiali, "A queueing-type birth-and-death process defined on a continous-time Markov chain," *Operations Research*, vol. 21, no. 2, 1973.

[92] F. Mehmeti and T. Spyropoulos, "Is it worth to be patient? Analysis and optimization of delayed mobile data offloading," in *Proc. of IEEE Infocom*, 2014.

[93] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Incentivizing time-shifting of data: a survey of time-dependent pricing for internet access," *IEEE Communications Magazine*, vol. 50, no. 11, 2012.

[94] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "Tube: time-dependent pricing for mobile data," in *Proc. of ACM SIGCOMM*, 2012.

[95] C. Joe-Wong, S. Ha, and J. Bawa, "When the price is right: Enabling time-dependent pricing for mobile data," *ACM SIGCHI 2013.*

[96] D.Y.Barrer, "Queuing with impatient customers and ordered service," *Operations Research*, 1957.

[97] R.E.Stanford, "Reneging phenomena in single channel queues," *Mathematics of Operations Research*, 1979.

[98] N. Perel and U. Yechiali, "Queues with slow servers and impatient customers," *European Journal of Operations Research*, no. 201, 2010.

[99] E. Altman and U. Yechiali, “Analysis of customers’ impatience in queues with server vacations,” *Queueing Systems*, vol. 52, 2006.

[100] A. Sharma, V. Navda, R. Ramjee, V. N. Padmanabhan, and E. M.Belding, “Cool-tether: Energy efficient on-the-fly WiFi hot-spots using mobile phones,” in *Proc. of ACM CoNEXT*, 2009.

[101] J. P. Singh, T. Alpcan, P. Agrawal, and V. Sharma, “A Markov decision process based flow assignment framework for heterogeneous network access,” *Wireless Networks*, no. 16, 2010.

[102] Y. Im, C. J. Wong, S. Ha, S. Sen, T. Kwon, and M. Chiang, “AMUSE: Empowering users for cost-aware offloading with throughput-delay tradeoffs,” in *Proc. of IEEE Infocom Mini-conference*, 2013.