TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Olga Galinina

Analytical Performance Evaluation of Cooperative and Multi-Radio Concepts in Emerging Wireless Networks

Julkaisu 1320 • Publication 1320

Tampere 2015

Tampereen teknillinen yliopisto. Julkaisu 1320
Tampere University of Technology. Publication 1320

Olga Galinina

Analytical Performance Evaluation of Cooperative and Multi-Radio Concepts in Emerging Wireless Networks

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Sähkötalo Building, Auditorium SJ 204, at Tampere University of Technology, on the 1st of September 2015, at 12 noon.

Tampereen teknillinen yliopisto – Tampere University of Technology
Tampere 2015

Supervisor:

Yevgeni Koucheryavy, Ph.D., Professor
Department of Electronics and Communications Engineering
Tampere University of Technology
Tampere, Finland

Instructor:

Sergey Andreev, Ph.D., Senior Researcher
Department of Electronics and Communications Engineering
Tampere University of Technology
Tampere, Finland

Pre-examiners:

Olav Tirkkonen, Ph.D., Associate Professor
Department of Communications and Networking
Aalto University
Helsinki, Finland

Luis Manuel de Jesus Sousa Correia, Ph.D., Professor
Instituto Superior Técnico
University of Lisbon
Lisbon, Portugal

Opponent:

Ekram Hossain, Ph.D., Professor
Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, Canda

ISBN 978-952-15-3564-2 (printed)
ISBN 978-952-15-4310-4 (PDF)
ISSN 1459-2045

ABSTRACT

For the past years, the analysts have been predicting a tremendous and continuous increase in mobile traffic, which substantiates the emergence of heterogeneous multi-radio networks supported by the trend in infrastructure and user densification. In the envisioned fifth-generation (5G) heterogeneous deployments, where each user device may utilize multiple radio access technologies (RATs) for communicating with the network infrastructure or other proximate devices, the challenge of *intelligent management* of such connections arises.

Assuming that cellular network is able to provide assistance in connectivity on multiple RATs, there potentially exists a variety of management strategies. In order to analyze them and subsequently optimize the network operation, an appropriate modeling tool is required, which would account for the specific properties of the involved communication technologies. In the light of this, it is essential to revisit the mathematical tools available today, augment them, and apply to the selected area.

While *queuing theory* has remained as one of the key solutions for network modeling for over a half-century, the contemporary wireless systems incorporated geographical location of a node as a crucial factor in determining the resulting quality of service. Consequently, shifting from the queuing theory to the area named *stochastic geometry* delivered useful operational insights for large macrocells. However, the continuing network densification as the mainstream trend in the development towards future 5G networks suggests focusing attention on the user loading together with its traffic dynamics. As a result, it is important to revisit the legacy tools of queuing analysis and *combine* them with appropriate methods of stochastic geometry in order to efficiently evaluate the ultimate system performance.

This thesis is devoted to developing a *novel space-time methodology* that flexibly combines spatial approaches of stochastic geometry with the investigation of network dynamics by queuing theoretic methods. This general approach allows for building useful *first-order performance estimations* for a wide range of system assumptions and constraints, which result from a particular multi-radio network configuration.

Conveniently offered in the form of a construction set, the proposed methodology has significant advantages over the existing methods, and may be successfully applied to a large number of novel problems in the field of wireless networking analysis, where each of them finds its own practical application in the emerging 5G ecosystem.

Preface

The research work described herein has been conducted in the Department of Electronics and Communications Engineering of Tampere University of Technology (Finland) over the years 2011-2015. This manuscript is to the best of my knowledge original, and neither this nor substantially similar dissertation has been submitted for any other degree or another qualification at any other university.

This work would not have been possible without the contribution of many wonderful people. Without their support, I would have never completed this dissertation. I thank all of them, and particularly the ones below.

I have been extremely fortunate to work under the supervision of Prof. Yevgeni Koucheryavy, who has improved my research capabilities significantly. I would like to thank him for invaluable lessons about the importance of guidance and for the freedom he granted to me in selecting my research direction. Also I would like to express my deepest appreciation to Dr. Sergey Andreev for his patience, being supportive and, most importantly, for being always there for me during many years. He has been encouraging me to not only grow as a researcher but also as an independent thinker with wide-open eyes.

I am indebted to Prof. Andrey Turlikov (Russia) for sharing expertize, his sincere and valuable guidance and teaching me the importance of using curiosity as the driving force behind research. Despite growing responsibility for his own department, he has been always willing to take time for our long discussions and share inspiring ideas.

I would like to express my gratitude to the reviewers of this thesis, Prof. Olav Tirkkonen (Finland) and Prof. Luis Manuel de Jesus Sousa Correia (Portugal) for sharing their views on my work. The manuscript has definitely benefited from their broad perspective, valuable suggestions, and constructive feedback. A special mention goes to Prof. Ekram Hossain (Canada) for agreeing to act as opponent at my defense.

As financial stability is an important aspect of any solid research, I gratefully acknowledge the generous support received from Graduate School in Electronics, Telecommunications and Automation (GETA) for the four years of my research work in Tampere and Centre for International Mobility (CIMO) program which had given me a life-changing opportunity to discover Tampere University of Technology.

I would like to thank all my colleagues and wonderful friends for creating warm and supportive environment. This includes (in alphabetical order) Amir Mehdi Ahmadian, Roman Florea, Mikhail Gerasimenko, Regina Gumenyuk, Dmitri Moltchanov, Ekaterina Olshannikova, Alexandr Ometov, Vitaly Petrov, Alexey Ponomarenko, Andrey Samuilov, Dmitry Solomitskiy, Natalia Troitskaya, Jani Urama. Special thanks to Alexander Pyattaev, with whom we have been puzzling over many problems, for sharing expertize, warm support and his truthful and illuminating views on a number of issues related to my research.

Naturally, I would like to extend my appreciation to Ulla Siltaloppi, Soile Lönnqvist, Elina Orava, Tuija Grek and Heli Ahlfors for their responsiveness, prompt assistance with practical matters, and friendly support.

Finally and most importantly, I take this opportunity to express my gratitude to my family and thank my parents, Lioudmila and Sergey, for their everlasting love, unfailing encouragement, for their faith in me, understanding and support.

Olga S. Galinina

September 1, 2015, Tampere, Finland

Table of Contents

List of Publications

This thesis is mainly based on the following publications:

[P1] O. Galinina, S. Andreev, and Y. Koucheryavy, ”Performance Analysis of Client Relay Cloud in Wireless Cellular Networks”, in *Proc. of the 10th International Conference on Wired/Wireless Internet Communications (WWIC)*, 2012.

[P2] O. Galinina, A. Trushanin, V. Shumilov, R. Maslennikov, Z. Saffer, S. Andreev, and Y. Koucheryavy, ”Energy-Efficient Operation of a Mobile User in a Multi-Tier Cellular Network”, in *Proc. of the 20th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA)*, 2013.

[P3] O. Galinina, A. Anisimov, S. Andreev, and Y. Koucheryavy, ”Performance Analysis of Uplink Coordinated Multi-Point Reception in Heterogeneous LTE Deployment”, in *Proc. of the 11th International Conference on Wired/Wireless Internet Communications (WWIC)*, 2013.

[P4] O. Galinina, A. Turlikov, S. Andreev, and Y. Koucheryavy, ”Stabilizing Multi-Channel Slotted Aloha for Machine-Type Communications,” in *Proc. of the IEEE International Symposium on Information Theory (ISIT)*, 2013.

[P5] O. Galinina, S. Andreev, M. Gerasimenko, Y. Koucheryavy, N. Himayat, S. Yeh, and S. Talwar, ”Capturing Spatial Randomness of Heterogeneous Cellular/WLAN Deployments With Dynamic Traffic”, *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 1083-1099, 2014.

[P6] O. Galinina, S. Andreev, A. Turlikov, and Y. Koucheryavy, ”Optimizing Energy Efficiency of a Multi-Radio Mobile Device in Heterogeneous Beyond-4G Networks”, *Performance Evaluation*, vol. 78, pp. 18-41, 2014.

[P7] S. Andreev, M. Gerasimenko, O. Galinina, Y. Koucheryavy, N. Himayat, S. Yeh, and S. Talwar, ”Intelligent Access Network Selection in Converged Multi-Radio Heterogeneous Networks”, *IEEE Wireless Communications*, vol. 21, pp. 86-96, 2014.

[P8] O. Galinina, A. Pyattaev, S. Andreev, M. Dohler, and Y. Koucheryavy, ”5G Multi-RAT LTE-WiFi Ultra-Dense Small Cells: Performance Dynamics, Architecture, and Trends”, *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 1224-1240, 2015.

List of Abbreviations

3GPP	3G Partnership Project
5G	Fifth Generation
AP	Access Point
BS	Base Station
CSMA	Carrier Sense Multiple Access
D2D	Device-to-Device
DL	Downlink
HetNet	Heterogeneous Network
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
LTE	Long Term Evolution
M2M	Machine-to-Machine
MIMO	Multiple-Input and Multiple-Output
PPP	Poisson Point Process
QoS	Quality of Service
RAT	Radio Access Technology
RV	Random Variable
SNR	Signal to Noise Ratio
SINR	Signal to Interference-plus-Noise Ratio
UE	User Equipment
UL	Uplink
WAP	Wireless Application Protocol
WLAN	Wireless Local Area Network
WPAN	Wireless Personal Area Network

List of Symbols

Symbol	**Units**	**Description**
p_i	W	Transmit power of the user i
w	Hz	Channel bandwidth
N_0	W	Noise level
I	W	Interference level
κ	-	Propagation exponent
G	m^{κ}	Propagation constant
r_i	bps (nats/s)	Instantaneous data rate of the user i
$r_i^{\max}$	bps	Instantaneous data rate of the user i at the maximum power
$r_{\lim}$	bps	Maximum achievable data rate
R	m	Circular cell radius
T	-	Number of tiers in heterogeneous network
r_0	bps	Target bitrate
λ	user/s^{-1}	User arrival rate to the system
λ_i	user/s^{-1}	User arrival rate at the tier number i
λ_T	user/s^{-1}	User arrival rate at the final macro tier T
μ^{-1}	s	Average session duration
L_i	BS/m^{-2}	Infrastructure density for the tier i
$d_{i,j}$	m	Distance between the transmitter i and the receiver j
δ	-	Resource available for sharing at the receiver

List of Figures

Chapter 1

Introduction

1.1 INTRODUCTION AND RESEARCH MOTIVATION

Looking at more than 40 years[1] of mobile communications history, from analogue to LTE and beyond, we confirm that network generations evolve over approximately 20-year intersecting cycles. On average, during around a half of that time, a new technology develops all the way to its peak, and through the rest of that time – from its peak to the last subscriber. Given that the beginning of every generation is based on several years of prior research and development, the emerging fifth generation (5G) of wireless systems is expected to be deployed sometime around 2020 [1]. However, yet there is no consensus on what comes after the state-of-the-art networking technologies, and what exactly 5G is.

Historically, every next generation is pushed by the consumer demands and the weaknesses of the previous generation. To this end, see examples of mobile networking evolution in Table 1.1. Hence, we expect that 5G will develop similarly and thus the associated performance criteria are already being shaped, with some of them available as [2]:

- ”anytime, anywhere” connectivity
- high data rate in the field (1-10 Gbps)
- extremely low latency (1 ms or less)
- high availability and reliability in some scenarios (99.999%)
- reduced power consumption (up to 90%)
- massive connectivity (10-100 times growth in connected devices)

[1] The first mobile call was made on April 3, 1973.

Table 1.1 Evolution of technology generations in terms of services and performance. Source: GSMA Intelligence [2]

	Primary services	Key differentiator	Weakness
1G	Analogue phone calls	Mobility	Poor spectral efficiency, major security issues
2G	Digital phone calls and messaging	Secure, mass adoption	Limited data rates – difficult to support demand for Internet
3G	Phone calls, messaging, data	Better Internet experience	Real performance failed to match hype, failure of WAP for Internet access
3.5G	Phone calls, messaging, broadband data	Broadband Internet, applications	Tied to legacy, mobile specific architecture and protocols
4G	All-IP services (including voice, messaging)	Faster broadband Internet, lower latency	...not yet defined

Some of these technical requirements, such as high rate and low latency, are already underway as a part of the development by the network operators. The rest of them are excessively diverse for a single breakthrough technology, and therefore 5G is expected to become different from other past generations (see [3], [4], and [5]) – one radio access technology (RAT) might not be sufficient. Today, many agree [6], [7] that prospective 5G networks will rather become a *blend* based on the convergence of several existing RATs (such as 3G, 4G, WiFi, and others), delivering higher user rates, better connectivity and coverage.

Hence, what emerges as a natural key question and the fundamental challenge for the 5G design, is the appropriate RAT management strategies [8], which help to make sure that the network operates in the best possible way. This implies that the network would need to show a certain system flexibility, or *intelligence*, to adjust to the diverse set of requirements, environmental and deployment conditions.

In order to develop and implement the demanded intelligent system, at the research stage we need a powerful methodology, which could focus on the most challenging 5G targets, such as low latency and high data rates, as well as would be able to quickly predict the performance of the considered 5G system, depending on its various conditions. Although the simulation tools may become of a considerable help here, the envisioned network densification and connectivity requirements call for a numerically-simple *first-order performance estimator* able to deliver fine-grained operational statistics. We continue by reviewing some methods of applied mathematics in a search of the appropriate methodology.

Turning back to the history of *analysis* in the broad area of communications, we notice that the *queuing theory* has been a key solution in respective performance evaluation for over half a century. However, the introduction of *wireless* technology brought novel research demand, so that the geographical locations of communication entities have become a crucial factor in determining the resulting quality of service (QoS). Accordingly, shifting from the queuing theoretic tools (see e.g., [9], [10], [11], and [12]) to the area of *stochastic geometry* [13] suits well for assessing large macrocells. However, with further network densification, we need to draw attention to the actual user loading and its uplink traffic dynamics, reiterating the importance of queuing analysis and thus *combining* these two crucial methodologies, queuing theory and stochastic geometry, in order to efficiently estimate the 5G system performance.

1.2 GENERAL BACKGROUND

Since 1909, when Erlang published his first paper [14], communications systems have been mostly assessed with the methods of "classical" queuing theory (see [15], [16], and [17]), which has been extensively developed driven by the technology evolution. One century later, however, with the emergence of wireless technology, this well-studied area became insufficient, since wireless networks are fundamentally limited by the intensity of the signal and interference (see [18], [19], and [20]), which in a wireless channel heavily depend on the geographical location. Driven further by wireless networking problems for several past decades, the research community has been (re)discovering the mathematical methods that allow taking into account the effects of network geometry – eventually arriving at stochastic geometry formulations, random graphs theory, and their sub-branches, such as point process theory, geometric probability, and percolation theory.

The term "stochastic geometry" has appeared already in 1963 [21], when studying the average system behavior over many spatial realizations. However, the underlying point process theory originates from the same year as Erlang publishes his paper, in the pioneering work [22], where the basic shot-noise process statistics have been delivered. Additional historical facts on geometric probability date back to as far as 18th century, and the mentioning of stochastic geometry may be found in [23] or [24], whereas one may refer to [25] for the comprehensive introduction into the stochastic geometry.

Therefore, both revived and novel spatial methods have been widely investigated in their application to ad-hoc and cellular networks, femtocells, cognitive radio, etc. They have thus provided important insights into characterizing the link budget (including interference, pathloss, and fading), as well as covered the connectivity and outage probability, capacity and other fundamental limits of wireless networks. However, with the current trend for network densification, 5G cellular networks may become substantially underutilized [26] except for highly congested areas during specific times of the day (commuting hours, public events, etc.). Hence, the performance of the network is increasingly dependent on the user loading, in addition to their geographical locations. As a result, it becomes evident that capturing the network dynamics together with its spatial features develops into a pressing demand.

Inspired by the above, the thesis proposes a flexible and simple approach to handling network spatiality with the use of stochastic geometry, while explicitly investigating network dynamics by queuing theoretic methods.

1.3 SCOPE OF THE THESIS

This thesis concentrates on assessing the cooperative and multi-radio concepts in the emerging heterogeneous networking architectures for stochastic user loads and locations, and by considering various degrees of *cellular assistance.* Along these lines, both (i) real-time data sessions with fixed bitrate and (ii) elastic traffic with files of random size are taken into account with the emphasis on the uplink performance in terms of the blocking probabilities[2], the average numbers of users or the average transmission times, and the energy consumption.

As a result, we deliver a unified *space-time methodology* for characterizing the operation of a heterogeneous network that is capable of cooperative transmission across multiple RATs. The proposed methodology is broad enough to accommodate various offloading scenarios, radio selection algorithms, user performance characteristics, and advanced wireless technologies.

Summarizing, this thesis targets various aspects of cooperative and multi-radio communications in the context of first-order analysis of next-generation wireless networks. The ultimate goal of this research is to extend the past, disjoint evaluation methodologies with respect to the requirements of emerging 5G systems. In order to achieve the objectives of this study, we extensively rely on advanced analytical techniques and confirm their applicability with supportive simulations. By that, we address the following challenges:

1. *General methodology for network performance evaluation:* We formulate a novel methodology that employs spatial processes and explicitly captures network dynamics providing distinct classification of emerging wireless systems.

2. *Example performance characterization for several types of heterogeneous networks:* We model an environment, where the conventional, multi-radio, and cooperative networks employ admission control, power control, and scheduling of ongoing user sessions as well as take advantage of interference coordination.

The output of this work provides useful insights into the integrated methods of heterogeneous networking analysis, while the accompanying analytical framework may serve as a useful reference point for ongoing 5G discussions in both academia and industry (e.g., in the 3GPP community), as well as offer a range of novel 5G-centric problem formulations.

1.4 THESIS OUTLINE AND MAIN RESULTS

This thesis consists of an introductory part comprising *seven* chapters and of *eight* main publications referred to as [P1]-[P8]. Additionally, the scope of this work is closely related to several other publications by the author, which are referred to in the bibliographical section of this manuscript.

[2] That is, when a user session is not admitted by the network.

In Chapter 1, we start with the core motivation behind our research and then continue with the scope of this work by highlighting the key problems addressed in the thesis.

In Chapter 2, we emphasize the importance of multi-radio and cooperative communications and provide the reader with the related background.

In Chapter 3, we classify the types of cooperative and multi-radio heterogeneous networks, as well as outline our generic system model assumptions.

In Chapter 4, we conduct analytical performance evaluation of some characteristic types of emerging heterogeneous networks for different structures of traffic arrival processes.

Chapter 5 offers supplementary system-level simulation results confirming the validity of the adopted analytical assumptions.

Chapter 6 concludes the introductory part.

Finally, Chapter 7 summarizes the publications constituting the second part of this thesis and highlights the author's contribution to them.

In summary, the **main contribution** of this thesis is a novel space-time mathematical model for traffic arrivals in a heterogeneous network, which includes (i) an integrated methodology for *assisted* network selection capturing the spatial randomness of cooperative and multi-radio systems together with dynamic uplink traffic; (ii) an example of in-depth *analytical* characterization of dynamic interactions between co-existing heterogeneous network tiers.

Chapter 2

Cooperative and Multi-Radio Networks

2.1 GENERAL BACKGROUND

Today, existing RATs are seeing high diversity in the data rates and suffering from excessive time delays, or sometimes even service outage due to poor coverage and harsh interference conditions. Cellular coverage also remains unsatisfactory in indoor environments despite aggressive spectrum reuse and very sophisticated techniques for interference coordination [27], [28]. To make matters worse, unprecedented numbers of diverse machine-type devices [29] connect to the network forming what is known as the Internet of Things (IoT) and thus reshape the global networks. All these imbalances with respect to the aforementioned 5G requirements accentuate the need to explore novel, more efficient technologies [30].

One solution to mitigate the increasing disproportion between the desired user QoS and the available wireless resources may be found in deploying the higher density of *small cells* in current cellular architecture (see e.g., [31], [32], [33], and [34]). This improves network capacity by increasing the frequency reuse per unit area and the average data rate per transmission (i.e., smaller cells yield shorter radio links and thus improve data rates). Moreover, as cell sizes shrink, the footprints of cellular, local, and personal area networks are increasingly overlapping, which creates an opportunity to simultaneously utilize multiple RATs for improved capacity and connectivity [35], [36].

Further, in the presence of continuously growing mobile traffic demand (for detailed numbers, see the report in [37]), another solution may be seen in offloading some user sessions onto direct device-to-device (D2D) radio links, which are generally shorter and lower-to-the-ground than small cell connections [38]. With D2D, neighboring wireless devices can communicate without the use of network infrastructure, enabling a dramatic improvement in spectral reuse [39]. In addition, the proximity of user devices promises higher data rates, lower transfer delays, and reduced power consumption [40]. The potential applications of D2D in cellular networks are numerous [41] and include local voice service (offloading calls between proximate users), multimedia content sharing, gaming, group multicast, context-aware applications, and public safety.

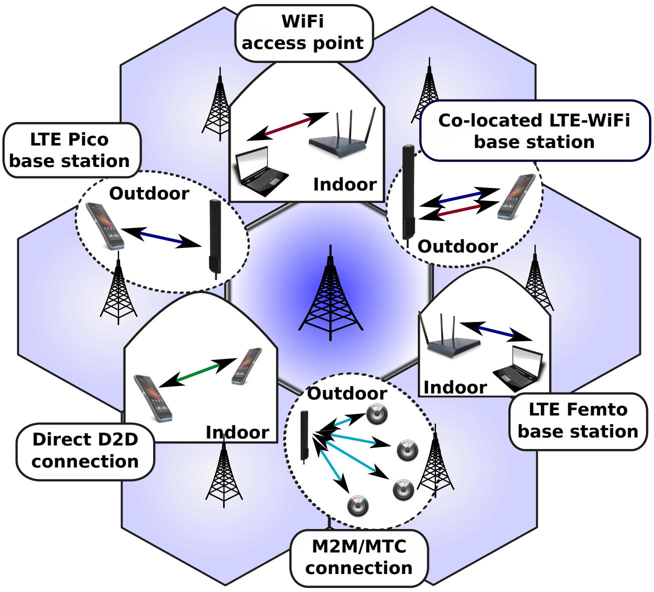


Figure 2.1 Envisioned 5G heterogeneous network.

Consequently, the incentive to efficiently coordinate between the alternative RATs is growing stronger [42]. To this end, the distributed unlicensed-band network (e.g., Wireless Local Area Network, WLAN) may take advantage of the *centralized* control function residing in the cellular network to effectively perform dynamic multi-RAT network association. However, very limited research attention has been dedicated to the *assisted* joint use of multiple networks, whereas much effort has been invested into optimizing the performance of individual radio technologies.

In summary, it is currently expected that the majority of near-term capacity gains will come from advanced architectures and protocols that would leverage the unlicensed spectrum and take advantage of the intricate interactions between the device and the network, as well as between the devices themselves, across the converged *heterogeneous* deployments (see Figure 2.1). To this end, intelligent coupling between multiple RATs may leverage several dimensions of diversity, where both short- and long-range technologies may need to work collaboratively to realize the desired improvements in capacity and service experience.

2.2 5G TECHNOLOGY TRENDS: HETEROGENEOUS NETWORKS

A transformation of mobile user experience requires revolutionary changes in both network infrastructure and device architecture, where the user equipment (UE) is jointly optimized with the surrounding network context. Consequently, tighter interworking between various RATs has been receiving more momentum over the past few years [43]. While previously cellular and WLAN technologies were developing largely independently, today WiFi is becoming an integral part of an operator's cellular network. As a result, it becomes crucial to aggregate different radio technologies as part of a common *converged* radio network, in a manner transparent to the end user, and develop techniques that can efficiently utilize the radio resources available across different spectral bands potentially using various RATs [26], [44].

In light of the above, heterogeneous networks (HetNets) represent advanced networking architectures (see Figure 2.1) enabling capacity and coverage improvements towards future 5G systems [45], [46], [47]. These architectures comprise hierarchical deployments of wide-area macro cells for basic connectivity and coverage augmented with (densely deployed [48], [49], [50], [51], [52], [53]) small cells of various footprints and by different RATs (femto and pico cells [54], [55], [56], WiFi access points [57], relay nodes [38], integrated WiFi-LTE small cells [P8], etc.) to boost capacity [58]. In particular, unlicensed-band technologies are increasingly managed as part of an operator's cellular network to unlock advanced levels of interworking between cellular and WLAN RATs. This is on the one hand due to the fact that contemporary consumer devices massively support WiFi together with other RATs. On the other hand, mobile network operators increasingly rely on WLAN-based offloading to relieve congestion on their cellular networks [59] and hence desire more control of how WLAN is utilized and managed.

Not surprisingly, recent literature has been very rich in addressing the important aspects of load balancing and access network selection for multi-RAT HetNets [60], [61], [62], [63], [64], [65]. The existing publications range from considering simpler user-centric network selection strategies (known as vertical handover [57]) to full multi-tier and multi-radio cooperation [66], [67], e.g., where WiFi becomes a "virtual carrier" anchored on the cellular network. However, the focus has been mostly on centrally-managed systems with full control at the base station or totally distributed solutions, but not so much on network-assisted schemes. Most recently, the concept of LTE-unlicensed has attracted interest of industry and academia alike with the goal of allowing LTE systems to utilize bandwidth-rich unlicensed spectrum around 5 GHz band to augment their capacity [68]. Another emerging industry trend considered in latest publications is multi-radio small cells with *co-located* cellular and WLAN interfaces able to reduce deployment costs and leverage common infrastructure across heterogeneous cells [69], [70], [71].

Reacting to this recent interest, 3GPP is becoming increasingly active in developing new interworking solutions between their cellular technologies, such as UMTS or LTE, and WiFi (IEEE 802.11 [72]) technology. However, given that co-located cellular/WLAN deployments are presently not common, current standardization efforts focus more on user-centric interworking architectures, while only assuming limited degrees of cooperation/assistance across the HetNet [73], [74]. The field of investigation spans across (i) schemes for trusted access to 3GPP services with WLAN devices, (ii) support for Access Network Discovery and Selection functions, and (iii) seamless mobility between cellular and WLAN technologies.

More recently, several new study/work items have been open targeting the interworking solutions that involve cooperation within the Radio Access Network (RAN) [75] by contrast to prior schemes that have loosely defined functions within the 3GPP core network (such as security and inter-RAT mobility) [76]. This shift is dictated by the need to support improved QoS on WLAN networks as prescribed by a consortium of network operators with their tighter requirements for carrier-grade WiFi. The WLAN community has also responded with their new initiatives on Hot Spot 2.0, as well as an emerging "High Efficiency WLAN" effort by the IEEE 802.11 work group. Therefore, we expect the trend for tighter integration of cellular and WLAN technologies to continue by potentially encompassing other radio technologies beyond current WiFi (e.g., mmWave-based small cells) and additional use cases beyond spectrum aggregation.

However, introducing an increasing number of serving stations to bridge the capacity gap incurs extra complexity due to more cumbersome interference management [77], [78], higher rental fees, and increased infrastructure maintenance costs [79]. Even when additional spectrum is allocated, these new frequencies are likely to remain fragmented and could require diverse transmission technologies. We, therefore, expect that the majority of near-term gains will be made available by efficient architectures and protocols leveraging the unlicensed spectrum. For example, mobile users with direct D2D communication capability may take advantage of their unlicensed-band radios and cooperate with other proximate users to locally improve access in a cost-efficient way [80].

2.3 5G TECHNOLOGY TRENDS: DIRECT COMMUNICATIONS

Currently, a major portion of the expected mobile traffic growth comes from peer-to-peer (P2P) services that involve clients in close proximity [81]. Hence, we envision that whenever possible, neighboring client devices will use their direct connectivity capabilities, instead of their cellular links. Consequently, D2D connections are believed to become an effective solution that would unlock substantial gains in capacity [82] and relieve congestion [83] in future 5G networks. For mobile network operators, D2D connectivity is becoming vital to enable traffic offloading from the core network [80] as well as realize efficient support of social networking through localization.

Over the last decade, much research effort has been invested into the characterization of D2D connections as part of LTE cellular technology [84] by 3GPP in licensed bands, where a license grants a network operator the right to use spectrum exclusively. Driven by a wealth of potential practical applications, the concept of D2D communication as an underlay to a cellular network has been developed by the seminal work in [85] and numerous subsequent papers. As in cognitive radio, D2D underlay is operating on the same resources as the cellular network and D2D users control their transmit power to suppress the resultant interference to the cellular users [86].

With the increasing number of cellular users, network-assisted D2D communication is becoming an essential next step to achieve enhanced resource utilization as the traditional methods to improve the use of licensed spectrum approach their theoretical limits [87]. Consequently, there has already been some coverage in literature on direct user connectivity with different levels of network involvement, ranging from the minimal degrees of assistance (such as in Aura-net/FlashLinQ) [88] to the fully controlled solutions (such as in cellular underlay) [85]. The latter is definitely more challenging and generally requires interference control to enable simultaneous direct connections [89].

For the underlay to work, the network should employ proper admission and power control on D2D transmitters as well as allocate radio resource to them. As a result, D2D links may (i) reuse resources reserved for cellular use, (ii) use free resources not allocated for cellular use, or (iii) relay traffic through the infrastructure network avoiding direct transmissions. The choice between these alternatives is known as transmission mode selection [90] and has attracted many researchers focusing on various optimization targets, from signal to interference plus noise ratio (SINR) and throughput to energy efficiency, data delay, fairness, and outage probability. The general difference between existing works is in the considered numbers of communicating entities of each type (base stations, cellular and D2D users), emphasis on uplink (UL) or downlink (DL) connection and the resulting interference, orthogonal vs. non-orthogonal resource sharing, degree of available network assistance, and network/D2D duplexing mode.

Given its growing importance, the licensed-bands D2D is becoming an attractive research area, where many fundamental questions still remain open including the information-theoretic capacity of the D2D underlay. However, the corresponding standardization efforts are developing slowly, such that the respective products, which are employing the licensed-band D2D underlay, may not be on the market until a long time from now.

Alternatively, unlicensed bands can be used freely, which gives opportunity to leverage D2D benefits almost immediately. Whereas there already exists a plethora of unlicensed spectrum protocols to technically enable direct connectivity, there is no centralized control of radio resources to manage QoS on D2D links [91]. Augmenting the current technology, we envision that devices be continually associated with the cellular network and use this connectivity to help manage their D2D connections in unlicensed bands. Therefore, as has been the case for HetNets, in the near-term we expect that the majority of gains will come from advanced architectures and protocols that would leverage the unlicensed spectrum.

In other words, in conventional WLANs, the access point (AP) has no measures to control the resources used by ad hoc user connections, which contend for the same channel. This is where the LTE network can be of much help. If clients are continuously connected to the LTE network, it knows which cell(s) they are associated with, which tracking area(s) they are in, and their locations within a few meters (if location services are enabled). Therefore, the network can quickly and without significant overhead determine if/when clients are potentially within D2D range and inform them accordingly. Additionally, network assistance can help with mode selection (LTE/WiFi), power control, and selecting transmission format (modulation and coding rates, MIMO transmission mode, etc.).

2.4 FOCUS AND CONTRIBUTIONS

In what follows, the focus is set on integration between multiple RATs within the envisioned 5G-grade HetNet architecture. As our case study, we consider convergence of WLAN- and D2D-based connections with operator-controlled cellular deployment, assuming that they belong to an operator deployed and *managed* multi-RAT HetNet.

We emphasize that interworking between different RATs has already been considered in the past, but largely from the perspective of inter-network (vertical) hand-off [57]. In addition, various specific concepts have been discussed to this end [92], [93], [94], including UMA/GAN (Unlicensed Mobile Access network later renamed to Generic Access Network), WWRF (Wireless World Research Forum) multi-radio considerations, WLAN integration with 3G systems, etc. Regarding the latter, cellular standards community, represented by the 3GPP, has also been involved in developing specifications that address cellular/WLAN interworking for a number of years. Several new study and work items have recently emerged to develop specifications towards tighter integration of WLAN with cellular networks.

However, we make a step ahead with respect to the current 3GPP and other efforts and consider *intelligent assistance* from the cellular network in the RAT selection process, when a new coordinating entity in the cellular RAN is made to receive relevant information from multi-radio devices (e.g., their position, QoS requirements, how much interference/load they sense on the nearby WLAN networks, etc.) and then advises the users on the attractive connectivity options.

For consistency with current network deployments, we concentrate on distributed small cell overlay with standalone WiFi access points as well as pico cell base stations [95], assuming that there is no direct interface between the cellular and WLAN radio networks. However, the presented methodology may also characterize co-located cellular/WLAN deployments as well as more advanced technologies and scenarios to become appealing in the context of 5G networks [96], [P8]. More specifically, we focus on *uplink* performance as it has not been fully addressed in existing literature due to more challenging interference-related aspects.

To further advance the state-of-the-art research primarily focusing so far on static (full-buffer) steady-state formulations, we target *flow-level* performance and consider stochastic traffic loads. In particular, new data flows representing, e.g., real-time data sessions with the minimum target bitrate are arriving randomly and leave the system after the service has been received. Consequently, the number of active flows varies with time, which is often referred to as the flow-level dynamics. Analyzing dynamic setups is important to gain better understanding of real-world systems, but it also incurs extra complexity. Therefore, dynamic systems had received much less research attention than their static alternatives, that is, with a fixed set of backlogged users.

In what follows, we outline a *general methodology* for modeling the operation of a converged multi-radio network that is capable of offloading user sessions onto small-cell and D2D connections across both licensed and unlicensed spectrum. The proposed methodology is broad enough to accommodate various offloading scenarios, radio selection algorithms, user performance characteristics, and advanced wireless technologies (e.g., WiFi and LTE).

Chapter 3

Comprehensive Methodology for Space-Time Network Analysis

3.1 CAPABILITIES OF THE PROPOSED MATHEMATICAL APPROACH

3.1.1 Capturing System Dynamics

Modern wireless networks are constantly evolving to enable better support for heterogeneous multimedia applications [97]. Since the integration of diverse services within a single radio platform is expected to result in higher operator profits and at the same time reduce network management expenses, intensive research efforts have been invested into the design principles of such networks [98]. However, as wireless resources are limited and shared by clients, service integration may become challenging [99], especially in HetNets [100], [101]. A key element in these systems is the packet scheduler, which typically helps ensure that the individual QoS requirements of wireless clients are satisfied. Such schedulers may be made opportunistic, i.e., primarily serving clients, which experience favorable channel conditions. Several attempts to investigate efficient opportunistic behavior while meeting diverse QoS demands of wireless clients have been made in [102], [103], [104], and [105].

More advanced QoS-constrained opportunistic frameworks for wireless cellular networks focus on *flow-level* performance [106] and consider stochastic traffic loads [107]. In particular, new data flows representing either real-time sessions or file transfer requests are arriving randomly and leave the system after the service has been received. Consequently, the number of active flows varies with time, which is referred to as the flow-level dynamics [108]. Analyzing dynamic setups is important to gain better understanding of real-world systems, but it also incurs extra complexity. Therefore, dynamic systems receive much less research attention than their static alternatives e.g., with a fixed set of backlogged clients.

Every data flow in a dynamic network may generally represent a stream of packets corresponding to a new file transfer, web-page browsing, or real-time voice/video session [109]. Originally, flow-level frameworks were helpful investigating flexible bandwidth allocation mechanisms in the context of wired systems. Extending their applicability to wireless networks, it was concluded that the throughput experienced by a dynamic user population can substantially differ from that received by a fixed number of users [110]. Consequently, studying dynamic wireless systems is becoming increasingly important and we concentrate on characterizing HetNet dynamics in what follows.

3.1.2 Capturing Topological Randomness

As it has already been mentioned above, another crucial aspect of HetNets is in that relative locations of the network users highly impact the resulting system performance [111]. Indeed, given that users are not regularly spaced, there may be a high degree of spatial randomness, which needs to be considered explicitly. Coupling such topological randomness with system dynamics requires a fundamental difference in characterizing user signal power and interference. Fortunately, the field of stochastic geometry provides us with a rich set of powerful results and analytical tools that can capture the network-wide performance of a random user deployment [20].

The use of stochastic geometry (that is, statistical modeling of spatial relationships) has become increasingly popular over the last decades to analyze network performance averaged over multiple spatial realizations. As part of a more recent surge, it has also been useful in assessing many important aspects of current cellular technology, from conventional macro cell deployments to hyper-dense heterogeneous and small cell networks [112], [113]. The application of stochastic geometry typically features a particular spatial point process to statistically capture, e.g., user locations yielding insights on the impacts of user density, transmit power, path loss, and interference.

On the other hand, the application of the queuing theory makes it possible to model user sessions arriving at random and leaving the system after being served. A *session* is a real-time data flow from one user to another; in this work, sessions are initiated according to a *Poisson point process* (PPP). This and other spatial processes have been used extensively to investigate the coexistence of cellular and mobile ad-hoc networks [101], study device discovery aspects of FlashLinQ [114], assess the performance of multi-tier heterogeneous cellular systems [115], cognitive femtocells [46], and even capture the distributions of transmit power and SINR in D2D networks [116]. However, in most cases, the use of stochastic geometry does not directly enable system dynamics.

As the overall performance of a converged HetNet depends considerably on the geographical locations of user devices, studying dynamic wireless systems *jointly* with their spatial features is becoming increasingly important. Therefore, the main target of this thesis is the development of the unified *space-time* evaluation methodology and the associated comprehensive system models that may be used for the performance assessment of the emerging 5G communication technologies.

3.2 PRELIMINARY DESCRIPTION

Here, we collect the well-known basic facts in order to advance the understanding of further discussions, as well as provide some examples, which are currently used in network modeling.

3.2.1 Point Processes

To capture the spatial diversity in the network under investigation, we exploit the powerful and well-studied tools of *point process theory*, which describes the stochastic processes in multidimensional space. The most important baseline example here is the PPP in two- or three-dimensional space, a particular case of which (in one-dimensional space) is the conventional Poisson process.

Let us assume that $N(A)$ is a finite random number of isolated points within the compact set $A \subset \mathcal{R}^n$, where n is the space dimension. The definition of the PPP comprises the following:

- for non-intersecting sets A_1 and A_2, the numbers of isolated points $N(A_1)$ and $N(A_2)$ are independent random variables (RVs),
- the distribution $\Pr\{N(A) = k\}, k \geq 0$ of RV $N(A)$ depends only on the area or volume $S(A)$ of A,
- if the area/volume $S(A)$ of A tends to zero, then $\Pr\{N(A) > 1\} = o(S(A))$, where $o(S(A))$ is infinitesimal with respect to $S(A)$.

As a consequence of these conditions, the number of points for the PPP is distributed according to a Poisson process with the parameter $\mu(A) = \lambda S(A)$, where λ is the density of PPP, and the expected number of points $\mu(A)$ is a measure of A. The PPP is thus a fundamental process, and for the subsequent network analysis it represents the easiest option in terms of calculation. It is thus often taken as the first modeling choice – mostly due to the fact that the realizations of the PPP within any bounded and closed region follow a uniform distribution.

Given the uniform nature of points within an area of interest, the PPP may be applied wherever the positions of network entities (transmitters and receivers) may be assumed "purely" random. Similarly, there also exists a Binomial Point Process with a difference in that it offers a fixed number of nodes within the given region. However, in the real networks the positions of the communicating entities might have a more complex structure and include a certain level of dependence in node locations. For example, 3GPP specifications suggest that the base stations cannot be deployed closer to each other than a particular distance threshold, which may be modeled according to a Hard-Core Point Process, where the minimum distance between the points is set.

Other highlights from the practical point of view are Matern and Thomas Cluster Processes, where the points are grouped around the independent cluster centers according to some distribution (uniform and Gaussian, correspondingly), which might be very useful in modeling clustered user deployments around WiFi access points

(APs). Finally, one example of a fairly tractable point process is Ginibre Process that models a so-called determinant repulsion between nodes. A variety of other area-interaction processes and simulation approaches may be found in [117] or more specific literature on stochastic geometry and point processes (see [118], [119], [120], [121], and the references therein). A detailed description of spatial point processes applied exclusively to the area of wireless communications may be found in [20] and [111].

3.2.2 Signal Propagation and Path Loss Model

The received signal quality significantly depends on the distance between the transmitter and its respective receiver, which in terms of network modeling directly follows from the corresponding point process. However, the distance is not the only factor determining the resulting signal quality. In order to simplify the modeling and omit the details of wave propagation, a range of "classical" solutions predicting the *average* signal strength has emerged over the years. Following the goals of this thesis, that is, to provide a first-order time-averaged performance evaluation, we only concentrate on considering the large-scale propagation models, by omitting the fading-related details in what follows. To pursue this topic further, the interested reader is advised to consult with [122].

The simplest propagation model is used in the line-of-sight (LOS) channels under ideal conditions and is given by the Friis free-space equation for the received power level $p_{rx}(d)$:

$$p_{rx}(d) = p_{tx}G_{tx}G_{rx}\frac{1}{L} = p_{tx}G_{tx}G_{rx}\left(\frac{\lambda}{4\pi d}\right)^2, \tag{3.1}$$

where d is the distance, p_{tx} is the transmit power, G_{tx}/G_{rx} is the transmitter/receiver gain, L is the path loss on the linear scale, and λ is the wavelength. The last part of the above constitutes the free-space path loss model $L = \left(\frac{\lambda}{4\pi d}\right)^{-2}$.

However, ideal LOS conditions are seldom the case in evaluating the more advanced cellular networks. As an example (more complex, but still classical), let us consider a widely-used Hata model for urban areas, which is applicable in the range from 150 MHz to 1500 MHz [122]. The equation for the path loss on the decibel scale is given as follows:

$$L_{dB} = 69.55 + 26.16\lg f_c - 13.82\lg h_{tx} - \alpha(h_{rx}) + (44.9 - 6.55\lg h_{tx})\lg d, \tag{3.2}$$

where f_c is the carrier frequency (in MHz), h_{tx}/h_{rx} is the transmitter/receiver antenna height, and $\alpha(h_{rx})$ is a correction factor:

$$\begin{aligned}
&\alpha(h_{rx}) = (1.1\lg f_c - 0.7)h_{rx} - (1.56\lg f_c - 0.8), \quad \text{small/medium city,}\\
&\alpha(h_{rx}) = 8.29(\lg 1.54h_{rx})^2 - 1.1, \quad \text{large city, } f_c < 300 \text{ MHz,}\\
&\alpha(h_{rx}) = 3.2(\lg 11.75h_{rx})^2 - 4.97, \quad \text{large city, } f_c \geq 300 \text{ MHz.}
\end{aligned} \tag{3.3}$$

Given a large number of related models, and depending on the considered RAT, any other methodology may be used, including the standardized urban models from 3GPP documentation. We note here that in the MHz band, such models have been well-investigated in the past and are built on extensive field measurements.

3.2.3 Data Transfer Rate

The use of an appropriate modulation and coding scheme allows for mitigating channel errors through adding some redundancy, therefore lowering the transmission rate. The connection between the channel conditions and the achievable channel capacity may thus be given by the Shannon's formula [123]:

$$c = w \log_2 \left(1 + \frac{p}{N}\right), \tag{3.4}$$

where p is the received signal power based on the path loss between the transmitter and the receiver, e.g., (3.1) or similar, while w is the bandwidth and N is the noise power, which in turn may include both noise and interference $N = N_0 + I$, where interference I is a sum of signals from the interfering transmitters.

3.3 GENERAL DESCRIPTION OF PROPOSED METHODOLOGY

We deliver our methodology in the form of a simple *construction set*, i.e., a collection of building blocks, which may be arbitrarily combined to construct the required system model and produce the corresponding solution. Below, we provide a brief description of how the proposed construction set works.

The main structure of our methodology, as well as its intermediate and related outputs are given in Figure 3.1. In particular, we differentiate between the *space unit* (SU), which corresponds to the user and infrastructure deployment (left), and the *time unit* (TU) related to the user arrivals, service, and departures (right). Both units consist of several sequential building blocks (e.g., distribution of the infrastructure), where for each block we have several alternative options to select from (e.g., PPP). After choosing exactly one option for each block in our classification, the outcomes of the SU (such as the distributions of distances, SINR, and the resource required for service) could be transferred to the TU in a form of transition probabilities for the core Markov process.

The general structure of the underlying Markov process is generated based on the corresponding selection in the TU. Combining the output distributions from the SU with the structure of the process in the TU, we arrive at the queuing theoretic problem formulation with known transition probabilities. Depending on the complexity of both components: (i) the expressions for transition probabilities and (ii) the structure of the Markov process, for the resulting problem formulation we may obtain a numerical or a closed-form solution. We emphasize here that depending on the selection of blocks, both components might lead to nontrivial formulations.

In what follows, we illustrate how the proposed construction set may be applied to analyzing the integrated HetNets, which may imply various density of infrastructure as well as different mechanisms for interaction between the communicating entities. In the absence of prior information about their locations, we exemplify below based on the PPP, as a simplest solution for both user and infrastructure deployment.

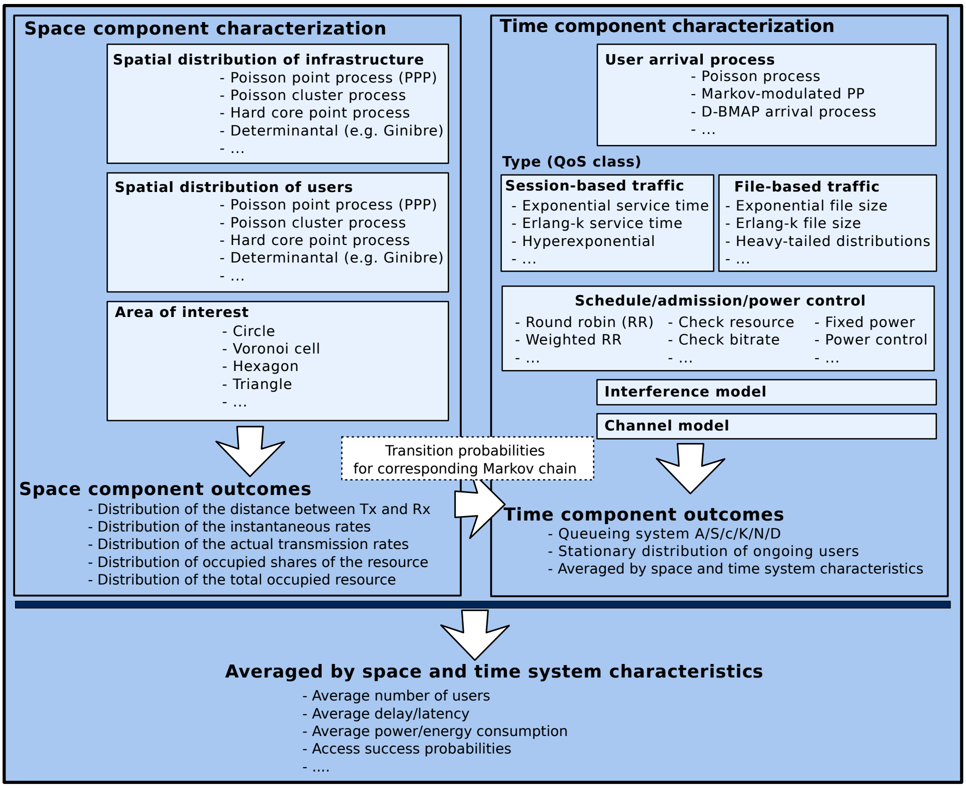

Space component characterization
Spatial distribution of infrastructure
- Poisson point process (PPP)
- Poisson cluster process
- Hard core point process
- Determinantal (e.g. Ginibre)
- ...
Spatial distribution of users
- Poisson point process (PPP)
- Poisson cluster process
- Hard core point process
- Determinantal (e.g. Ginibre)
- ...
Area of interest
- Circle
- Voronoi cell
- Hexagon
- Triangle
- ...
Space component outcomes
- Distribution of the distance between Tx and Rx
- Distribution of the instantaneous rates
- Distribution of the actual transmission rates
- Distribution of occupied shares of the resource
- Distribution of the total occupied resource
Transition probabilities for corresponding Markov chain
Time component characterization
User arrival process
- Poisson process
- Markov-modulated PP
- D-BMAP arrival process
- ...
Type (QoS class)
Session-based traffic
- Exponential service time
- Erlang-k service time
- Hyperexponential
- ...
File-based traffic
- Exponential file size
- Erlang-k file size
- Heavy-tailed distributions
- ...
Schedule/admission/power control
- Round robin (RR)
- Weighted RR
- ...
- Check resource
- Check bitrate
- ...
- Fixed power
- Power control
- ...
Interference model
Channel model
Time component outcomes
- Queueing system A/S/c/K/N/D
- Stationary distribution of ongoing users
- Averaged by space and time system characteristics
Averaged by space and time system characteristics
- Average number of users
- Average delay/latency
- Average power/energy consumption
- Access success probabilities
- ....

Figure 3.1 Construction set structure and outputs of the proposed methodology.

3.4 PROPOSED TAXONOMY FOR HETNETS

Here, we consider a characteristic HetNet example as well as demonstrate the methodology discussed above in its application to the generic HetNets. We begin with describing our proposed *classification* for the distinct types of HetNet tiers, embracing their distinguishing features from the analytical perspective.

To this end, we consider one tagged macrocell with the base station (BS) located at the center, which collects all the relevant control information and feedback as well as performs *network assistance* by making decisions and employing *admission control* mechanisms. Moreover, *for every tier* in such a HetNet, we differentiate between *three* main components (see axes in Figure 3.2) that primarily determine the corresponding mathematical constructs, namely:

- interference (insignificant, so that it may be neglected, vs. significant, which is to be accounted for explicitly),

- power control/resource allocation (varies from the fixed power allocation, i.e., absence of power control, to round-robin resource allocation, while an optimal power allocation may be located somewhere between these extremes [108]),

- resource utilization (dedicated resources per a communication link vs. shared channel access by several links).

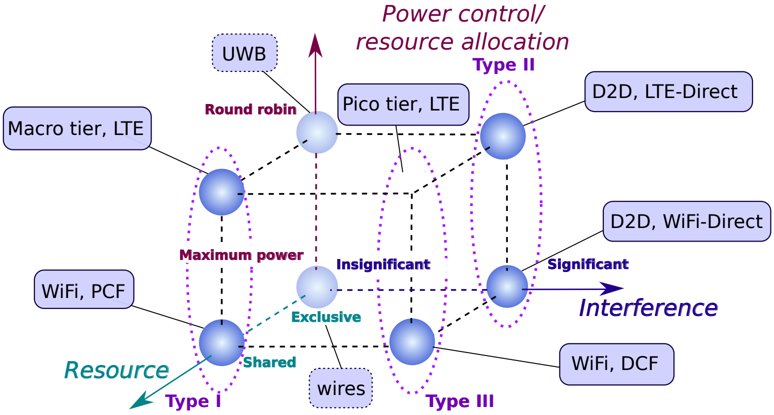


Figure 3.2 Proposed taxonomy of various HetNet types.

In Figure 3.2, we illustrate the three-dimensional space formed by the above three criteria and specify various HetNet types as examples in thus introduced space. Particularly, we split the provided points (examples) into three groups and further consider these groups separately as individual models:

- Type I (conditionally termed ’macro’): resource is shared between several links, impact of interference may be neglected due to the technology-related features (such as coordination, frequency reuse planning, tight beamforming, as well as other more recent and advanced techniques).

- Type II (conditionally termed ’D2D’): resource is exclusive for one link, but interference has to be taken into account.

- Type III (conditionally termed ’small cell’): resource is shared between several links, and interference between the neighboring cells has to be taken into account.

In more detail, we have enumerated the above *types* according to their increasing complexity (see Figure 3.3). The simplest, type I (’macro’), is equivalent to one cell under a macro BS coverage (one entity), where interference from other entities can be treated as the background noise due to sophisticated interference control procedures, which constrain the resource at the BS to be shared across all the users in service. A more complex scenario is named type II (’D2D’), when the resource is exclusive for a transmitter-receiver pair and thus cannot be shared with others, but the interference from other connections (entities) is considerable and has to be taken into account. Finally, the most complicated option, type III (’small cell’), implies significant interference between the communicating entities, as well as sharing the common resource by several transmitters.

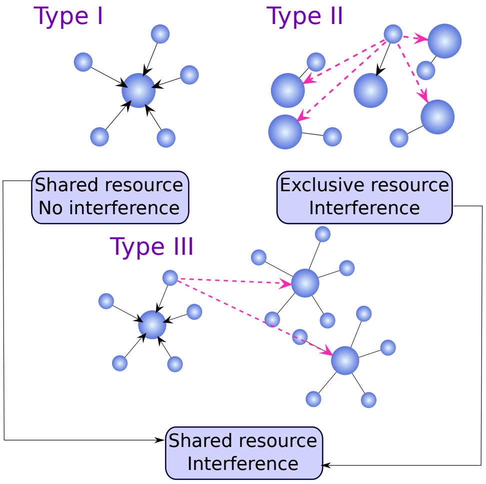


Figure 3.3 Illustration of our classification for various HetNet types.

Due to a wide variety of alternative power control mechanisms, we combine all the power allocation schemes into one group vertically (see Figure 3.2), and hereinafter refer to the *system types* as specified above. As an example, the point *WiFi, DCF* is based on IEEE 802.11 WLAN standard and corresponds to significant interference between the WiFi "cells", while the resource of a single AP is shared between several users with fixed transmit power. Another example is the point *WiFi, LTE-Direct*, where interference between the neighboring links may be high enough, but up to the entire uplink resource per link may be allocated to one transmitter exclusively.

3.5 GENERAL ASSUMPTIONS OF THE MODEL

In what follows, we introduce our integrated system model comprising a number of cellular macro- and small cells, WLAN, as well as D2D connections, which we refer to as tiers. Below, we summarize the core assumptions made throughout this thesis, by reflecting the elements of our construction set (see Figure 3.1).

We study one (typical) cell of a macro network with the radius of R, featuring a macro BS in its center. We assume that T radio networks (RATs) are available within the macrocell coverage. Every network is an instance of one of the above three categories (see, for example, Figure 3.4), that is, belongs to the type I, II, or III. All the networks are serving *uplink* data from their wireless users concurrently. For the sake of exposition, the considered traffic is characteristic of real-time sessions with the target bitrate of r_0.

Based on the recent specifications in [75], we further assume *non-overlapping* frequency bands for all tiers. Therefore, user transmissions on one tier do not interfere with those on the others. However, all links of types II and III (for instance, cellular small cells, WLAN, or D2D) share the frequency bands of their respective tiers and thus interfere, whereas the tier type I is interference-free. Our general system model is illustrated in Figure 3.4 representing the area of the macrocell as type I tier, small cell and WLAN coverage as type III tier, and D2D system as type II tier together with all the corresponding users and infrastructure nodes.

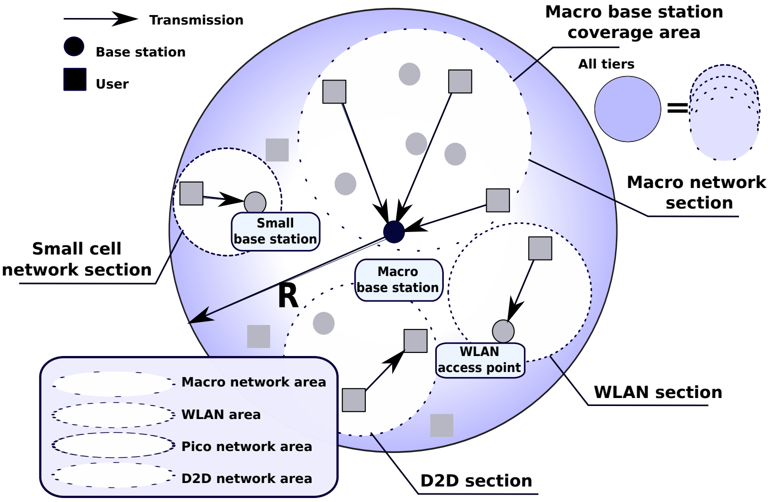


Figure 3.4 System model of a four-tier HetNet within a macro cell of radius R: the *cuts* demonstrate different network tiers.

We assume that the transmitting users (or, *sessions*) with some uplink traffic demand arrive into the joint network according to one-dimensional Poisson process of rate λ in time. We thus associate a newly arrived user with its session and its location, which is assumed fixed throughout the lifetime of the session. For the sake of tractability, we also assume that the duration of a user session is distributed exponentially with the mean of μ^{-1}, which may correspond to, e.g., a real-time voice/video call. To explicitly model system spatiality, we further make the following principal assumption.

Assumption 1. ***Spatial distribution of users.*** *The locations of arriving users follow a Poisson point process on the two-dimensional plane. The area of our interest is limited by the considered macro cell (e.g., circle of radius R), resulting in uniform distribution of users within the circle.*

Additionally, we adopt the following assumption about the locations of users and network infrastructure.

Assumption 2. ***Spatial distribution of infrastructure.*** *Type II. For every session i of the tier type II, we differentiate between the transmitting user T_i, which is the data originator, and the receiving user R_i, which is the respective destination. We further assume that for the transmitting user T_i, the corresponding receiving user R_i arrives simultaneously with T_i, such that the location of R_i is distributed uniformly within the same circle of radius R.*

Type III. The locations of receivers (e.g., APs/BSs) on the tier type III are independent and spatially distributed according to a PPP on the two-dimensional plane with the rates of L_i, where i is the sequential number of the corresponding tier ($1 \leq i < \infty$).

We note that in practice the constraint of deploying users within a particular area may be dictated by, e.g., maximum transmit power restrictions and/or channel degradation factors. Moreover, uniform distribution for the tier type II is only assumed here as a baseline example. Generally, we may consider any other distribution $f(x, y)$ of user locations, which would somewhat complicate further analysis technically, but without significant impact on the derivation methodology.

Assumption 3. ***Signal propagation.*** *For all tiers, we assume for tractability that for the ongoing session the wireless channel gain $\gamma_{k,j}$ between the user k and the receiver j depends on the distance $d_{k,j}$ separating them, and therefore it may be expressed as:*

$$\gamma_{k,j} = \frac{G}{{d_{k,j}}^{\kappa}}, \tag{3.5}$$

where $d_{k,j}$ is the distance between the receiver and the transmitting user, κ is the propagation exponent, and G is the propagation constant determined by a particular RAT and accounting for the corresponding channel model.

The above expression may directly follow from the selected propagation model (e.g., offered by 3GPP or as it has been described in Section 3.2.2). We continue by specifying the power-rate mapping.

Assumption 4. ***Power-rate mapping.*** *We assume that the data rate is continuous and that the power/rate mapping is given by the Shannon's formula [123], [124]. This consideration has been shown in [P2] to remain very accurate for current wireless networks. Hence, the transmit power p_i of a user i and its data rate r_i (in [nats/s]) are coupled by the generalized Shannon's capacity theorem:*

$$r_i = \min\{B \log\left(1 + Ap_i\right), r_{\mathrm{lim}}\}, \tag{3.6}$$

where p_i is the output power of the radio-frequency power amplifier, whereas A and B are the scaling coefficients that depend on the particular RAT used. For the sake of an example, these are given as $A = \frac{\eta\gamma_{i,i}}{N_0+I}, B = w$, where $\gamma_{i,i}$ is the path gain between the user and the receiver for session i, η is the fading margin, w is the channel bandwidth, N_0 is the noise level, and I is the interference level at the receiver.

The constant r_{lim} defines the constraint of the data rate growth based on the fixed set of modulation and coding schemes. Hence, further increase in the SNR does not yield the unbounded data rate increase after a certain value of $d_0 = \left[\frac{G \cdot p}{(N_0+I)\left(e^{r_{\mathrm{lim}}/w}-1\right)}\right]^{1/\kappa}$.

While random network topology is the primary focus of our model, we also investigate flow-level system dynamics. This involves an appropriate queuing model, where the session arrives and leaves the system after being served (the service time is determined by the random session length). When a new session arrives or a served session leaves the system, the centralized assisting entity in the cellular network performs *admission* and *power control* on all tiers by deciding whether the session would be admitted to a particular tier or not (admission control) and/or advising on the user's transmit power (power control).

For each of the three considered tier types, the corresponding data link is governed by applying certain *transmission policies.* A particular policy generally decides on user admission, scheduling, and transmit power. Whenever admitted, a transmitting user occupies a fraction of the time-frame resource and sets its power as commanded by the BS to achieve the data rate given by (3.6). The BS makes a new decision on scheduling allocations and transmission power for all active users at every new arrival or when an existing session is served and leaves the system.

Assumption 5. ***Power control and scheduling.*** *The considered transmission policies are the following.*

1. The Maximum Rate (MR) policy assumes no explicit power control and sets a fixed transmit power, which is the allowable maximum for a particular RAT. Then, admission control checks if the target bitrate can be achieved with this maximum power. Given the relationship in (3.6), the instantaneous data rate for the session k *is determined by the maximum transmit power* $p_{\max}$ *as:*

$$r_k^{\max} = \min\{w \log\left(1 + \gamma_{k,k} p_{\max}\right), r_{\lim}\}. \tag{3.7}$$

2. The Round Robin (RR) or Full Utilization (FU) policy ensures that the system resource is always shared between the users equally. Each admitted session out of n *running sessions is allocated an equal portion of the total time resource, i.e.,* $\frac{r_0}{r_k} = \frac{1}{n}$*. Then, the users adjust their transmit power to match their required target bitrate as long as it does not exceed the maximum allowed power level. Clearly, in case of* n *active sessions,* $r_k = r_0 n, \forall k = 1, ..., n$*.*

In summary, both MR and RR policies offer a flexible choice between more system capacity (resulting also in higher power consumption) and better network resource utilization (enabling some transmit power savings). By considering them below, we ensure that the HetNet may offer a good balance between network- and user-side performance.

Assumption 6. ***Admission control.***

Types I and III. Since a real-time session requires the bitrate of r_0*, the system admits a newly-arrived session if there still remains sufficient resource to serve it. In other words, each ongoing session* k *has to occupy exactly* r_0/r_k*-fraction of the system time (where the overhead is accounted for later), while for all the active sessions it holds the following:*

$$\sum_{all\ sessions} \left(\frac{r_0}{r_k^{\max}}\right) \leq \delta, \tag{3.8}$$

where δ *is the resource available for sharing at a particular receiver (AP or BS) (e.g., excluding resources allocated for fading compensation),* $r_k \leq r_k^{max}$ *is the instantaneous data rate depending on the distance between the user and the receiver,* *and r_k^{max} is the highest achievable data rate at the maximum power level. Admission control of tier type III also incorporates interference control, which is detailed below (see Assumption 8).*

Type II. Since the resource is exclusive for the link of this tier, admission control focuses on ensuring that the rate r_0 is achievable for the newly-arrived link, so that the interference is controlled as well.

Figure 3.5 illustrates an example of flow-level dynamics and admission control mechanism for the tier type I. For the tier types II and III, admission control procedure has to determine whether the current interference exceeds a particular given threshold or not.

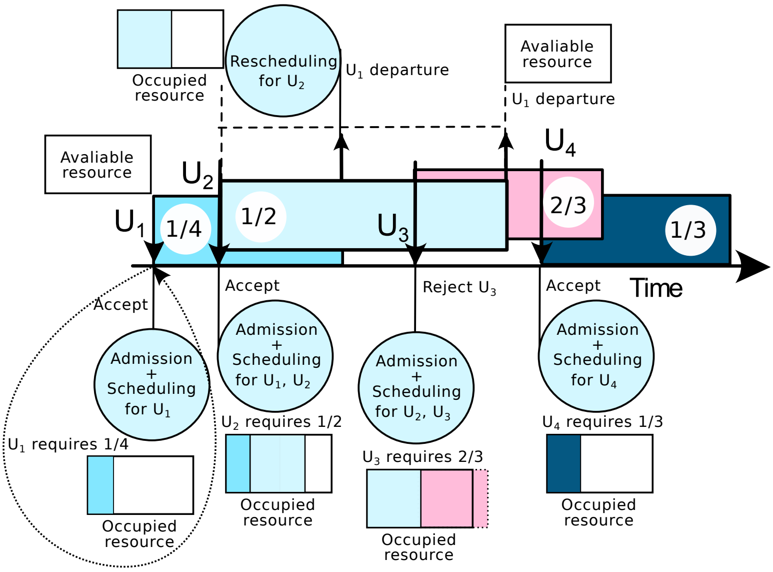


Figure 3.5 An example of flow-level dynamics.

Assumption 7. *Interference margin.* *We also assume that the noise plus interference has the form of $N_0 + I = KN_0$, where the value of K is a scaling factor fixed across the network in question. It has the meaning of interference margin per receiver.*

The latter corresponds to the well-known concept of interference-over-thermal, which has been widely used in analyzing the uplink cellular networks for the open-loop power control, see [125] or WiFi CSMA/CA mechanisms. We may thus aggregate the individual interferences created by the proximate users of a particular tier into a cumulative background noise level, which in the practice of network planning is taken into account as a particular interference margin.

Assumption 8. *Interference assessment.* *Type II. We assume that the noise plus interference power does not exceed some network-wide threshold, i.e., $N_0 + I \leq KN_0$. Further, it is assumed that the tier type I with $n-1$ active users admits a* *new session n if for the set $\{T_k\}_{k=1}^{n}$ of transmitters the following conditions hold at each receiver R_j, $j = 1, ..., n, j \neq k$:*

$$\frac{p_k\gamma_{k,k}}{KN_0} \geq e^{r_0/(w\delta_{i,k})} - 1 \quad \text{and} \quad p_k\gamma_{k,j} \leq N_0, \quad \forall k, \quad j \neq k, \tag{3.9}$$

where r_0 is the target bitrate, $\delta_{i,k}$ is the actually available resource for a given link of the tier of sequential number i (after removing the overheads and signaling), and the value of K is fixed for this tier.

These conditions imply that the required bitrate r_0 can be achieved on each link k (see the left part of (3.9)) and that the interference on R_j produced by T_k does not exceed the given threshold N_0 (see the right part of (3.9)).

Type III. Further, it is imposed that a tier with $n - 1$ active users admits a new session n if for the set $\{U_k\}_{k=1}^{n}$ of all users the following condition holds at each receiver:

$$\frac{p_k\gamma_{k,k}}{KN_0} \geq e^{r_0/w} - 1 \quad \text{and} \quad p_k\gamma_{k,j} \leq N_0, \quad \forall j, \quad j \neq k, \tag{3.10}$$

where $\gamma_{k,j}$ is the path gain between the user k and its receiver j and p_k is the corresponding allocated power.

By that, the admission control function ensures that the required minimum bitrate can be achieved by a user, and that the interference at the receiver R_k produced by the transmitter T_j does not exceed a given threshold depending on the technology-related features. We also note here that our interference and rate estimation has predictive character and assists the network in making a guided decision on whether a user should be admitted or not.

3.6 CONSIDERED HETNET OPERATION

In the remainder of this work, we begin by studying the individual performance of tiers belonging to the above main types. Further, for the entire converged HetNet, we explore a particular *sequential* mechanism of user admission and network selection as a characteristic example of future 5G operation. It is illustrated in Figure 3.6, where we assume the "cascade" service for any new session arriving into the system. Correspondingly, the network selection assistance entity attempts to offload the newly arrived session onto the *initial* network according a particular RAT priority. In case when a RAT operates over a *shared resource*, the network selection entity attempts to offload the user session to the *nearest* receiver (i.e., the closest AP/BS) by performing the corresponding admission control, which is managed centrally. We note that the nearest receiver may also be located outside of the circle R.

If the session is accepted on the current tier, it is served there without interruption until when it successfully leaves the system. Otherwise, if this session cannot be admitted onto the attempted tier, the network admission function attempts the following RAT in the order of decreasing priority. Hence, either the session is accepted onto one of the $T - 1$ tiers and served there, or the macro network T (which is always attempted the *last*) tries to serve this session. Eventually, if the session cannot be admitted onto the macro network either, it is considered *permanently blocked* and leaves the system unserved without any impact on the new arrivals.

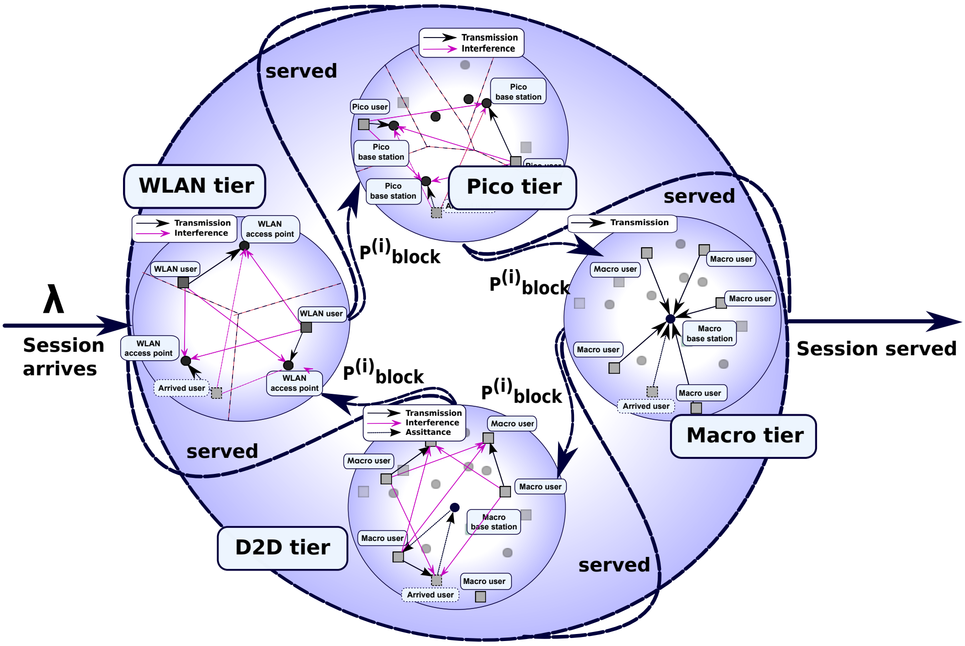


Figure 3.6 Considered operation of a multi-RAT HetNet.

Whenever admitted, a transmitting user exploits a fraction of the system time resource and sets its power either fixed or as commanded by the power control function to achieve its required data rate. The system makes a new decision on scheduling and transmit power allocation for all the active users at every new arrival or when an existing session is served and leaves the system. For each tier, we introduce the corresponding blocking probability $P^{(i)}_{block}$ and acceptance probability $P^{(i)}_{a} = 1 - P^{(i)}_{block}$, where i is the index corresponding to one of the tiers, respectively. Moreover, we remind that the session arrival rate on the initial tier is λ (see Assumption 1).

Assumption 9. ***Decoupling assumption.*** *To preserve analytical tractability of our mathematical model, we assume that all types of network tiers serve their users independently, which results in a random thinning of the arrival process with the corresponding acceptance probabilities. Therefore, due to the Poisson property of the thinned flow, the arrivals on the following tier (those not accepted by the current tier) follow a Poisson process of density* $\lambda_{i+1} = \lambda_i \left(1 - P^{(i)}_{a}\right)$*, where* $P^{(i)}_{a}$ *is the tier* i *accept probability.*

The above assumption is a natural methodological move to decompose the system into a set of tractable and well-defined components, which may be easily replaced and/or interchanged.

Abstracting away the locations of users for analytical tractability, we further assume that the arrivals on the subsequent tiers are also placed uniformly within the circle of radius R (following from the initial PPP). This latter consideration *does not* actually hold in reality. Instead, there is some pattern in which users are taken for service by the tiers. However, our simulation results (as reported below) reveal that the assumption of uniformity is surprisingly accurate. This makes the analysis of our system under the aforementioned assumptions to be an adequate approximation for the practical HetNet operation [P5].

Consequently, denoting the macro network (the "last resort", final tier) accept probability as $P_a^{(T)}$, we may easily establish the overall system *blocking* probability P_{block} as follows:

$$P_{block} = 1 - \sum_{i=1}^{T} \prod_{j=1}^{i-1} (1 - P_a^{(j)}) P_a^{(i)}. \tag{3.11}$$

Chapter 4

Mathematical Characterization of Emerging HetNets

4.1 ANALYSIS OF RANDOM DYNAMIC HETNETS

Below, we provide a summary of our rigorous analytical efforts to evaluate the important HetNet-related performance metrics. Hereinafter, we consider different tiers separately. We underline here that our system analysis is built on the *decoupling principle* as per Assumption 9. This technique is used widely and allows for evaluating even very complex systems by regarding them as an integrated set of tractable components. The following mathematical models and associated reasoning are divided into two large parts: for the tier types I and II, where the system is determined by the state of the links, and for the tier type III, where the additional information on the receivers is needed.

Here, we outline our general stochastic model for the tiers based on the assumptions introduced previously. Assume that the arrivals on all tiers follow a Poisson process with the rates $\lambda_1 = \lambda$ (to the overall system and thus the initial tier), λ_i, and λ_T (at the final tier). We observe the tier types I and II at the particular moments t of session (user) arrivals/departures. Since the arrivals follow a Poisson process and the service (session length) is distributed exponentially, our system behavior may be represented by a stochastic Markov process $S(t)$, where the future process evolution is determined by the set of the ongoing sessions that are currently served on a given tier.

4.1.1 Analyzing Tier Types I and II

For the tier types I and II, the state of the process $S(t)$ is determined by the characteristics of the ongoing sessions within the target circle. For convenience, we denote these abstract characteristics as ω and note that they depend on the distance between the transmitter and the receiver. Therefore, the system state is represented by the vector $(\omega_1, ..., \omega_n)$, where n is the number of sessions in service (see Figure 4.1 for details).

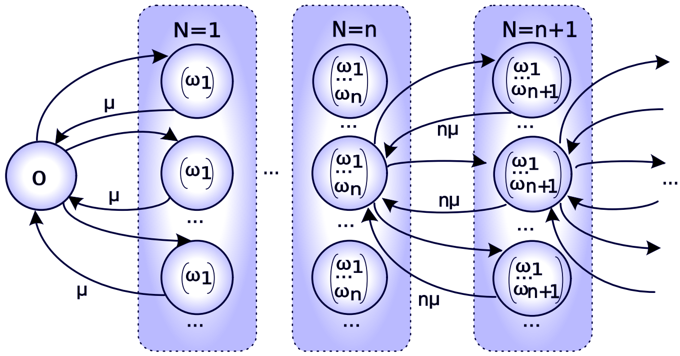


Figure 4.1 State diagram for the tier types I and II.

Let a tier have n running sessions in the state s. We denote the rejection probability at the state s for the newly-arrived session as $Q_{n+1|s}$. Then, transitions from the state $s = (\omega_1, ..., \omega_n)$ to the state $(\omega_1, ..., \omega_n, \omega_{n+1})$ and backwards have the rates of $\lambda_T \left(1 - Q_{n+1|s}\right)$ and $(n+1)\mu$, respectively.

4.1.2 Analyzing Tier Type III

The tier of type III having the sequential number i comprises several shared receivers (APs/BSs), which are all distributed on the plane with the densities of L_i. Moreover, such tier is interference-limited, and, hence, the respective stochastic processes show state-dependent properties, which are different from those discussed previously. The state of the stochastic Markov process $S(t)$ may be represented by the set of sessions with respect to the corresponding receivers. Similarly, we adopt the notation ω for the session characterization. Then, the state s of the tier type III is represented by:

$$(\omega_1, ..., \omega_{n_1}; \omega_{n_1+1}, ..., \omega_{n_1+n_2}; ...; \omega_{s_n+1}, ..., \omega_{s_n+n_{N_i}}),$$

where $s_n = \sum_{i=1}^{N_i-1} n_i$, as well as n_1, n_2, and n_{N_i} are the numbers of users associated with the first, second, and last AP/BS, respectively. The random variable N_i corresponds to the number of receivers in a certain area and follows the Poisson distribution. The state diagram of the considered system is illustrated in Figure 4.2.

We consider state s, where the tier type III is serving n ongoing sessions with a random number of N_i receivers. Similarly, we denote the rejection probability for the newly-arrived session as $Q_{n+1|s}$. Then, the transitions from the state s to the state of $n+1$ active sessions have the rate of $\lambda_i \left(1 - Q_{n+1|s}\right)$. The backward rate equals $(n+1)\mu$, since the service does not depend on the state, but rather on the number of sessions served simultaneously.

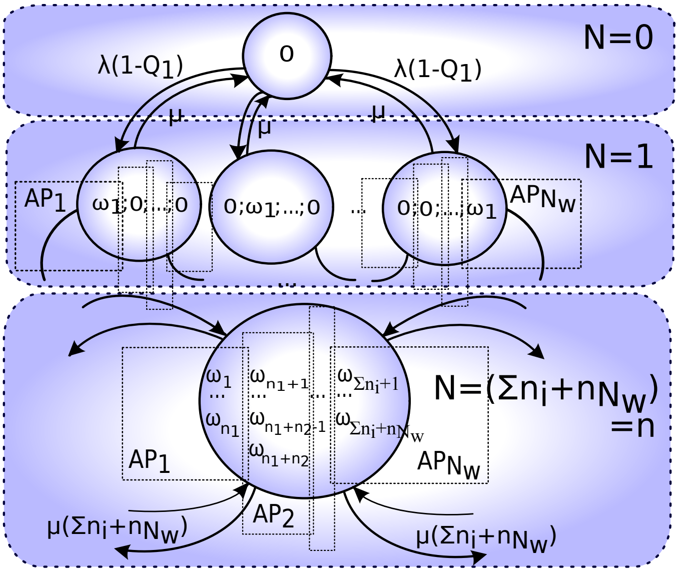


Figure 4.2 State diagram for the tier type III.

4.2 CALCULATING THE STEADY-STATE DISTRIBUTION

Due to the uncountable number of states in the considered system, it may be complicated to deliver the steady-state distribution straightforwardly. However, the corresponding Markov process may be simplified by employing the *state aggregation* technique.

Hence, for the tier types I and II, we aggregate the states $\{(\omega_1, ..., \omega_n)\}_{\omega \in \Omega}$ by n (where Ω is the space of all possible vectors $(\omega_1, ..., \omega_n)$, $n \in N$). For the more complex tier type III, we aggregate all possible states of the system (which contain n ongoing sessions) into the state n. The described aggregation process is demonstrated in Figure 4.2.

Assumption 10. ***State aggregation.***

1. For the tier types I and II, we aggregate all the states containing n sessions into the unifying state n, regardless of the actual locations of users.

2. For the more complex tier type III, we combine all possible states of the system (which contain n ongoing sessions) into the state n, regardless of the locations of the current users or their connections to a certain receiver.

3. In order to keep our system memoryless, we adopt a simplification, where the sessions at the state n, while keeping all of their other properties, do not preserve their locations from state to state. For the sake of analytical tractability, these locations are assumed to be generated anew at every particular state n.

We note, however, that the system still keeps track of the previously admitted sessions owing to the probabilities Q_{n+1} to reject the session arrived at the state n conditioning on the fact that the current n-th session satisfies the admission control criteria. As the result of the state aggregation, we arrive at the birth-death processes for all the tiers with the rates of $\lambda_i\,(1-Q_{n+1})$ and $(n+1)\mu$. We further denote the arrival rate λ_i into the system simply as λ and formulate the following proposition.

Proposition 1. *The steady-state distribution* $\{\pi_i\}_{i=0}^{\infty}$ *for the considered process* $S(t)$ *with the transitions* $\lambda\,(1-Q_{n+1})$ *and* $(n+1)\mu$ *can be closely approximated by:*

$$\pi_n = \pi_0 \frac{\lambda^n}{\mu^n} \frac{\prod_{i=1}^{n}(1-Q_n)}{n!}, \tag{4.1}$$

where

$$\pi_0 = \left(\sum_{i=0}^{\infty} \frac{\lambda^n}{\mu^n} \frac{\prod_{i=1}^{n}(1-Q_n)}{n!} \right)^{-1},$$

and Q_{n+1} *is the reject probability on the transition from the state* n *to the state* $n+1$.

Proof. Proof is straightforward and is thus omitted here. □

Based on the steady-state distribution and assuming that it exists, our approach empowers us to estimate a wide class of stationary characteristics in the considered system, such as the expected number of ongoing sessions, the probability of session's permanent blocking, and even its energy consumption. To this end, the average number of active sessions and the system blocking probability are defined as:

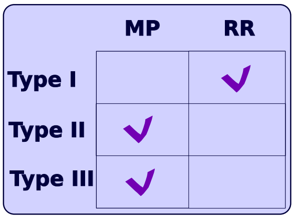
$$E[N] = \sum_{n=0}^{\infty} n\pi_n, \quad P_{block} = \sum_{n=0}^{\infty} Q_{n+1}\pi_n. \tag{4.2}$$

The average number of users may also be used as the system (area) capacity prediction for sufficiently high arrival rates.

In our analysis, we disregard the history of the system processes from the perspective of the ongoing sessions. We thus replace the initial "stateful" systems with the memoryless processes, for which we examine the arbitrary set of respective random variables at each point t. If the reject probabilities Q_{n+1} are known for all tiers, we easily obtain the steady-state distribution by using (4.1). Therefore, in what follows we concentrate on calculating the values of Q_{n+1}. In order to take into account the memory property that we have thus omitted, we will refer to the corresponding conditional probabilities further on.

The described methodology is rather general and can be applied to a wide range of RATs, provided that they are sufficiently simplified to meet our taxonomy. The distinctive features determining the selected model behavior are defined by the particular values of parameters as well as by the calculation of the probabilities Q_{n+1}, which, in turn, depend on the power allocation policy:

$$(1-Q_{n+1}) = \Pr\{\text{new session } n+1 \text{ is admitted } \mid n \text{ sessions are already active}\},$$



	MP	RR
Type I		✓
Type II	✓	
Type III	✓	

Figure 4.3 Selected systems for the tier types I, II, and III.

(4.3)

where by the above conditional probability we account for the previous system history, while the new system evolution process (n-based) is memoryless. In other words, we estimate the probability to share the resource between $n+1$ random sessions if n other stochastically different sessions have already been admitted at the previous state.

We restrict our further exploration to considering one system of each type by selecting a particular power policy (as it is noted in Figure 4.3). Alternative power allocation policies may be accounted for similarly.

4.3 CHARACTERIZING TRANSITIONS FOR IMPORTANT HETNET EXAMPLES

We continue by evaluating the probabilities Q_{n+1} and the transition rates $\lambda_i\left(1-Q_{n+1}\right)$ necessary for deriving the steady-state distribution. The rest of the text is organized in the following order. First, we illustrate our approach on the simpler tier types I and II, selecting as examples the RR (round robin) and the MP (maximum power) policies, correspondingly, which may characterize LTE macro cell and D2D over WiFi-Direct. Then, we continue with a more complex (due to the presence of interference) tier type III, represented by the system with the maximum power policy and corresponding to, e.g., WLAN operation.

4.3.1 Tier Type I Transitions

We begin with tier type I under the RR transmission policy and detail the calculations, which are necessary for characterizing this policy conditioning on (i) the absence of interference and (ii) equal sharing of the resource among all the links. Hence, the transitions from the state n to the state $n+1$ are defined by:

$$\lambda_m\left(1-Q_{n+1}\right)=\lambda_m\left(\Pr\left\{\frac{r_0}{r_i^{\max}}\le\frac{\delta_m}{n+1},\forall i=1,n+1 \,|\, \frac{r_0}{r_i^{\max}}\le\frac{\delta_m}{n},\forall i=1,n\right\}\right). \quad (4.4)$$

Further, we formulate the following Theorem.

Theorem 1. *For the tier type I under the RR policy, the accept probabilities* $\Pr\{accepted \mid arrived\} = 1 - Q_{n+1}$ *can be obtained by:*

$$1 - Q_{n+1} = \Pr\left\{r_{n+1}^{\max} \geq \frac{r_0 (n+1)}{\delta_m}\right\}\left(\frac{\Pr\left\{r_i^{\max} \geq \frac{r_0}{\delta_m}(n+1)\right\}}{\Pr\left\{r_i^{\max} \geq \frac{r_0 n}{\delta_m}\right\}}\right)^n, \tag{4.5}$$

where

$$\Pr\{r \geq x\} = 1 - \Pr\{r < x\} = 1 - F_r(x), \tag{4.6}$$

and

$$F_r(x) = 1 - \frac{1}{R^2}\left[\frac{Gp_{\max}}{N_0}\right]^{2/\kappa}\left(e^{x/w} - 1\right)^{-2/\kappa}, r_R \leq x < r_{\lim}; \quad F_r(x) = 1, x \geq r_{\lim}.$$

Proof. Proof is based on calculating distribution of the functional transform, similar discussion may be found in [126] □

The latter completes the expression (4.4) and delivers the steady-state distribution (4.1), as well as other relevant stationary metrics (4.2).

4.3.2 Tier Type II Transitions

We continue by considering the tier type II, for which the resource access is exclusive, the admission control regulates interference, while the transmit power is set to its maximum. First, let n sessions already exist in the system. Hence, for all $i = 1, ..., n$ we require the following target data rate condition to hold:

$$r \leq w \log\left(1 + \frac{p_{\max}\gamma_{i,i}}{KN_0}\right) \Leftrightarrow p_{\max}\gamma_{i,i} \geq KN_0\left(e^{\frac{r}{w}} - 1\right). \tag{4.7}$$

Therefore, the following Theorem can be formulated.

Theorem 2. *For the tier type II under the MP policy, if admission control is performed according to (3.9) and, in particular, accounting for (4.7), then the reject probabilities* Q_{n+1} *can be closely approximated by:*

$$Q_{n+1} = 1 - \Pr\{accepted \mid arrived\} = \left[F_\gamma\left(\frac{N_0}{p_{\max}}\right)\right]^{2n-1}\left[1 - F_\gamma\left(\frac{\theta_0}{p_{\max}}\right)\right], \tag{4.8}$$

where $\theta_0 = KN_0\left(e^{\frac{r}{w}} - 1\right)$ *and the cumulative distribution function (CDF) for the SNR per a power unit* γ *is given as:*

$$F_\gamma(\gamma) = 1 + \frac{G^{\frac{4}{k}}\gamma^{-\frac{4}{k}}}{8R^4} - \frac{G^{\frac{2}{k}}\gamma^{-\frac{2}{k}}}{R^2}\ln 2, \ \textit{if}\ \frac{G}{(2R^2)^{\frac{k}{2}}} \leq \gamma \leq \gamma_{\max}, \gamma_{\max} = \frac{KN_0}{p_{\max}}\left(e^{\frac{r_{\max}}{w}} - 1\right)$$

$$F_\gamma(\gamma) = 1 - \frac{1}{R^2}\left(\frac{G^{\frac{4}{k}}\gamma^{-\frac{4}{k}}}{8R^2} + G^{\frac{2}{k}}\gamma^{-\frac{2}{k}}\ln\frac{4R^2\gamma^{\frac{2}{k}}}{G^{2k}}\right), \ \textit{if}\ \frac{G}{(2R)^k} \leq \gamma \leq \frac{G}{(2R^2)^{\frac{k}{2}}}.$$

Proof. Proof may be found in [126] □

4.3.3 Tier Type III Transitions

We proceed with characterizing the tier type III and detailing the calculations, which are necessary for capturing the MP transmission policy. The transitions from the state n to the state $n+1$ are thus defined by:

$$\lambda_w(1-Q_{n+1})=\lambda_w \Pr\left\{A_j^{(n+1)}, j=1,...,n+1 | A_j^{(n)}, j=1,...,n\right\}, \tag{4.9}$$

where event $A_j^{(n)}$ is given as:

$$A_j^{(n)} = \left\{ \frac{r_0}{r_j^{\max}} \leq \delta_w - \sigma_n \text{ and } \gamma_{j,k} p_{\max} \leq N_0, \forall k \neq j \right\},$$

where δ_w is a share of the available resource at the receiver (without the signaling overhead and collisions) and σ_n is a part of the resource given to other sessions at the same receiver in the current state. We further denote $r_0/(\delta_w - \sigma_n)$ as $\tilde{r}_{0,n}$.

The calculation of σ_n is based on the following assumption and the subsequent Theorem.

Assumption 11. ***AP link abstraction.*** *Here, to abstract away the session-receiver details at the state n, we assume that upon its arrival into the system, a session observes the average (typical) number of users at the nearest receiver (see Theorem 3). This average number depends on the number of ongoing sessions, i.e., on the state index n, as well as on the parameter $\tilde{r}_{0,n}$.*

Theorem 3. *For the tier type III, the average number of sessions per receiver (AP/BS) n_0 tends to $\frac{n}{L_i(\pi R^2)} = \frac{n}{E[N_i]}$ for large areas, where $E[N_i]$ is the expected number of receivers of tier i within the circle R.*

Proof. Proof may be found in [P5] □

Note that Theorem 3 above is similar in its meaning to the research findings obtained previously in [127]. Then, basing on these results, we may reformulate the following as stated in Assumption 11. A newly-arrived session observes the system, where on average every receiver already serves $n_0 = \frac{n}{L_i(\pi R^2)}$ sessions.

Theorem 4. *For the tier type III under the MP policy, the corresponding transition rates may be calculated as* $\Pr\{$*accepted* $|$ *arrived*$\}$ $= 1 - Q_{n+1}$ *accounting for the following:*

$$1-Q_{n+1}=\frac{\left(1-e^{-\pi L_w d_{r,n+1}^2}\right)^{n+1}}{\left(1-e^{-\pi L_w d_{r,n}^2}\right)^{n}}\left(L_w \pi d_{thr}^2 e^{-L_w \pi d_{thr}^2}+e^{-L_w \pi d_{thr}^2}\right), \tag{4.10}$$

where $d_{thr}=\left[\frac{G p_{\max}}{N_0}\right]^{\frac{1}{\kappa}}$ *and the constant value* $d_{r,n}$ *is defined as* $\left(\frac{p_{\max} G}{K N_0}\right)^{\frac{1}{\kappa}} \left(e^{\frac{\tilde{r}_{0,n}}{w}}-1\right)^{-\frac{1}{\kappa}}$.

Proof. Proof may be found in [P5] □

In summary, by introducing $d_{r,n}$ we emphasize that it depends on $\tilde{r}_{0,n} = r_0/(\delta_w - \sigma_n)$. This, in turn, is a function of the number of sessions on the tier type III via the occupied resource σ_n representing the average share of the resource exploited at the state n and is given by:

$$\sigma_n = E\left[\frac{r_0}{r_i^{\max}} \middle| \frac{r_0}{r_i^{\max}} \leq \delta_w\right]\frac{n}{E[N_i]} = E[y|y \leq \delta_i]\frac{n}{E[N_i]}, \tag{4.11}$$

and $E[y|y \leq \delta_i]$ may be found as:

$$\begin{aligned} E[y|y \leq \delta_i] = \int\limits_{y_0}^{\delta_i} y f_y(y|y \leq \delta_i)dy = \tfrac{2\pi L_w}{C_3}\int\limits_{y_0}^{\delta_i} y d(y) d'(y) e^{-\pi L_w \left(\frac{p_{\max} G}{K N_0}\right)^{\frac{2}{\kappa}} \left(e^{\frac{r_0}{w y}} - 1\right)^{-\frac{2}{\kappa}}} dy + \\ + \tfrac{2\pi L_i}{C_3} y_0 \int\limits_{0}^{y_0} d(y) d'(y) e^{-\pi L_w \left(\frac{p_{\max} G}{K N_0}\right)^{\frac{2}{\kappa}} \left(e^{\frac{r_0}{w y}} - 1\right)^{-\frac{2}{\kappa}}} dy, \end{aligned} \tag{4.12}$$

where $C_3 = \Pr\{y \leq \delta_i\} = F_y(\delta_i)$ and y_0 is assumed to be less than δ_i.

The expression (4.10) finally enables us to derive the key performance metrics of interest, such as the expected number of ongoing sessions and the overall blocking probability (4.2).

4.4 DISCUSSION ON POSSIBLE EXTENSIONS

The above system has been constructed entirely for the example of ”session-based” traffic, but the case of ”file-based” transmission deserves separate attention. Let us consider the so-called elastic traffic transmission, where upon arrival users generate files of exponential size with the average of θ, but the actual bitrate is not constrained and depends on the location and the current number of users that equally share the available resource. We note that the exponential size is considered here for the sake of preserving the memoryless property. However, other alternatives are possible as well (e.g., r-Erlang or hyperexponential process), but they would complicate the core Markov process diagram.

The approach described above should thus be modified by taking into account the varying service rate. Therefore, if the admission control keeps the same forward transitions and the backward transitions are modified by the average service rate, the stationary distribution $\pi = \{\pi_n\}_{n=0}^{\infty}$ of the aggregated process $S(t)$ could be obtained as:

$$\pi_n = \pi_0 \prod_{j=1}^{n} \frac{\lambda_i(1 - Q_{j-1})}{b_j}, \tag{4.13}$$

where λ_i is the arrival rate to the considered tier i, $1 - Q_{j-1}$ is the transition probability from the state $j-1$ to the state j upon the user arrival, b_j is the transition rate from the state j to the state $j-1$, and π_0 may be calculated from the normalization condition.

This statement enables the calculation of all the system metrics of interest, averaged across time and space (such as the average number of users, their transmission times, etc.). However, it would also require an additional derivation of the transition rates b_j, which are determined by the distribution of the actual service rate, i.e., the file size and the instantaneous data rate given by the Shannon's formula. For example, if j active users have the same instantaneous data rate of r, as well as share equally the available resource and hence receive the actual data rate of $\frac{r}{j}$, then the transition $(j) \to (j-1)$ is $b_j = j\frac{1}{\theta j/r} = \frac{r}{\theta}$. In case of varying instantaneous rates, $r_i \neq const$, the transitions of the aggregated process are defined by:

$$b_j = j\frac{1}{E[\text{service time}]} = j\frac{1}{E\left[\frac{s}{r/j}\right]} = \frac{1}{\theta E\left[\frac{1}{r}\right]}, \tag{4.14}$$

where s is the random file size, θ is the average file size, and r is the instantaneous data rate, which may be calculated through the functional transform $r(d)$ and the probability density function (PDF) $f_d(d)$ of the distances between the transmitter and the receiver, as discussed above.

The corresponding calculations for the example two-tier HetNet (WiFi + LTE RATs) as well as for the single-tier co-located WiFi/LTE system may be found in [P8]. These expressions comprehensively describe the steady-state distribution and, hence, define the average number of active users, their average time spent in service, and the average effective data rate per a served user:

$$E[N] = \sum_{i=0}^{\infty} n\pi_n, \quad E[T] = \frac{E[N]}{\lambda(1-P_{bl})}, \quad E[r] = \overline{r}\frac{\sum\limits_{n=0}^{\infty}\frac{1}{n}\pi_n}{1-\pi_0}, \tag{4.15}$$

where P_{bl} is the probability of not being accepted to the tier and $\overline{r}$ is the spatially-averaged instantaneous rate:

$$\overline{r} = \int\limits_{r_R}^{r_{\lim}} r f_d(d(r))|d_r'(r)|dr + r_{\lim}F_d(d_{\lim}). \tag{4.16}$$

We note that here we do not assume any particular distribution $f_d(d)$ – it may be taken as needed by the specified scenario.

Chapter 5

Quantifying Performance with System-Level Evaluations

5.1 FEATURES OF OUR 5G SYSTEM-LEVEL SIMULATOR

To complement our analytical study, we exploit an advanced system-level simulator (SLS) based on the up-to-date 3GPP LTE evaluation methodology (3GPP LTE Release-12 FDD) and current IEEE 802.11 specifications (IEEE 802.11-2012 supporting WiFi-Direct features). Presently, neither free nor commercially-available simulation platforms are readily applicable for evaluating 5G-grade multi-RAT systems, as they are missing the necessary features, as well as lacking scalability to adequately capture the dependencies between the studied variables. By contrast, our SLS is a flexible tool designed to support diverse deployment strategies, traffic models, channel characteristics, and wireless protocols.

To this end, we construct a multi-RAT simulation model representative of an urban deployment, where WiFi and D2D small ”cells” are overlaid on top of the multi-tier 3GPP LTE network. Outdoor deployments are considered and are based on the recommendations in [128] combining that with varying pico BS and WLAN AP densities (as per [129]). Hence, our scenario represents a harmonized 3GPP vision of a characteristic HetNet deployment. A part of it concentrates on an *area of interest*, in which co-located cellular and D2D/WLAN networks cover a limited region with many users requiring service (e.g., shopping mall, business center, etc.). For the D2D/WLAN systems, the simulation is largely based on IEEE 802.11 medium access control procedure with carrier sensing. We also assume that all APs and their respective users run the same version of the technology as WiFi-Direct clients, namely, IEEE 802.11-2012. For calibration purposes, we employ reliable results from publications on ad-hoc WLAN deployments.

For the LTE system, the simulation captures the following practical features (as opposed to the above analytical methodology): data frame structure, bandwidth requests, and scheduling by the BS. Here, our example scenario comprises 19 hexagonal cells supporting 3GPP LTE Release-10 technology, and a wrap-around technique is used to improve precision of the simulation at the edges of the deployment area. The system works over two 10-MHz bands for FDD operation (for both UL and DL), shared by all cells with 3 sectors in each, resulting in a 1x3x1 reuse pattern. For more details on the configuration of the reference LTE network, the interested reader is directed to the relevant standardization documents (e.g., 3GPP TR 36.814-900 and ITU-R M.2135-1). For performance verification purposes, we also implemented a calibration scenario from 3GPP TR 36.814-900, Table A-2.1, and ran the corresponding tests. Our simulation results fall well within the required limits for both cell-center and cell-edge spectral efficiency targets.

5.2 DISCUSSING REPRESENTATIVE NUMERICAL RESULTS

5.2.1 Three-Tier and Two-Tier HetNet Study

We begin our evaluations with Figure 5.1, where we detail the blocking probabilities (or the proportion of service requests that cannot be served by the network) for the integrated HetNet as well as for the three tiers individually: macrocell, picocell, and WLAN tier. Our observation is that with two additional overlay tiers, the HetNet performance improves significantly over what can be achieved in the macro-only networks (cellular baseline). Even though this conclusion is generally not unexpected, we note that our results here are fundamentally different from most past work, as they consider user dynamics based on stochastic traffic loads. We continue with a more detailed analysis.

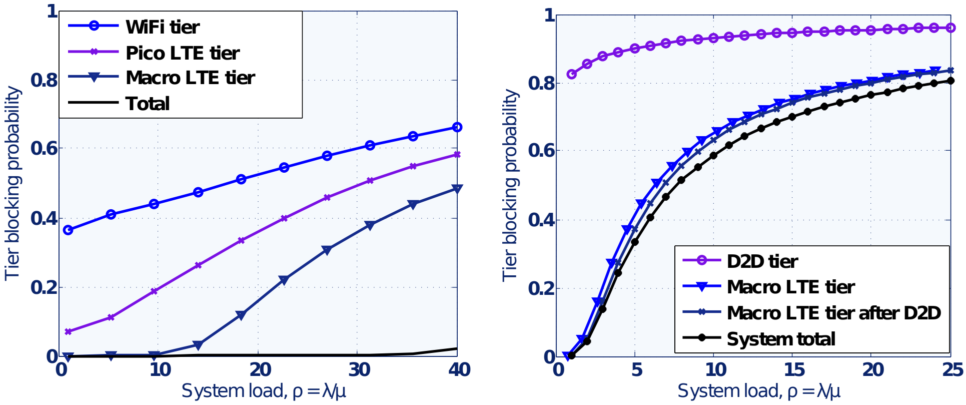


Figure 5.1 Blocking probabilities in three-tier (left) and two-tier (right) HetNets.

By employing our advanced SLS tool, we are able to demonstrate in detail how the components of the blocking probability P_{block} evolve with increasing load on the network (see Figure 5.2, left). The session is blocked if it cannot fit into the schedule at the time of arrival, and for the D2D network we differentiate between session rejections due to (i) prohibitive interference from the existing transmissions and (ii) excessive link length to support the required bitrate (given that the interference constraint has been satisfied). It is important to analyze the structure of the blocking processes for both systems. For the D2D system, at low loads the blocking is primarily caused by excessive link length, whereas as the load increases the probability of blocking due to high interference becomes dominant.

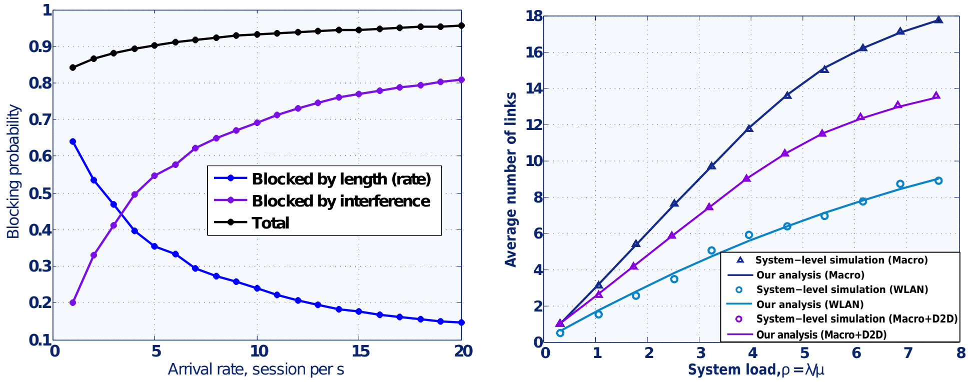


Figure 5.2 Session blocking/reject probabilities, simulations (left); and system capacity, simulation vs. analysis (right).

In addition, Figure 5.2, right contrasts the simulation results against our above analysis for LTE, WiFi, and integrated LTE+D2D network to confirm good convergence between the analytical approximations and the SLS-based results. Here, continuous lines indicate analytical data, whereas symbols correspond to the simulated values. Clearly, the overall trend is the increase in the expected number of running links, up to the saturation point, which depends on the deployment, scheduling, and multiplexing methods used.

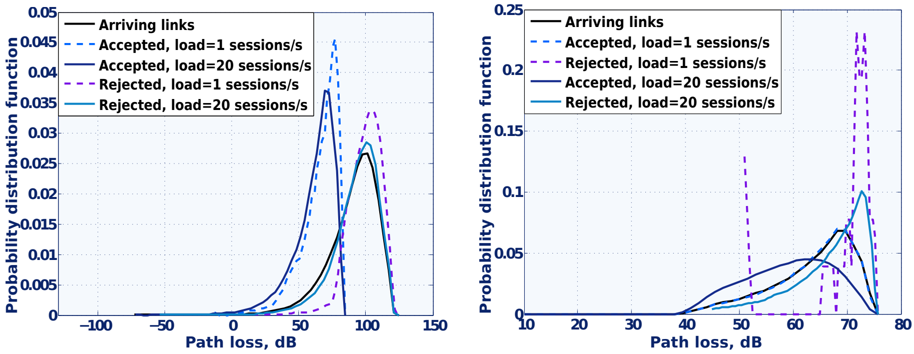


Figure 5.3 Link quality for D2D (left) and LTE (right) tiers.

In order to understand more subtle effects associated with the HetNet operation, let us examine the quality of the links in our system of interest (see Figure 5.3). When the cellular network is empty, it can afford accepting all links, no matter the quality. Under such conditions, the link quality for arrivals and accepted links is similar, and there are almost no discards (see Figure 5.3, left). As the cellular system becomes loaded, however, we see that it takes only shorter links in – as those have significantly better chances to fit into the schedule (refer to Figure 5.3, right).

5.2.2 Two-Tier HetNet Study of Densification Limits

We further illustrate the operation of an ultra-dense two-tier LTE+WiFi network, where the network selection is assumed to be managed according to the WiFi-preferred logic. The parameters of the considered scenario are similar to our above SLS study and are deemed typical for future ultra-dense small cell deployments in light of ongoing 3GPP discussions. In particular, when characterizing the system geometry, we introduce a new measure (termed network's *specific density*) and define it as the number of coverage areas that are encountered within a coverage area of an "average" cell. For example, in the conventional cellular network, the $D_s = 7$ is a typical specific density, which results in the hexagonal grid of cells, and each cell thus has exactly 6 neighbors. Consequently, we name a particular deployment ultra-dense, if $D_s > 7$, as this results in a system that is no longer conventional cellular, and cannot be represented by a regular grid on a plane.

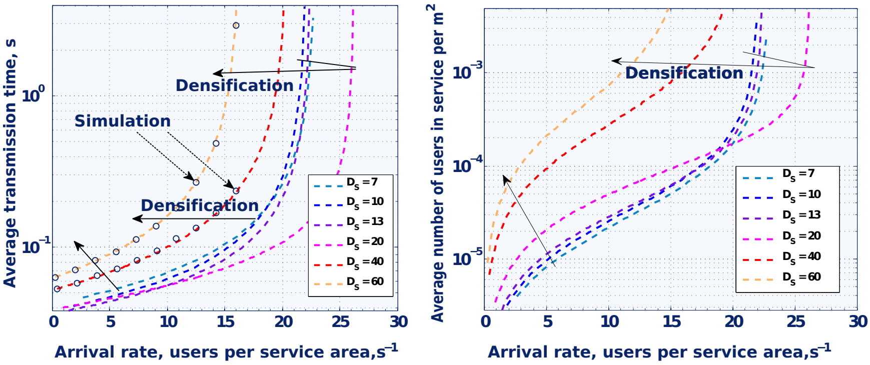


Figure 5.4 Average transmission time (left) and number of users per m^2 (right).

Let us first investigate how our considered ultra-dense HetNet system reacts to various user loads. In Figure 5.4, left, circle markers correspond to our simulation results, which selectively verify the obtained analytical dependencies. As Figure 5.4 generally suggests, the network has a very notable response to overloads. Essentially, the moment we reach the overload intensities, the transmission times grow exponentially, along with the number of backlogged users. Interestingly enough, the point where the system hits overload is sometimes inversely proportional to the deployment density. In essence, providing more access points than necessary may have a negative effect on network capacity.

To illustrate this important effect better, let us study what happens when each user has a small cell of its own. With a fixed coverage area of a small BS, the majority of the resources will become allocated to protect from excessive interference, thus decreasing the amount of resource actually available to a particular "tagged" user. On the other hand, this user's SINR will be exceptionally high. The practical limitation, however, is that the UE can only make use of around 25 dB SINR;

anything above that is essentially useless due to the limitation in the modulation and coding schemes. Hence, unless the power of a small BS is reduced appropriately, over-densification may have a visible negative impact on the system capacity, which calls for further research in this area.

In Figure 5.5, we reiterate the discussed effect, but under a different angle. One can clearly see that for higher user densities, an advanced transmission scheme (utilization of both LTE and WiFi networks simultaneously) with the specific density of 7 enjoys the best performance (which effectively corresponds to "almost" regular lattice layout). On the other hand, when system remains essentially idle, it is still beneficial to have more small cells. As this makes the UE-BS links shorter, the effect is the greater attainable data rates at lower loads.

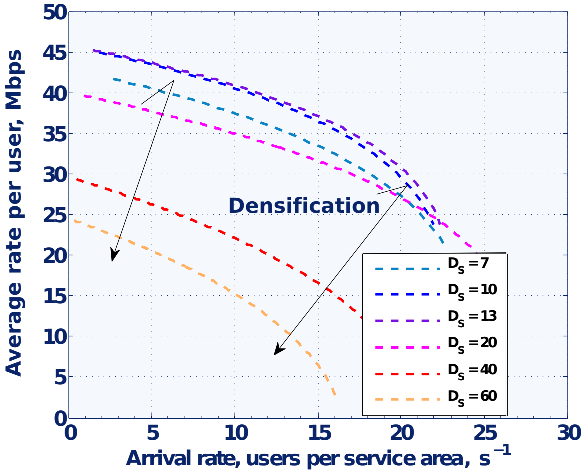


Figure 5.5 Average data rate per user.

Contrarily to how the more advanced simultaneous transmission scheme operates, the baseline WiFi-preferred system typically benefits from densification much less: at some point all of the UEs are forced to use WiFi by their RAT selection policy. This indicates, in turn, that whenever a choice of multiple alternative RATs is available, the UE should not be restricted to using either one of those, irrespective of its position relative to the small BS.

Chapter 6

Conclusions and Future Directions

In this thesis, we provided a *unified mathematical methodology* allowing for capturing traffic dynamics together with the geometrical randomness of realistic user deployments for various complex scenarios, as well as delivered a first-order evaluation of important HetNet-related metrics. To this end, we proposed a comprehensive classification of practical HetNet scenarios embracing the envisioned 5G network types in a systematic way. Abstracting away less impactful system properties, we thus focused on the generic HetNet examples of access technology groups within the proposed classification. Further, we thoroughly described the analytical methods required to calculate important performance-related parameters. By doing that, we rigorously covered a broad range of attractive HetNet configurations by providing relatively simple and accurate approximations for the stationary metrics of interest, and selectively verified these with advanced system-level simulations.

More generally, studying the ultimate capacity of 5G multi-radio HetNets remains an open problem in the field of information theory, and our methodology has the potential to shed light on it given that it can explicitly capture new interference situations and hence the achievable data rates. This challenging objective may require novel advanced analytical tools to interconnect and apply techniques and methods coming from the area of point processes, probability theory, queuing theory, and percolation theory, as well as modern engineering insights. Correspondingly, the framework developed in this thesis is not restricted to the network types considered for the discussion of results, but has flexibility to be extended for other systems of interest that may emerge in the future. Possible further applications of the proposed methodology embrace, e.g., novel UL/DL decoupling schemes [130], [131] and emerging "HetHetNet" concepts (focusing on the spatially-heterogeneous traffic in HetNets) [132]. Another interesting possible avenue is the consideration of mmWave tier within the next-generation HetNets. However, due to the complex nature of the mmWave signal propagation and much ongoing research on the appropriate channel models, respective modeling work might require careful verification through extensive practical measurements.

Chapter 7

Summary of Publications

7.1 DESCRIPTION OF PUBLICATIONS

The second part of this thesis includes *eight* publications referred to as [P1]-[P8]. None of these publications have been used as part of any other thesis. Works [P5], [P6], [P7], and [P8] are articles published in scientific journals and the rest are conference papers. The major contribution of each of the *main* publications is clarified below.

- **[P1]** O. Galinina, S. Andreev, and Y. Koucheryavy, ”Performance Analysis of Client Relay Cloud in Wireless Cellular Networks”, in *Proc. of the 10th International Conference on Wired/Wireless Internet Communications (WWIC)*, 2012.

  Description

  In [P1], we build an originator-centric model and study the performance of a relay cloud with respect to the main performance metrics: throughput, packet delay, and energy efficiency. We obtain closed-form analytical expressions for the sought metrics and verify our results via extensive simulations. In particular, we considered a wireless cellular network that enables the distributed control over cooperative communication via a client relay cloud to enhance system performance through the support of cell-edge mobile clients with poor communication links. The main performance metrics were studied, including throughput, mean packet delay, and energy efficiency. Accurate closed-form analytical expressions have been derived and verified by extensive system-level simulations. The results indicate significant promise of the relay cloud, which is able to recover the performance of the mobile clients with degraded wireless links.

  This paper is a collaborative work of the author and her supervisor with Dr. Sergey Andreev from the same research group at Tampere University of Technology (Finland).

- **[P2]** O. Galinina, A. Trushanin, V. Shumilov, R. Maslennikov, Z. Saffer, S. Andreev, and Y. Koucheryavy, "Energy-Efficient Operation of a Mobile User in a Multi-Tier Cellular Network", in *Proc. of the 20th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA)*, 2013.

  Description

  In [P2], we have considered the problem of energy efficient power control when the mobile user may communicate on several uplink wireless channels at the same time. We propose a new power control scheme suitable for a multi-tier wireless network, which maximizes the energy-efficiency of a mobile device transmitting on several communication channels while at the same time ensures the required minimum quality of service. As the result, a good compromise between improving the data rate and extending the battery lifetime is provided. In order to enable energy-efficiency maximization, we formulate an optimization problem basing on the Shannon's capacity formula. The optimal transmit power is thus obtained from the direct solution of this optimization problem under several practical constraints, such as minimum bitrate and maximum transmit power. In the second part of the paper, we apply extensive simulations to calibrate the key parameters of our optimization framework. We have also calibrated our analytical solution with the detailed link-level LTE-A simulations. The numerical results suggest the benefit of the proposed analytical solution by comparing it against intuitive (heuristic) power control strategies.

  This paper is a collaborative work of the author and her supervisor with Alexey Trushanin, Vyacheslav Shumilov, Dr. Roman Maslennikov from Lobachevsky State University of Nizhny Novgorod (Russia) and Dr. Sergey Andreev from the same research group at Tampere University of Technology (Finland), as well as with Dr. Zsolt Saffer from (formerly) Budapest University of Technology and Economics (Hungary).

- **[P3]** O. Galinina, A. Anisimov, S. Andreev, and Y. Koucheryavy, "Performance Analysis of Uplink Coordinated Multi-Point Reception in Heterogeneous LTE Deployment", in *Proc. of the 11th International Conference on Wired/Wireless Internet Communications (WWIC)*, 2013.

  Description

  In [P3], we consider a heterogeneous 3GPP LTE deployment where neighboring low power (pico) nodes may assist the macro-associated user (UE) by independently receiving its UL data packets and forwarding the successful outcomes to the serving base station. However, at the cell edges, a macro-associated user may still suffer from poor performance due to low uplink channel quality. This is when the neighboring low power nodes can help by independently trying to receive data packets from the macro user and share the result with the base station if successful. Such CoMP scheme is known as selection combining and is believed to considerably improve user cell-edge performance. With our evaluation methodology, we combine analysis and simulations to account for the UE mobility, power control, and dynamic traffic load and depending on the user proximity to the serving base station. We confirm that the expected energy efficiency and packet delay gains remain significant and consistent even for the low number of available LPNs.

This paper is a collaborative work of the author and her supervisor with Dr. Alexey Anismov from (formerly) Nokia Siemens Networks, MBB LTE (Russia) and Dr. Sergey Andreev from the same research group at Tampere University of Technology (Finland).

- **[P4]** O. Galinina, A. Turlikov, S. Andreev, and Y. Koucheryavy, "Stabilizing Multi-Channel Slotted Aloha for Machine-Type Communications," in *Proc. of the IEEE International Symposium on Information Theory (ISIT)*, 2013.

  Description

  In [P4], we consider a wireless cellular system with an unbounded population of contending machine-type users. The system provides a number of non-interfering slotted-time channels which users contend for when sending their uplink data packets subject to a common channel access probability advertised by the base station. Whereas we demonstrate that the optimal control of such probability is not feasible, we also detail a practical adaptive procedure that provably maintains a finite number of unserviced users in the system. With the increasing number of channels, the proposed procedure quickly converges to the optimal solution. We, therefore, conclude that our stabilized multi-channel slotted Aloha algorithm is naturally suitable for future machine-type systems with large user population and our solution demonstrates near-optimum performance.

  This paper is a collaborative work of the author and her supervisor with Prof. Andrey Turlikov from St. Petersburg State University of Aerospace Instrumentation (Russia) and Dr. Sergey Andreev from the same research group at Tampere University of Technology (Finland).

- **[P5]** O. Galinina, S. Andreev, M. Gerasimenko, Y. Koucheryavy, N. Himayat, S. Yeh, and S. Talwar, "Capturing Spatial Randomness of Heterogeneous Cellular/WLAN Deployments With Dynamic Traffic", *IEEE Journal on Selected Areas in Communications* vol. 32, pp. 1083-1099, 2014.

  Description

  In [P5], we provide an emerging vision of heterogeneous networks, which exploits the potential of a diverse range of devices requiring connectivity at different scales to augment available system capacity and improves the user connectivity experience. We proposed our novel integrated methodology for assisted (managed) radio network selection capturing spatial randomness of converged cellular/WLAN deployments together with dynamic uplink traffic from their users. To this end, we employ tools coming from stochastic geometry to characterize performance of macro and pico cellular networks, as well as WLAN, mindful of user experience and targeting intelligent network selection/assignment. We complement our analysis with system-level simulations providing deeper insights into the behavior of future heterogeneous deployments.

This paper is a collaborative work of the author and her supervisor with Dr. Sergey Andreev and Mikhail Gerasimenko from the same research group at Tampere University of Technology (Finland), as well as Dr. Nageen Himayat, Dr. Shu-ping Yeh, and Dr. Shilpa Talwar from Wireless Communications Laboratory, Intel Corporation (USA).

- **[P6]** O. Galinina, S. Andreev, A. Turlikov, and Y. Koucheryavy, "Optimizing Energy Efficiency of a Multi-Radio Mobile Device in Heterogeneous Beyond-4G Networks", *Performance Evaluation*, vol. 78, pp. 18-41, 2014.

  Description

  In [P6], we address energy efficient power control for a wireless deployment with multiple available radio access technologies. The problem of strict energy efficiency maximization at a mobile user device has been solved analytically for an arbitrary number of RATs and under several practical restrictions, such as minimum target bit-rate and maximum allowed transmit power. Our illustrative numerical examples for two and three RATs confirm that the proposed power control scheme reduces mobile device's power expenditure, while at the same time maintaining the required level of user data rate. By contrast to the previous work, the use of our approach establishes support regions where two or more RATs work collaboratively to result in more energy efficient device operation when compared against simpler power control techniques. Our results suggest that the proposed power control strategy might become an attractive choice for the future integrated beyond-4G wireless systems and thus contribute to the related research.

  This paper is a collaborative work of the author and her supervisor with Dr. Sergey Andreev from the same research group at Tampere University of Technology (Finland) and Prof. Andrey Turlikov from St. Petersburg State University of Aerospace Instrumentation (Russia).

- **[P7]** S. Andreev, M. Gerasimenko, O. Galinina, Y. Koucheryavy, N. Himayat, S. Yeh, and S. Talwar, "Intelligent Access Network Selection in Converged Multi-Radio Heterogeneous Networks", *IEEE Wireless Communications*, vol. 21, pp. 86-96, 2014.

  Description

  In [P7], we consider heterogeneous multi-radio network deployments, where each user device may employ multiple radio access technologies to communicate with network infrastructure. We review major challenges in delivering uniform connectivity and service experience to converged multi-radio heterogeneous deployments. We envision that multiple radios and associated device/infrastructure intelligence for their efficient use will become a fundamental characteristic of future 5G technologies, where the distributed unlicensed-band network (e.g., WiFi) may take advantage of the centralized control function residing in the cellular network (e.g., 3GPP LTE). Illustrating several available architectural choices for integrating WiFi and LTE networks, we specifically focus on interworking within the radio access network and detail feasible options for intelligent access network selection. Both network- and user-centric approaches are considered, wherein the control rests with the network or the user. In particular, our system-level simulation results indicate that load-aware user-centric schemes, which augment SNR measurements with additional information about network loading, could improve the performance of conventional WiFi-preferred solutions based on minimum SNR threshold. Comparison with more advanced network-controlled schemes has also been completed to confirm attractive practical benefits of distributed user-centric algorithms. Building on extensive system-wide simulation data, we also propose novel analytical space-time methodology for assisted network selection capturing user traffic dynamics together with spatial randomness of multi-radio heterogeneous networks.

This paper is a collaborative work of the author and her supervisor with Dr. Sergey Andreev and Mikhail Gerasimenko from the same research group at Tampere University of Technology (Finland), as well as Dr. Nageen Himayat, Dr. Shu-ping Yeh, and Dr. Shilpa Talwar from Wireless Communications Laboratory, Intel Corporation (USA).

- **[P8]** O. Galinina, A. Pyattaev, S. Andreev, M. Dohler, and Y. Koucheryavy, "5G Multi-RAT LTE-WiFi Ultra-Dense Small Cells: Performance Dynamics, Architecture, and Trends", *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 1224-1240, 2015.

  In [P8], we address the ongoing densification of small cells integrating both cellular and WiFi technology families, since users in future 5G systems will most likely be able to use 3GPP, IEEE, and other technologies simultaneously to maximize their quality of experience. We perform a novel performance analysis specifically taking the system-level dynamics into account and thus giving a true account on the uplink performance gains of an integrated multi radio access technology (RAT) solution versus legacy approaches. In terms of the mathematical framework, we were able to capture the spatial randomness of the users' distribution jointly with their uplink data dynamics. For LTE, we proposed a novel concept of phantom users, which makes interference coordination and scheduling analytically tractable. Consequently, this allowed us to obtain the stationary distribution and transition rate of the aggregated service process, as well as the resulting resource slicings in WiFi, LTE, and joint deployments. We also analyzed the important practical case of mixed PPP-cluster user distributions. The obtained equations are mostly in closed form, and thus easy to apply. Also, the mathematical model has then been verified by means of a 3GPP-compliant simulator, where we contrasted the baseline scenario to the truly integrated HetNet approach with flow splitting. The baseline refers to the case where offload to WiFi is always preferred to LTE service. The performance parameters considered were the average number of users per unit area, the average transmission time of the uplink file transmission, and the rate per user under loaded system conditions.

This paper is a collaborative work of the author and her supervisor with Dr. Sergey Andreev and Alexander Pyattaev from the same research group at Tampere University of Technology (Finland), as well as Prof. Mischa Dohler from Kings College London (UK).

7.2 AUTHOR'S CONTRIBUTION

The research work summarized in this thesis has been carried out in the Department of Electronics and Communications Engineering, Tampere University of Technology, Finland. The author of this thesis is the main contributor to [P1]-[P6], [P8]. The reported research has been done by the author, supervised and guided by her supervisor Prof. Yevgeni Koucheryavy and by her instructor Dr. Sergey Andreev, as well as deeply supported by her colleagues from the same research group in Tampere. Numerous discussions with the supervisor, instructor, and co-authors helped the author shape the ideas presented in this thesis, as well as improve the quality and the style of her writing. Further, many particular features published in [P1]-[P8] have been developed in tight collaboration between the author, the research team, and the international colleagues. Below we detail the author's contribution to each one of the referred main publications.

In [P1], the author has been responsible for the system-level simulations as well as for developing the analytical part. In [P2], the author has formulated the general problem, introduced the system model, and derived all the resulting analytical findings. In [P3], the author has contributed the system model, derived the underlying mathematical expressions for system performance metrics, and developed system-level simulation to verify the results. In [P4], the author has formulated the research hypothesis and provided the necessary mathematical proofs as well as built the numerical solution. In [P5], the author has developed the novel analytical framework, as well as implemented the system level simulation. In [P6], the author has extended the system model from [P2] and provided an algorithm-based analytical solution for the extended model. In [P7], the author has formulated the analytical space-time methodology and has contributed in outlining the overall vision of the paper. In [P8], the author has formulated the system model and general analytical framework, stated and proved the main and underlying mathematical results, and provided part of numerical results, which is based on analysis/simplified simulation.

Bibliography

[1] D. Raychaudhuri and N. B. Mandayam, "Frontiers of Wireless and Mobile Communications," *Proceedings of the IEEE*, vol. 100, pp. 824–840, 2012.

[2] GSMA Intelligence, *Understanding 5G: Perspectives on Future Technological Advancements in Mobile*, 2014.

[3] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart, "Networks and devices for the 5G era," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 90–96, 2014.

[4] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, 2014.

[5] W. Chin, Z. Fan, and R. Haines, "Emerging technologies and research challenges for 5G wireless networks," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 106–112, 2014.

[6] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?," *Selected Areas in Communications, IEEE Journal on*, vol. 32, no. 6, pp. 1065–1082, 2014.

[7] E. Hossain and M. Hasan, "5G cellular: key enabling technologies and research challenges," *Instrumentation & Measurement Magazine, IEEE*, vol. 18, no. 3, pp. 11–21, 2015.

[8] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, and Y. Selen, "5G radio access," *Ericsson Review*, vol. 6, pp. 2–7, 2014.

[9] P. P. Bocharov, C. D'Apice, and A. Pechinkin, *Queueing theory.* Walter de Gruyter, 2003.

[10] A. Borovkov, *Stochastic processes in queueing theory*, vol. 4. Springer Science & Business Media, 2012.

[11] G. I. Ivchenko, V. Kashtanov, and I. Kovalenko, "Queueing theory," *Vysshaya Shkola, Moscow*, 1982.

[12] V. Kalashnikov, *Mathematical Methods in Queuing Theory.* Springer Science & Business Media, 1993.

[13] D. Stoyan, "Applied stochastic geometry: A survey," *Biometrical Journal*, vol. 21, no. 8, pp. 693–715, 1979.

[14] A. K. Erlang, "The theory of probabilities and telephone conversations," *Nyt Tidsskrift for Matematik B*, vol. 20, no. 33-39, p. 16, 1909. English translations of both articles in Brockmeyer, E. et. al. (1948), The Life and Works of A.K.Erlang, The Copenhagen Telephone Company, Copenhagen.

[15] L. Takács, "An introduction to queueing theory," 1962.

[16] L. Kleinrock, "Queueing theory," *John Wiley & Sons, Inc. NY*, vol. 75, p. 76, 1975.

[17] B. V. Gnedenko and I. N. Kovalenko, *Introduction to queueing theory.* Birkhauser Boston Inc., 1989.

[18] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 7, pp. 1029–1046, 2009.

[19] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective," *Wireless Communications, IEEE*, vol. 21, no. 3, pp. 118–127, 2014.

[20] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic Geometry for Modeling, Analysis, and Design of Multi-Tier and Cognitive Cellular Wireless Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 15, pp. 996–1019, 2013.

[21] H. Frisch and J. Hammersley, "Percolation processes and related topics," *Journal of the Society for Industrial & Applied Mathematics*, vol. 11, no. 4, pp. 894–918, 1963.

[22] N. Campbell, "The study of discontinuous phenomena," in *Proceedings of the Cambridge Philosophical Society*, vol. 15, 1909.

[23] B. Bollobas and O. Riordan, *Percolation.* Cambridge University Press, 2006.

[24] H. Solomon, *Geometric probability*, vol. 28. Siam, 1978.

[25] D. Kendall, "An introduction to stochastic geometry," *Stochastic Geometry*, vol. 65, p. 1, 1974.

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

[27] R. Amir, J. Martin, J. Deaton, L. A. DaSilva, A. Hussien, and A. Eltawil, "Balancing spectral efficiency, energy consumption, and fairness in future heterogeneous wireless systems with reconfigurable devices," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 5, pp. 969 – 980, 2013.

[28] D. Lopez-Perez, I. Guvenc, G. de la Roche, M. Kountouris, T. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 22–30, 2011.

[29] M.-Y. Cheng, G.-Y. Lin, H.-Y. Wei, and A. Hsu, "Overload control for machine-type-communications in LTE-Advanced system," *IEEE Communications Magazine*, vol. 50, pp. 38–45, June 2012.

[30] R. Baldemair, E. Dahlman, G. Fodor, G. Mildh, S. Parkvall, Y. Selén, H. Tullberg, and K. Balachandran, "Evolving wireless communications: Addressing the challenges and expectations of the future," *IEEE Vehicular Technology Magazine*, vol. 8, no. 1, pp. 24–30, 2013.

[31] W. Ni and I. Collings, "A New Adaptive Small-Cell Architecture," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 5, pp. 829–839, 2013.

[32] T. Nakamura, S. Nagata, A. Benjebbour, Y. Kishiyama, T. Hai, S. Xiaodong, Y. Ning, and L. Nan, "Trends in small cell enhancements in LTE Advanced," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 98–105, 2013.

[33] Q. Li, H. Niu, A. Papathanassiou, and G. Wu, "5G Network Capacity: Key Elements and Technologies," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 71–78, 2014.

[34] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, 2014.

[35] M. Gerasimenko, N. Himayat, S.-p. Yeh, S. Talwar, S. Andreev, and Y. Koucheryavy, "Characterizing Performance of Load-Aware Network Selection in Multi-Radio (WiFi/LTE) Heterogeneous Networks," in *GLOBECOM Workshops (GC Wkshps)*, 2013.

[36] A. Y. Panah, S.-p. Yeh, N. Himayat, and S. Talwar, "Utility-based radio link assignment in multi-radio heterogeneous networks," in *GLOBECOM Workshops (GC Wkshps)*, pp. 618–623, IEEE, 2012.

[37] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014-2019*, February 2015.

[38] D. Lee, S.-I. Kim, J. Lee, and J. Heo, "Performance of multihop decode-and-forward relaying assisted device-to-device communication underlaying cellular networks," in *Proc. of the International Symposium on Information Theory and its Applications (ISITA)*, pp. 455–459, 2012.

[39] J. Lehtomaki, I. Suliman, J. Vartiainen, M. Bennis, A. Taparugssanagorn, and K. Umebayashi, "Direct communication between terminals in infrastructure based networks," in *Proc. of the ICT-MobileSummit*, 2008.

[40] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklos, and Z. Turanyi, "Design aspects of network assisted device-to-device communications," *IEEE Communications Magazine*, vol. 50, pp. 170–177, March 2012.

[41] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in LTE-Advanced networks," *IEEE Wireless Communications*, vol. 19, pp. 96–104, June 2012.

[42] K. Andersson and C. Ahlund, "Optimized Access Network Selection in a Combined WLAN/LTE Environment," *Wireless Personal Communications*, vol. 61, pp. 739–751, 2011.

[43] S.-P. Yeh, S. Talwar, G. Wu, N. Himayat, and K. Johnsson, "Capacity and coverage enhancement in heterogeneous networks," *IEEE Wireless Communications*, vol. 18, pp. 32–38, 2011.

[44] T. Shuminoski and T. Janevski, "Radio Network Aggregation for 5G Mobile Terminals in Heterogeneous Wireless and Mobile Networks," *Wireless Personal Communications*, vol. 78, no. 2, pp. 1211–1229, 2014.

[45] S.-P. Yeh, A. Y. Panah, N. Himayat, and S. Talwar, "QoS aware scheduling and cross-RAT coordination in multi-radio heterogeneous networks," in *IEEE VTC*, 2013.

[46] H. ElSawy and E. Hossain, "Two-Tier HetNets with Cognitive Femtocells: Downlink Performance Modeling and Analysis in a Multichannel Environment," *IEEE Transactions on Mobile Computing*, vol. 13, no. 3, pp. 649–663, 2014.

[47] S. Lee and K. Huang, "Coverage and Economy of Cellular Networks with Many Base Stations," *IEEE Communications Letters*, vol. 16, no. 7, pp. 1038–1040, 2012.

[48] S. Tombaz, A. Vastberg, and J. Zander, "Energy- and cost-efficient ultra-high-capacity wireless access," *IEEE Wireless Communications*, vol. 18, pp. 18–24, 2011.

[49] J. Park, S.-L. Kim, and J. Zander, "Asymptotic behavior of ultra-dense cellular networks and its economic impact," in *Global Communications Conference (GLOBECOM)*, pp. 4941–4946, IEEE, 2014.

[50] Q. Li, G. Wu, and R. Hu, “Analytical study on network spectrum efficiency of ultra dense networks,” in *Proc. of the IEEE Symposium on Personal, Indoor, and Mobile Radio Communications*, pp. 2764–2768, 2013.

[51] I. Hwang, B. Song, and S. Soliman, “A holistic view on hyper-dense heterogeneous and small cell networks,” *IEEE Communications Magazine*, vol. 51, no. 6, pp. 20–27, 2013.

[52] A. G. Gotsis and A. Alexiou, “On Coordinating Ultra-Dense Wireless Access Networks: Optimization Modeling, Algorithms and Insights,” *arXiv preprint: 1312.1577*, 2013.

[53] J. Xu, J. Wang, Y. Zhu, Y. Yang, X. Zheng, S. Wang, L. Liu, K. Horneman, and Y. Teng, “Cooperative distributed optimization for the hyper-dense small cell deployment,” *IEEE Communications Magazine*, vol. 52, no. 5, pp. 61–67, 2014.

[54] W. Cheung, T. Quek, and M. Kountouris, “Throughput Optimization, Spectrum Allocation, and Access Control in Two-Tier Femtocell Networks,” *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 561–574, 2012.

[55] J. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. Reed, “Femtocells: Past, Present, and Future,” *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, 2012.

[56] M. Bennis, S. Perlaza, P. Blasco, Z. Han, and V. Poor, “Self-Organization in Small Cell Networks: A Reinforcement Learning Approach,” *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3202–3212, 2013.

[57] M. Bennis, M. Simsek, W. Saad, S. Valentin, and M. Debbah, “When cellular meets WiFi in wireless small cell networks,” *IEEE Communications Magazine*, vol. 51, no. 6, pp. 44–50, 2013.

[58] A. Prasad, O. Tirkkonen, P. Lundén, O. N. C. Yilmaz, L. Dalsgaard, and C. Wijting, “Energy-efficient inter-frequency small cell discovery techniques for LTE-Advanced heterogeneous network deployments,” *IEEE Communications Magazine*, vol. 51, pp. 72–81, 2013.

[59] C. Sankaran, “Data offloading techniques in 3GPP Rel-10 networks: A tutorial,” *IEEE Communications Magazine*, vol. 50, pp. 46–53, 2012.

[60] S. Singh, H. Dhillon, and J. Andrews, “Offloading in Heterogeneous networks: Modeling, analysis, and design insights,” *IEEE Transactions on Wireless Communications*, vol. 12, pp. 2484–2497, 2013.

[61] L. Wang and G. S. Kuo, “Mathematical Modeling for Network Selection in Heterogeneous Wireless Networks – A Tutorial,” *IEEE Communications Surveys & Tutorials*, vol. 15, pp. 271–292, 2013.

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

[63] S. Singh, F. Baccelli, and J. Andrews, "On Association Cells in Random Heterogeneous Networks," *IEEE Wireless Communications Letters*, vol. 3, no. 1, pp. 70–73, 2014.

[64] H. Dhillon, R. Ganti, and J. Andrews, "Load-Aware Modeling and Analysis of Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 4, pp. 1666–1677, 2013.

[65] C. de Lima, M. Bennis, and M. Latva-aho, "Statistical Analysis of Self-Organizing Networks with Biased Cell Association and Interference Avoidance," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 1950–1961, 2013.

[66] T. Novlan, H. Dhillon, and J. Andrews, "Analytical modeling of uplink cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, pp. 2669–2679, 2013.

[67] S. Singh and J. Andrews, "Joint resource partitioning and offloading in Heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, pp. 888–901, 2014.

[68] *U-LTE: Unlicensed Spectrum Utilization of LTE, Huawei white paper, 2014.*

[69] N. Himayat, S.-P. Yeh, A. Y. Panah, S. Talwar, M. Gerasimenko, S. Andreev, and Y. Koucheryavy, "Multi-Radio Heterogeneous Networks: Architectures and Performance," in *Proc. of IEEE International Conference on Computing, Networking and Communications (ICNC)*, 2014.

[70] Y. Kojima, J. Suga, T. Kawasaki, M. Okuda, and R. Takechi, "LTE-WiFi Link Aggregation at Femtocell Base Station," in *Proc. of the World Telecommunications Congress*, pp. 1–6, 2014.

[71] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. Thomas, J. Andrews, P. Xia, H. Jo, H. Dhillon, and T. Novlan, "Heterogeneous cellular networks: From theory to practice," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 54–64, 2012.

[72] "IEEE 802.11-2012, Part 11: Local and metropolitan area networks," 2012.

[73] "Performance benefits of RAN level enhancements for WLAN/3GPP," *3GPP R2-133604*, 2013.

[74] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP Heterogeneous Networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, 2011.

[75] "Study on WLAN/3GPP Radio Interworking," *3GPP TR 37.834*, 2013.

[76] "Architecture enhancements for non-3GPP accesses," *3GPP Technical specification (TS) 23.402*, 2013.

[77] W. Nam, D. Bai, J. Lee, and I. Kang, "Advanced interference management for 5G cellular networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 52–60, 2014.

[78] S. Deb, P. Monogioudis, J. Miernik, and J. Seymour, "Algorithms for Enhanced Inter-Cell Interference Coordination (eICIC) in LTE HetNets," *IEEE/ACM Transactions on Networking*, vol. 22, no. 1, pp. 137–150, 2014.

[79] P. Marsch, B. Raaf, A. Szufarska, P. Mogensen, H. Guan, M. Farber, S. Redana, K. Pedersen, and T. Kolding, "Future Mobile Communication Networks: Challenges in the Design and Operation," *IEEE Vehicular Technology Magazine*, vol. 7, pp. 16–23, 2012.

[80] S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, and Y. Koucheryavy, "Cellular Traffic Offloading onto Network-Assisted Device-to-Device Connections," *IEEE Communications Magazine*, vol. 52, pp. 20–31, 2014.

[81] F. Fitzek, M. Katz, and Q. Zhang, "Cellular controlled short-range communication for cooperative P2P networking," *Wireless Personal Communications*, vol. 48, pp. 141–155, January 2009.

[82] A. Ozgur, O. Leveque, and D. Tse, "Hierarchical Cooperation Achieves Optimal Capacity Scaling in Ad Hoc Networks," *IEEE Transactions on Information Theory*, vol. 53, pp. 3549–3572, 2007.

[83] C.-H. Yu, K. Doppler, C. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *IEEE Transactions on Wireless Communications*, vol. 10, pp. 2752–2763, August 2011.

[84] *3GPP LTE Release 10 & beyond (LTE-Advanced).*

[85] K. Doppler, M. Rinne, C. Wijting, C. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to LTE-Advanced networks," *IEEE Communications Magazine*, vol. 47, pp. 42–49, December 2009.

[86] B. Kaufman, J. Lilleberg, and B. Aazhang, "Spectrum sharing scheme between cellular users and ad-hoc device-to-device users," *IEEE Transactions on Wireless Communications*, vol. 12, pp. 1038–1049, March 2013.

[87] M. Dohler, R. Heath, A. Lozano, C. Papadias, and R. Valenzuela, "Is the PHY layer dead?," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 159–165, 2011.

[88] M. Corson, R. Laroia, J. Li, V. Park, T. Richardson, and G. Tsirtsis, "Toward proximity-aware internetworking," *IEEE Wireless Communications*, vol. 17, pp. 26–33, December 2010.

[89] S. Xu, H. Wang, T. Chen, T. Peng, and K. Kwak, "Device-to-device communication underlaying cellular networks: Connection establishment and interference avoidance," *KSII Transactions on Internet and Information Systems*, vol. 6, pp. 203–228, January 2012.

[90] H. Min, W. Seo, J. Lee, S. Park, and D. Hong, "Reliability improvement using receive mode selection in the device-to-device uplink period underlaying cellular networks," *IEEE Transactions on Wireless Communications*, vol. 10, pp. 413–418, February 2011.

[91] L. Al-Kanj, Z. Dawy, and E. Yaacoub, "Energy-Aware Cooperative Content Distribution over Wireless Networks: Design Alternatives and Implementation Aspects," *IEEE Communications Surveys & Tutorials*, vol. 15, pp. 1736–1760, 2013.

[92] H. Luo, R. Ramjee, P. Sinha, L. Li, and S. Lu, "UCAN: A unified cellular and Ad-Hoc network architecture," in *Proc. of the International Conference on Mobile Computing and Networking (MobiCom)*, pp. 353–367, 2003.

[93] J. Luo, R. Mukerjee, M. Dillinger, E. Mohyeldin, and E. Schulz, "Investigation of radio resource scheduling in WLANs coupled with 3G cellular network," *Communications Magazine, IEEE*, vol. 41, no. 6, pp. 108–115, 2003.

[94] H. Luo, X. Meng, R. Ramjee, P. Sinha, and L. Li, "The design and evaluation of unified cellular and ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 6, pp. 1060–1074, September 2007.

[95] D. Lopez-Perez, X. Chu, and I. Guvenc, "On the Expanded Region of Picocells in Heterogeneous Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 3, pp. 281–294, 2012.

[96] M. Peng, D. Liang, Y. Wei, J. Li, and H.-H. Chen, "Self-configuration and self-optimization in LTE-Advanced heterogeneous networks," *IEEE Communications Magazine*, vol. 51, pp. 36–45, 2013.

[97] D. Raychaudhuri and N. Mandayam, "Frontiers of wireless and mobile communications," *Proceedings of the IEEE*, vol. 100, no. 4, pp. 824–840, 2012.

[98] F. Meshkati, H. Poor, and S. Schwartz, "Energy-efficient resource allocation in wireless networks," *IEEE Signal Processing Magazine*, vol. 24, pp. 58–68, May 2007.

[99] F. Meshkati, H. Poor, S. Schwartz, and R. Balan, "Energy-efficient resource allocation in wireless networks with quality-of-service constraints," *IEEE Transactions on Communications*, vol. 57, pp. 3406–3414, November 2009.

[100] W. Song and W. Zhuang, "Multi-service load sharing for resource management in the cellular/WLAN integrated network," *IEEE Transactions on Wireless Communications*, vol. 8, pp. 725–735, February 2009.

[101] K. Huang, V. Lau, and Y. Chen, "Spectrum sharing between cellular and mobile ad hoc networks: Transmission-capacity trade-off," *IEEE Journal on Selected Areas in Communications*, vol. 27, pp. 1256–1267, September 2009.

[102] X. Liu, E. Chong, and N. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks*, vol. 41, pp. 451–474, March 2003.

[103] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *IEEE/ACM Transactions on Networking*, vol. 13, pp. 636–647, June 2005.

[104] S. Patil and G. de Veciana, "Reducing feedback for opportunistic scheduling in wireless systems," *IEEE Transactions on Wireless Communications*, vol. 6, pp. 4227–4232, December 2007.

[105] S. Patil and G. de Veciana, "Measurement-based opportunistic scheduling for heterogeneous wireless systems," *IEEE Transactions on Communications*, vol. 57, pp. 2745–2753, September 2009.

[106] D. Ohmann, A. Fehske, and G. Fettweis, "Transient flow level models for interference-coupled cellular networks," in *Proc. of the Annual Allerton Conference on Communication, Control, and Computing*, pp. 723–730, 2013.

[107] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed alpha-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 20, pp. 177–190, February 2012.

[108] H. Kim and G. de Veciana, "Leveraging dynamic spare capacity in wireless systems to conserve mobile terminals' energy," *IEEE/ACM Transactions on Networking*, vol. 18, pp. 802–815, June 2010.

[109] H. Kim, *Exploring Tradeoffs in Wireless Networks under Flow-Level Traffic: Energy, Capacity and QoS.* PhD thesis, University of Texas at Austin, 2009.

[110] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *Proc. of IEEE INFOCOM*, 2013.

[111] J. G. Andrews, R. Ganti, M. Haenggi, and N. Jindal, "A primer on spatial modeling and analysis in wireless networks," *IEEE Communications Magazine*, vol. 48, pp. 156–163, 2010.

[112] J. G. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, pp. 136–144, 2013.

[113] A. G. Gotsis, S. Stefanatos, and A. Alexiou, "Spatial coordination strategies in future ultra-dense wireless networks," in *Wireless Communications Systems (ISWCS), 2014 11th International Symposium on*, pp. 801–807, IEEE, 2014.

[114] F. Baccelli, N. Khude, R. Laroia, J. Li, T. Richardson, S. Shakkottai, S. Tavildar, and X. Wu, "On the design of device-to-device autonomous discovery,"

in *Proc. of the International Conference on Communication Systems and Networks (COMSNETS)*, 2012.

[115] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, pp. 550–560, April 2012.

[116] M. Erturk, S. Mukherjee, H. Ishii, and H. Arslan, "Distributions of transmit power and SINR in device-to-device networks," *IEEE Communications Letters*, vol. 17, pp. 273–276, February 2013.

[117] W. S. Kendall, "Perfect simulation for the area-interaction point process," in *Probability towards 2000*, pp. 218–234, Springer, 1998.

[118] D. Stoyan, W. S. Kendall, J. Mecke, and L. Ruschendorf, *Stochastic geometry and its applications*, vol. 2. Wiley New York, 1987.

[119] A. J. Baddeley and M. Van Lieshout, "Area-interaction point processes," *Annals of the Institute of Statistical Mathematics*, vol. 47, no. 4, pp. 601–619, 1995.

[120] J. Moller and R. P. Waagepetersen, *Statistical inference and simulation for spatial point processes.* CRC Press, 2003.

[121] V. Isham, "An introduction to spatial point processes and markov random fields," *International Statistical Review/Revue Internationale de Statistique*, pp. 21–43, 1981.

[122] T. Rappaport, *Wireless communications: principles and practice*, vol. 2. Prentic Hall PTR, New Jersey, 1996.

[123] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[124] P. Mogensen, W. Na, I. Z. Kovács, F. Frederiksen, A. Pokhariyal, K. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, "LTE capacity compared to the Shannon bound," in *Vehicular Technology Conference (VTC2007-Spring)*, pp. 1234–1238, IEEE, 2007.

[125] N. Himyat, S. Talwar, A. Rao, and R. Soni, "Interference management for 4G cellular standards," *IEEE Communications Magazine*, pp. 86–94, 2010.

[126] S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson, and Y. Koucheryavy, "Analyzing assisted offloading of cellular user sessions onto D2D links in unlicensed bands," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 1, pp. 67–80, 2015.

[127] S. Foss and S. Zuyev, "On a Voronoi Aggregative Process related to a Bivariate Poisson Process," *Advances in Applied Probability*, vol. 28, pp. 965–981, 1996.

[128] *3GPP TR 36.814, Further advancements for E-UTRA physical layer aspects*, 2010.

[129] *3GPP TR 36.819. Coordinated multi-point operation for LTE physical layer aspects*, September 2013.

[130] K. Smiljkovikj, H. Elshaer, P. Popovski, F. Boccardi, M. Dohler, L. Gavrilovska, and R. Irmer, "Capacity analysis of decoupled downlink and uplink access in 5G heterogeneous systems," *arXiv preprint arXiv:1410.7270*, 2014.

[131] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, "Downlink and uplink decoupling: a disruptive architectural design for 5g networks," in *Global Communications Conference (GLOBECOM)*, pp. 1798–1803, IEEE, 2014.

[132] M. Mirahsan, R. Schoenen, and H. Yanikomeroglu, "HetHetNets: Heterogeneous Traffic Distribution in Heterogeneous Wireless Cellular Networks," *IEEE Journal on Selected Areas in Communications*, 2015.

Publications

Publication 1

© 2012 Springer Science + Business Media. Reprinted, with permission, from

O. Galinina, S. Andreev, and Y. Koucheryavy, ”Performance Analysis of Client Relay Cloud in Wireless Cellular Networks”, in *Proc. of the 10th International Conference on Wired/Wireless Internet Communications (WWIC)*, 2012.

Performance Analysis of Client Relay Cloud in Wireless Cellular Networks

Olga Galinina, Sergey Andreev, and Yevgeni Koucheryavy

Tampere University of Technology (TUT), Finland
{olga.galinina,sergey.andreev}@tut.fi,
yk@cs.tut.fi

Abstract. Cooperative communication is a promising concept to mitigate the effect of fading in a wireless channel and is expected to improve performance of next-generation cellular networks in terms of client throughput and energy efficiency. With recent proliferation of smart phones and machine-to-machine communication, so-called 'client relay' cooperative techniques are becoming more important. As such, a mobile client with poor channel quality may take advantage of other neighboring clients, who would relay data on its behalf. In the extreme, the aggregate set of available client relays may form a relay cloud, and members of the cloud may opportunistically cooperate with the data originator to improve its uplink channel quality. The key idea behind the relay cloud is to provide flexible and distributed control over cooperative communication by the wireless clients themselves. By contrast to centralized control, this will minimize extra protocol signaling involved and ensure simpler implementation. In this work, we build an originator-centric model and study the performance of a relay cloud with respect to the main performance metrics: throughput, packet delay, and energy efficiency. We obtain closed-form analytical expressions for the sought metrics and verify our results via extensive simulations.

Keywords: client relay cloud, cellular networks, performance analysis, throughput, energy efficiency.

1 Introduction and Related Work

Various diversity techniques aim at mitigating the negative effects of multipath channel fading in order to improve the reliability of wireless communication link. In particular, one of the most promising techniques for next-generation mobile systems (3GPP LTE-Advanced, IEEE 802.16m) is *spatial transmit diversity* exploiting two or more transmit antennas to enhance the link quality [1]. However, mobile terminals with multiple transmit antennas may be costly due to their size or hardware limitations. For that reason, a concept of *cooperative communication* has been introduced allowing single-antenna mobiles to take advantage of spatial diversity gain and provide so-called *cooperative diversity.*

Y. Koucheryavy et al. (Eds.): WWIC 2012, LNCS 7277, pp. 40–51, 2012.
© Springer-Verlag Berlin Heidelberg 2012

Historically, the core ideas behind cooperative communication were firstly introduced in the fundamental work [2], where a simplified three-terminal system model containing a *sender*, a *receiver*, and a *relay* was studied within the context of mutual information. More thorough capacity analysis of the relay channel was conducted later in [3]. These pioneering efforts focused on the similar three-node case and suggested a number of relaying strategies. They also established achievable regions and upper bounds on the capacity of what we now call the 'classical' relay channel.

With recent proliferation of smart phones and machine-to-machine communication, wireless technology is rapidly evolving toward 4G mobile systems. As such, a renewed surge of interest has come with rapidly expanding literature on cooperation. For example, [4] addressed some further information-theoretic aspects of the relay channel bringing new important insights.

More specifically, cooperative diversity was described in [5] as a relatively new class of spatial diversity techniques that is enabled by relaying and cooperative communication. In [6], authors proposed an efficient cooperation strategy and also explored the concept of cooperation together with some practical issues of its implementation. A good tutorial on cooperative communication may be found in [7].

Some recent works have also addressed a more complicated usage model with multiple wireless clients that may be selected as relays. The problem of relay selection (when data originator may practically have more than one relay to partner with) has been elaborated upon in [8], where the availability of a centralized cooperation-aware controller was assumed. Thus, it brings the concept of cooperation into the scope of wireless cellular networks with a *base station* controlling the activity of its clients.

Further, in [9] several efficient protocols for the relay selection were proposed to recover the multiplexing loss in relay networks, while requiring additional feedback. Evidently, most recent works study cooperation from the perspective of centralized control, which increases extra protocol signaling involved and results in more difficult implementation for the existing systems. By contrast, we concentrate on a more practical scenario with flexible and distributed control over cooperative communication by the wireless clients themselves.

In our previous work [10], we tailored the 'classical' three-node cooperative model to contemporary wireless networks and analyzed primary QoS parameters together with the most important energy-related metrics. In this paper, we continue our efforts by considering a system with multiple clients. The data originator may opportunistically partner with some of those to improve its uplink channel quality. In the extreme, the aggregate set of available relays may form a *relay cloud*. Thus, the goal of our research is to investigate the benefits of the relay cloud and to develop and assess algorithms that will maximize the impact of cooperative communication.

The rest of the paper is organized as follows. Section 2 describes the system model, while giving the main notations and assumptions. In Section 3, we provide theoretical analysis of the client relay cloud and establish the expressions for the main performance metrics. In Section 4, we consider some numerical results verified by simulation. Finally, Section 5 concludes the paper.

2 System Model

In this section, we model a wireless cellular network consisting of the base station B and several mobile clients (see Figure 1 for the topology and the Table 1 for the notations).

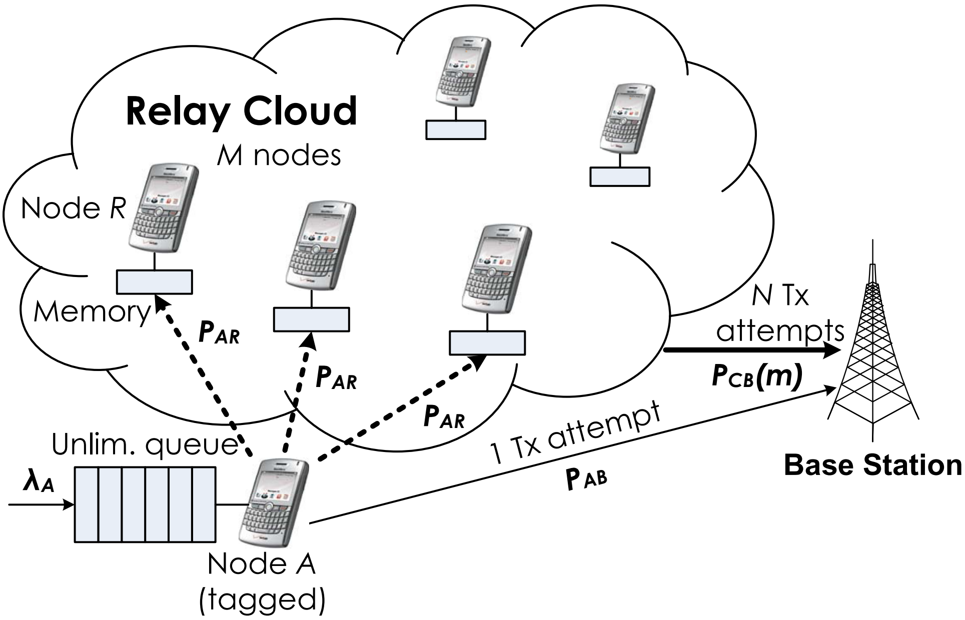


Fig. 1. Illustration of the relay cloud system topology

We define the *cooperative system* as follows. The wireless clients acting as relay nodes are allowed to eavesdrop on the data packets from the originator. As mentioned previously, the aggregate set of available client relays forms a relay cloud. After successful eavesdropping, the members of the cloud may opportunistically transmit on behalf of the data originator to improve its uplink performance. The base station only provides time resources (slots) for such cooperative transmission, whereas opportunistic control resides at the client side. If cooperation is not possible, we term the system *non-cooperative.*

Further on, for the sake of simplicity and without loss of generality we consider the performance of the tagged node A (the data originator). It is assumed, for example, that A is a cell-edge mobile user and thus suffers from the low quality of its uplink channel to the base station. The rest of M neighboring wireless clients (the relay cloud) may potentially perform cooperation acting as relays.

While the originator transmits its initial data packet, each relay node in the cloud may eavesdrop on this packet and store it for subsequent retransmission.

Table 1. Analytical model notations

Notation	Parameter description
λ_A	Mean arrival rate of packets to node A
N	Maximum number of attempts provided by B for cloud transmissions
M	Number of relay nodes in the relay cloud
m	Number of relay nodes in the relay group
p_{AB}	Probability of successful reception at B when A transmits
p_{AR}	Probability of successful reception at R when A transmits
p_{tx}	Opportunistic cooperation probability
$p_{CB}(m)$	Probability of successful reception at B when cloud cooperates
τ_A	Mean service time of a packet from the node A
ρ_A	Queue load coefficient
P_{lossA}	Loss probability of the packet from A
δ_A	Mean packet delay of the packet from A
η_A	Mean throughput of node A
ϵ_A	Mean energy expenditure of node A
ϵ_R	Mean energy expenditure of relay group
ϕ	Mean energy efficiency of the system
P_{TX}	Power level for the transmitting node
P_{RX}	Power level for the eavesdropping node
P_I	Power level for the idle node

The size of extra memory location at each relay is assumed to equal one for every relay session, whereas the size of the outgoing originator buffer is unlimited. In case the originator fails its initial transmission and if eavesdropping is successful, the relay node R decides probabilistically whether to cooperate or not.

The successful relay nodes which decide to cooperate form a so-called *relay group*. We emphasize that the proposed scheme does not require explicit centralized control by the base station and thus minimizes the necessary signaling. The base station may be completely unaware of which nodes belong to the relay group at a particular time instant. Below we detail the system model.

Traffic assumptions. We consider a simple stochastic traffic model to assess the performance of the system and preserve the analytical tractability. As the first step of this research, we assume i.i.d. exponentially-distributed inter-arrival times at the originator (node A). We concentrate on the originator traffic only and abstract out the analysis of own traffic in the relay cloud, which may be more complex. Base station also has no outgoing traffic.

Scheduler assumptions. The system time is slotted. We assume that the packet size equals one and that the transmission of each packet takes exactly one time slot. Scheduling information is immediately available to all the clients (e.g., via a dedicated downlink control channel).

We consider the scheduler operation as follows. As the channel between the node A and the destination is poor, it is very likely that several packet retransmissions may not lead to success. As such, we assume only one attempt to transmit a packet by the originator to save some of its power. If the originator A

has packets, the next time slot is given to A. Upon the transmission, a potential relay may intercept the packet from the originator and store it.

In case node A fails its initial packet transmission, the base station assigns the following slot to the relay cloud so that it could assist the originator. Such assignment repeats until successful delivery or until the number of consecutive cloud retransmission attempts exceeds some maximum number N (a parameter controlled by the base station). In the latter case, all the members of the cloud may empty their memory location and the system considers the current packet as lost.

Channel assumptions. Throughout this paper, we assume immediate feedback over a reliable separate channel (e.g., in the downlink). We also account for the following probabilities of successful delivery p_{AB}, $p_{CB}(m)$ and the symmetric probability for each relay node p_{AR}:

- $p_{AB} = Pr\{packet\ from\ A\ is\ received\ at\ B|only\ A\ transmits\}$,
- $p_{CB}(m) = Pr\{packet\ from\ A\ is\ received\ at\ B|exactly\ m\ relays\ transmit\}$,
- $p_{AR} = Pr\{packet\ from\ A\ is\ received\ at\ a\ given\ relay|only\ A\ transmits\}$.

Let us now illustrate the discussed scheduler operation by an example for $N = 2$ as shown in Figure 2. Firstly, the transmission of packet no. 0 is successful and the system becomes idle. Then, the originator acquires a new packet no. 1 and attempts its uplink transmission, which fails with probability $1 - p_{AB}$. The members of the relay cloud eavesdrop on this transmission and each of them is successful with probability p_{AR} independently. In the following slot, the successful relays make a decision whether or not to help with probability p_{tx}. Those who have decided positively form a relay group that retransmits the eavesdropped packet to the base station simultaneously. As such, a 'virtual MIMO' link with better quality is created due to spatial transmit diversity [11] and the packet is transmitted successfully with probability $p_{CB}(m)$.

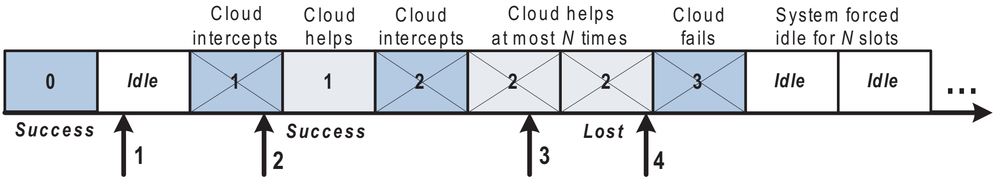


Fig. 2. Example time diagram for the relay cloud system

We generally note that due to the diversity gain the probability $p_{CB}(m)$ is expected to be a nondecreasing function of the relay group size m. Here, m depends on the probabilities p_{AR} and p_{tx}. We implicitly assume that the quality of the "originator-to-base station" channel is low, whereas the quality of the "originator-to-relay cloud" channel is quite good due to many neighboring clients available.

The following slot is given back to the originator A (packet no. 2) and during its unsuccessful transmission the relays intercept the packet again. However, this time the relay cloud is unsuccessful to transmit the packet for $N = 2$ times consecutively. As such, the base station considers the current packet as lost and assigns the next slot to the originator. Further, if the interception fails (packet no. 3) or all the successful relays decide not to transmit twice, $N = 2$ slots are assigned to the relay cloud anyway, but the system stays idle. This is a negative consequence of the distributed control over client relays.

In what follows, we study the mean packet delay, the throughput, and the packet loss probability. In particular, we are interested in the derivation of simple and exact closed-form expressions.

3 Performance Evaluation

This section presents analysis of the relay cloud system with respect to the main performance metrics, such as the mean *number of retransmissions*, the *throughput*, the *packet loss probability*, and the mean *packet delay*.

Firstly, we introduce the following definitions:

Definition 1. The service time is defined as the period of time between the beginning of the first transmission attempt and the moment packet reaches its destination. In case of packet loss, the service time is assumed to be equal N, so that the mean service time could account for the lost packets.

Definition 2. The saturation throughput is defined as the limit reached by the system throughput as the offered load increases [12].

Definition 3. The delay of a packet is defined as the time it takes the packet to reach the destination after it arrives in the system (includes both queueing time and service time).

Definition 4. The energy efficiency is defined as the amount of energy required to successfully transmit one data packet.

Our analytical approach is based on the notion of the service time. We define a stochastic variable T_A, which is the service time of a packet from A. We treat the considered system as an $M/G/1$ system due to the properties of the incoming traffic. Initially, we establish the service discipline and then continue by obtaining the closed-form expressions for the first and the second moments. It should be noted that the first moment is the mean number of the packet transmission attempts.

Knowing both moments, we derive the mean packet delay using the Pollacek-Khinchin formula and the Little's law. The other metrics of interest, such as the throughput, the packet loss probability, the energy expenditure, and the energy efficiency can also be derived from the obtained expressions.

After thorough analysis of all the possibilities for a packet transmission, we formulate the service discipline for the node A as follows:

$$
\begin{aligned}
&Pr\{T_A = 1\} = p_{AB},\\
&Pr\{T_A = n\} = (1 - p_{AB}) \sum_{m=1}^{M} \binom{M}{m} p_{AR}^{m} (1 - p_{AR})^{(M-m)} \times\\
&\qquad \times \left\{ \sum_{j=1}^{m} \binom{m}{j} p_{tx}^{j} (1 - p_{tx})^{(m-j)} \left(1 - p_{CB}(j)\right)^{(n-1)} p_{CB}(j) \right\}, n \leq N,\\
&Pr\{T_A = N+1\} = (1 - p_{AB}) \sum_{m=1}^{M} \binom{M}{m} p_{AR}^{m} (1 - p_{AR})^{(M-m)} (1 - p_{tx})^{m} +\\
&\qquad + (1 - p_{AB}) \sum_{m=1}^{M} \binom{M}{m} p_{AR}^{m} (1 - p_{AR})^{(M-m)} \times\\
&\qquad \times \left\{ \sum_{j=1}^{m} \binom{m}{j} p_{tx}^{j} (1 - p_{tx})^{(m-j)} \left(1 - p_{CB}(j)\right)^{N} \right\} +\\
&\qquad + (1 - p_{AB}) (1 - p_{AR})^{M}.
\end{aligned}
$$

Further, we omit massive transformations and give only the expressions for the first and the second moments of the stochastic variable T_A:

$$
\begin{aligned}
\tau_A = E[T_A] = p_{AB} + (N+1)(1 - p_{AB})(1 - p_{AR})^{M} +\\
+(N+1)(1 - p_{AB}) \cdot S_1 + (1 - p_{AB}) \cdot S_2,
\end{aligned}
\qquad (1)
$$

$$
E[T_A^2] = p_{AB} + (N+1)^2 (1 - p_{AB})((1 - p_{AR})^{M} + S_1) + (1 - p_{AB}) \cdot S_3, \qquad (2)
$$

where components S_1, S_2 and S_3 are given in the Appendix.

The mean load for the queue of the considered tagged node A can be established as:

$$
\rho_A = \lambda_A \tau_A. \qquad (3)
$$

The mean throughput of A may thus be calculated as:

$$
\eta_A = \lambda_A (p_{AB} + (1 - p_{AB}) S_4), \qquad (4)
$$

where component S_4 is also given in the Appendix.

Given the first and second moments, we use the Pollacek-Khinchin formula to obtain the accurate value for the mean packet delay:

$$
\delta_A = \tau_A + \frac{E[T_A^2]\lambda_A}{2(1 - \rho_A)}. \qquad (5)
$$

With the basic formulae for the two moments, we can also find other important metrics. In particular, the packet loss probability is given by:

$$
P_{lossA} = 1 - \eta_A \tau_A. \qquad (6)
$$

Furthermore, let us establish the expressions for the energy consumption. If power level P_i corresponds to a particular power state i, then the normalized energy expenditure per time slot equals $P_i \pi_i$. Here, $\boldsymbol{\pi}$ is the stationary distribution over power states and $i \subset G$, where G is the set of possible states. In the considered model, the three states are accounted for from the power perspective:

- the node is transmitting data with the power P_{TX};
- the node is receiving data with the power P_{RX};
- the node is idle with the power P_I.

Thus, energy expenditures of the tagged node A and of the relay group are calculated as:

$$\epsilon_A = P_{TX}\lambda_A + P_I(1 - \lambda_A \tau_A), \tag{7}$$

$$\begin{aligned} \epsilon_R = P_{RX} M \lambda_A + P_{TX} p_{AR} p_{tx} \lambda_A (\tau_A - 1) M + \\ + P_i M (1 - p_{AR} p_{tx} \lambda_A (\tau_A - 1) - \lambda_A), \end{aligned} \tag{8}$$

Therefore, the total system energy expenditure is given by:

$$\epsilon = \epsilon_A + \epsilon_R. \tag{9}$$

As mentioned above, we define energy efficiency as:

$$\phi = \frac{\eta_A}{\epsilon}. \tag{10}$$

4 Numerical Results

In this section, we use the extended system-level simulator described in our previous works [10] and [11] in order to verify the obtained analytical results. We borrow power consumption values from [13] as: $P_{TX} = 1.65$ W, $P_{RX} = 1.40$ W, and $P_I = 1.15$ W. We also assume that the size of the each slot equals 5 ms.

The main simulation parameters are set as $p_{AB} = 0.3$, $p_{AR} = 0.7$. For the vector of successful delivery probabilities p_{CB}, as an example, we consider a random nondecreasing linear function. However, in reality this function might be much more complicated and surely has a nonlinear structure. Note that solid curves stand for the analytical results, whereas symbols represent simulated data.

In Figure 3, we explore the behavior of the saturation throughput for different number of relay nodes in the cloud. As expected, we observe a monotonically increasing function of M. It is easy to see that beyond the point of $M = 5$ the curve becomes almost linear. Therefore, we set $M = 5$ in what follows.

Further, we explore the mean packet delay at the originator for different numbers relay of nodes available. For this purpose, we vary the arrival rate λ_A. In Figure 4, different curves for the various values of M (with appropriate asymptotes) are compared. Naturally, the delay drops significantly as the number of available relays grows.

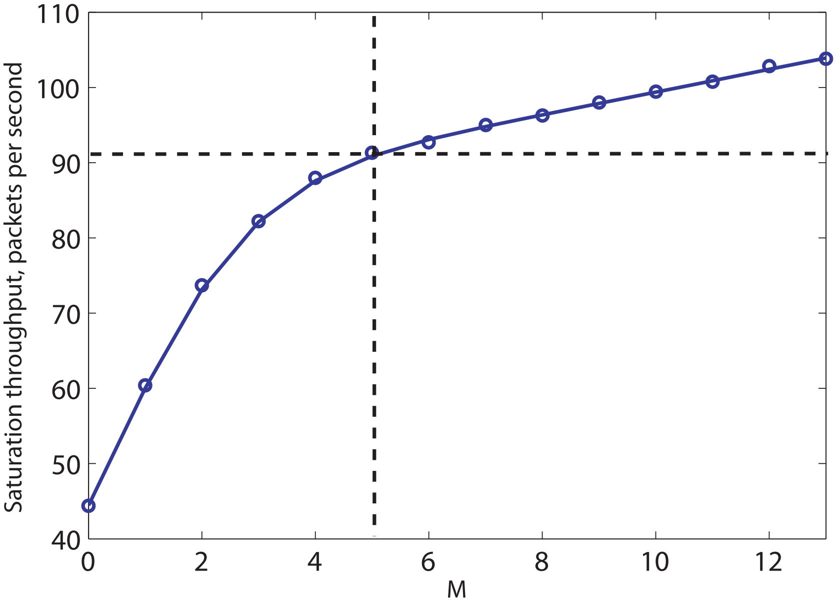


Fig. 3. Saturation throughput vs. number of nodes in relay cloud

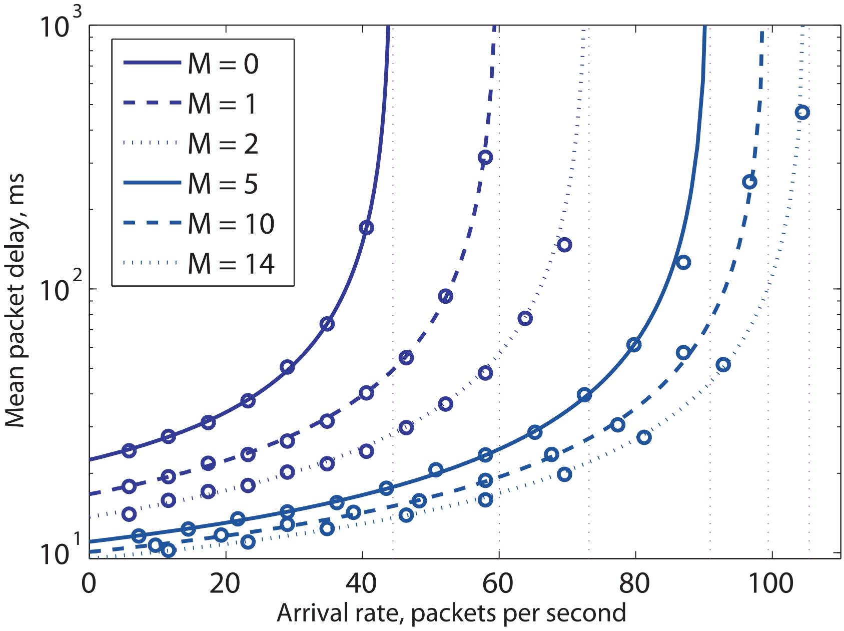


Fig. 4. Mean packet delay vs. arrival rate λ_A for different values of M

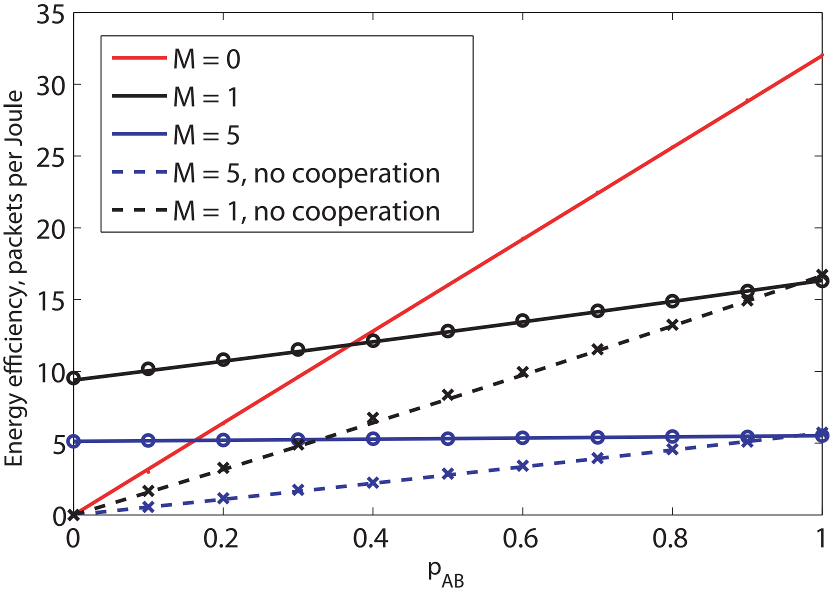


Fig. 5. Energy efficiency vs. probability of successful packet transmission

Also we study the energy efficiency dependence on e.g. the probability p_{AB} in Figure 5. Here, we contrast the non-cooperative mode (when there are no relay nodes) against the systems with $M = 1$ and $M = 5$. Evidently, the assistance of the relay cloud results in slightly higher energy expenditure, which is the cost of the increased originator performance.

5 Conclusion

In this paper, we considered a wireless cellular network that enables the distributed control over cooperative communication via a client relay cloud. The primary aim of such cloud is to enhance system performance through the support of cell-edge mobile clients with poor communication links. The main performance metrics were studied, including throughput, mean packet delay, and energy efficiency. Accurate closed-form analytical expressions have been derived and verified by extensive system-level simulations. The results indicate significant promise of the relay cloud, which is able to recover the performance of the mobile clients with degraded wireless links. As a future extension of this model, it would be reasonable to examine a more realistic arrival process and propose efficient decision-making algorithms on when to co-operate. Also it is important to establish practical scenarios where client relay cloud operation is benefiting the wireless system performance and where it is not.

References

1. LTE Release 10 & beyond (LTE-Advanced)
2. Van Der Meulen, E.C.: Three-terminal communication channels. Advances in Applied Probability 3, 120–154 (1971)
3. Cover, T.M., El Gamal, A.A.: Capacity theorems for the relay channel. IEEE Transactions on Information Theory 25(5), 572–584 (1979)
4. Kramer, G., Gastpar, M., Gupta, P.: Cooperative strategies and capacity theorems for relay networks. IEEE Transactions on Information Theory 51(9), 3037–3063 (2005)
5. Laneman, J., Tse, D., Wornell, G.: Cooperative diversity in wireless networks: Efficient protocols and outage behavior. IEEE Transactions on Information Theory 50(12), 3062–3080 (2004)
6. Sendonaris, A., Erkip, E., Aazhang, B.: User cooperation diversity. Part I, II. IEEE Transactions on Communications 51(11), 1927–1938 (2003)
7. Nosratinia, A., Hunter, T., Hedayat, A.: Cooperative communication in wireless networks. IEEE Communications Magazine 42(10), 74–80 (2004)
8. Nosratinia, A., Hunter, T.: Grouping and partner selection in cooperative wireless networks. IEEE Journal on Selected Areas in Communications 25(2), 369–378 (2007)
9. Tannious, R., Nosratinia, A.: Spectrally efficient relay selection with limited feedback. IEEE Journal on Selected Areas in Communications 26(8), 1419–1428 (2008)
10. Andreev, S., Galinina, O., Lokhanova, A., Koucheryavy, Y.: Analysis of Client Relay Network with Opportunistic Cooperation. In: Masip-Bruin, X., Verchere, D., Tsaoussidis, V., Yannuzzi, M. (eds.) WWIC 2011. LNCS, vol. 6649, pp. 247–258. Springer, Heidelberg (2011)
11. Andreev, S., Galinina, O., Vinel, A.: Performance evaluation of a three node client relay system. International Journal of Wireless Networks and Broadband Technologies, IJWNBT 1(1), 73–84 (2011)
12. Bianchi, G.: Performance analysis of the IEEE 802.11 distributed coordination function. IEEE Journal on Selected Areas on Communications 18, 535–547 (2000)
13. Andreev, S., Galinina, O., Koucheryavy, Y.: Energy-efficient client relay scheme for machine-to-machine communication. In: Proceedings of GLOBECOM (2011)

Appendix 1: Auxiliary Variables

We introduce the following auxiliary variables in order to simplify the expressions above.

$$S_1 = \sum_{m=0}^{M} \binom{M}{m} p_{AR}^{m}(1-p_{AR})^{(M-m)} \times \tag{11}$$
$$\times \left\{ \sum_{j=0}^{m} \binom{m}{j} p_{tx}^{j}(1-p_{tx})^{(m-j)}(1-p_{CB}(j)) + (1-p_{tx})^{m} \right\}^{N} .$$

$$S_2 = \sum_{m=0}^{M} \binom{M}{m} p_{AR}^{m}(1-p_{AR})^{(M-m)} \cdot X \cdot b, \tag{12}$$

where

$$X = a^{N} - \frac{a^{(N+1)}}{(1-a)^2} - \frac{(a-2)}{(1-a)^2} - a^{N}\frac{(N+2)}{(1-a)}. \tag{13}$$

$$S_3 = \sum_{m=0}^{M} \binom{M}{m} p_{AR}^{m}(1-p_{AR})^{(M-m)} \cdot Z \cdot b, \tag{14}$$

where

$$Z = X + \frac{2(2a - a^{(N+1)}(N+2))}{(1-a)^2} + \frac{2(a^2 - a^{(N+2)})}{(1-a)^3} + \frac{2 - a^{N}(N+1)(N+2)}{(1-a)}. \tag{15}$$

The following variable is the probability of the successful transmission by the relays:

$$S_4 = \sum_{m=0}^{M} \binom{M}{m} p_{AR}^{m}(1-p_{AR})^{(M-m)} \cdot Y \cdot b, \tag{16}$$

where

$$Y = \frac{(1-a^{N})}{(1-a)}. \tag{17}$$

Here, also for the sake of brevity the enlarged variables a and b are defined as:

$$a = \sum_{j=0}^{m} \binom{m}{j} p_{tx}^{j}(1-p_{tx})^{(m-j)}(1-p_{CB}(j)) + (1-p_{tx})^{m}, \tag{18}$$

$$b = \sum_{j=0}^{m} \binom{m}{j} p_{tx}^{j}(1-p_{tx})^{(m-j)} p_{CB}(j). \tag{19}$$

Publication 2

© 2013 Springer Science + Business Media. Reprinted, with permission, from

O. Galinina, A. Trushanin, V. Shumilov, R. Maslennikov, Z. Saffer, S. Andreev, and Y. Koucheryavy, ”Energy-Efficient Operation of a Mobile User in a Multi-Tier Cellular Network”, in *Proc. of the 20th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA)*, 2013.

Energy-Efficient Operation of a Mobile User in a Multi-tier Cellular Network

Olga Galinina[1], Alexey Trushanin[2], Vyacheslav Shumilov[2],
Roman Maslennikov[2], Zsolt Saffer[3],
Sergey Andreev[1], and Yevgeni Koucheryavy[1]

[1] Tampere University of Technology (TUT), Finland
{olga.galinina,sergey.andreev}@tut.fi, yk@cs.tut.fi
[2] Lobachevsky State University of Nizhny Novgorod (UNN), Russia
{alexey.trushanin,vyacheslav.shumilov,roman.maslennikov}@wcc.unn.ru
[3] Budapest University of Technology and Economics (BUTE), Hungary
safferzs@hit.bme.hu

Abstract. In this paper[1], we propose a new power control scheme suitable for a multi-tier wireless network. It maximizes the energy-efficiency of a mobile device transmitting on several communication channels while at the same time ensures the required minimum quality of service. As the result, a good compromise between improving the data rate and extending the battery lifetime is provided. In order to enable energy-efficiency maximization, we formulate an optimization problem basing on the Shannon's capacity formula. The optimal transmit power is thus obtained from the direct solution of this optimization problem under several practical constraints, such as minimum bitrate and maximum transmit power. In the second part of the paper, we apply extensive simulations to calibrate the key parameters of our optimization framework. The numerical results suggest the benefit of the proposed analytical solution by comparing it against intuitive (heuristic) power control strategies.

1 Introduction and Background

Wireless cellular networks have experienced essential growth over the last decades, eventually becoming an integrated part of our daily lives [1]. As market analysts predict, this steady development is expected to continue over the following years [2]. Hence, it comes as no surprise that users are increasingly interested in extending functionality of their mobile devices to run more demanding applications. The resulting advent of the high-rate fourth generation (4G) communication technologies combined with a wide variety of new mobile devices and services brings substantial increase in the amounts of user-generated data [3]. This, in turn, implies higher power consumption when transmitting this data, which may be harmful for the battery-powered mobile devices [4]. As a result,

[1] Part of this work had been completed when Alexey Trushanin and Vyacheslav Shumilov were on a research visit at Tampere University of Technology, Finland.

A. Dudin and K. De Turck (Eds.): ASMTA 2013, LNCS 7984, pp. 198–213, 2013.
© Springer-Verlag Berlin Heidelberg 2013

the gap between the user's need for higher data rates and the battery lifetime restrictions of small-scale user equipment grows considerably [5].

To address this gap, wireless industry reacts with a selection of solutions ranging from device battery innovation to advanced network architecture [6]. The latter suggests the use of small cells to augment the conventional cellular layout. Such integrated multi-tier deployments offer decisive benefits to the indoor user's connectivity, as well as in the areas with limited cellular coverage [7]. For example, a user in a two-tier in-building network (see Figure 1), e.g. in a shopping mall or in an office building, may receive improved service from the infrastructure low-power nodes (LPNs). However, when traveling from a small cell to a small cell, this user may also suffer from extra signaling when selecting the best LPN to transmit to. Furthermore, frequent cell re-selections may lead to excessive power consumption and drain the device battery. This problem receives increasing attention from wireless community [8], which recognizes the need for improved device power management mechanisms that would explicitly target small cell deployments.

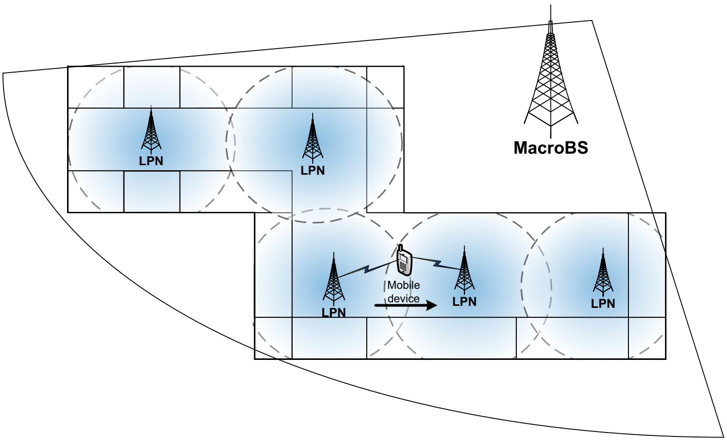


Fig. 1. Example topology of a multi-tier network

Whereas there has been much work on power control schemes for conventional cellular networks [9], it becomes crucial to address and account for the specific features of multi-tier networks. We believe that by intelligently allocating power on the available communication channels, a mobile user may considerably reduce its power consumption, while not compromising its desired quality of service. In this work, we propose a candidate power control strategy suitable for multi-tier wireless networks. We seek to maximize the energy efficiency of a user device to reach good balance between the required data rate and the resulting energy consumption.

More specifically, we consider a single user device which may *simultaneously* transmit its data to several neighboring LPNs centered at the surrounding small cells (as may be possible in future beyond-4G networks). Our goal is to advise this device on the optimal transmit power levels for each of the available LPN connections (which we term channels). We choose energy-efficiency as the optimization criterion, which is given as the relation between the data rate and the corresponding power expenses. Furthermore, we account for several important practical constraints of mobile user operation, such as the minimum required bitrate and the maximum allowed transmit power, which directly leads to an inequality-constrained optimization problem.

The energy-efficiency optimization at hand is based on the relationship between the transmit power and the resulting data rate. Instead of the actually achievable data rate, we study its theoretical maximum, that is, the capacity of a communication system. Here, the Shannon's capacity formula [10] for the channel with additive white Gaussian noise is the most widely used and popular approach. It helps abstract away the specific transmitter and receiver structures and, therefore, can be applied to most contemporary wireless networks, such as UMTS HSPA/HSPA+, WiMAX, 3GPP LTE/LTE-A, etc. In what follows, we choose 3GPP LTE-A (Long Term Evolution Advanced) as our example 4G technology [11] and demonstrate that its performance is reasonably close to the Shannon's limit, so that our capacity approximation is very precise. However, our solution is also applicable for alternative power-rate functions.

Furthermore, our approximation may be improved by generalizing the Shannon's formula through introducing several empirical factors. Then, these additional factors have to be calibrated, i.e. determined from the simulation results. Hence, in the second part of this paper, a detailed LTE-A link-level simulator is described and then used for adjusting the empirical parameters. Finally, numerical results are provided to conclude on the benefits of our energy-efficiency centric power control scheme by comparing its performance against two intuitive (heuristic) power control disciplines.

Our system model and the analytical solution to the constrained energy-efficiency optimization problem are presented in Section 2. We then augment this solution with extensive link-level simulation results. The used simulator mimics a realistic LTE-A Release-10 deployment and is detailed in Section 3. Further, the numerical results for the proposed power control scheme and the competitor heuristic strategies are given in Section 4. Finally, Section 5 concludes this paper.

2 Energy-Efficiency Optimization Problem

2.1 System Model and Assumptions

We consider an uplink data transmission of a single user device in a multi-tier wireless cellular network and study its achievable data rate, power, and energy-efficiency. This device may simultaneously use up to K available communication channels to the neighboring LPNs. In our model, every channel $i = \overline{1, K}$ may have different properties, and we only assume that the channels are mutually non-interfering. This may correspond to the practical scenario when the adjacent small cells are allocated non-overlapping radio frequencies, which is often the case in real-world deployments.

We also assume that the application-layer traffic of the considered user is saturated. The achievable data rate on the channel i is determined by the properties of the channel and by how much power is allocated by the user to transmit on it. The total user data rate can thus be obtained by aggregating the individual rates $r = \sum_{i=1}^{K} r_i$, where r_i is the data rate on the channel i. We impose a constraint on the total data rate r such that it must not drop below the minimum bitrate requirement r_0 given by e.g. a particular mobile application.

The total power consumption of the user device equals $p = p^{tx}(\mathbf{r}) + p^c = p^{tx}(\mathbf{r}) + \sum_{i=1}^{K} p_i^c$, where $p^{tx}(\mathbf{r})$ is the transmit power which is determined by a particular vector $\mathbf{r} = (r_1, ..., r_K)$ and p_i^c is the constant circuit power component (incurred by the active electronic circuitry) for the channel i. Further, we assume that the transmit power $p^{tx}(\mathbf{r})$ can also be aggregated over the individual powers $p^{tx}(\mathbf{r}) = \sum_{i=1}^{K} p_i^{tx}(r_i)$.

We introduce a variation of the Shannon's capacity formula, which would give us relationship between the transmit power and the maximum achievable data rate as:

$$p_i^{tx} = A_i(2^{B_i c_i} - 1), \tag{1}$$

where p_i^{tx} is the transmit power on the channel i, c_i is the theoretical capacity, while A_i and B_i are the additional parameters depending on wireless system implementation and configuration (including signal transmission mode, implementation-specific parameters, etc.). These can be given as:

$$B_i = \frac{1}{w_i}, \quad A_i = \frac{N_i}{g_i}, \tag{2}$$

where w_i is the channel bandwidth, $g_i = \rho_i/PL$ is the corresponding power gain, PL is path loss, ρ_i is the antenna gain and N_i is the total noise power over the given bandwidth. The coefficient B_i can also account for the overhead of the pilot signals, cyclic prefixes, and control channels occupying a portion of the system resources: $B_i = \frac{T_{total}}{w_i T_{data}}$, where T_{total} is the total amount of orthogonal (time-frequency) resources and T_{data} is the amount of resources allocated for the data transmission, w_i is the pure data channel (so-called PUSCH channel in LTE-A) bandwidth without guard bands and control channels.

The Shannon's formula relates the transmit power to the theoretical capacity, which may in reality differ from the actually achievable data rates. This may be due to the limited user knowledge about the wireless environment and thus non-optimal selection of modulation and coding schemes. In order to bring the relation (1) closer to the actual LTE-A system performance, we also introduce additional empirical factors α and β as:

$$B_i = \beta \frac{T_{total}}{w_i T_{data}}, \quad A_i = \alpha \frac{N_i}{g_i}. \tag{3}$$

Accounting for the above factors, we formally replace c_i by r_i, p_i^{tx} by p_i in (1) to summarize the considered relationship between the transmit power and the achievable data rate as:

$$p_i = A_i(2^{B_i r_i} - 1). \tag{4}$$

We name expression (4) the generalized Shannon's formula and give it as:

$$r_i = \frac{1}{B_i} log_2 \left(1 + p_i \frac{1}{A_i}\right). \tag{5}$$

We also note that the considered power-rate function (5) is bijective and monotonic with a continuous derivative, which will be used by the further analysis.

2.2 Optimization Problem

Our goal of energy-efficiency maximization can be achieved with the appropriate power control. Below, we formulate the constrained optimization problem where the argument is the achievable data rate on each available communication channel (which is equivalent to the corresponding transmit power).

We define and further optimize the energy efficiency of a user as the ratio between the total data rate r and the total power p:

$$\eta(\mathbf{r}) = \frac{r}{p} = \frac{\sum_{i=1}^{K} r_i}{\sum_{i=1}^{K} p_i(r_i) + \sum_{i=1}^{K} p_i^c}. \tag{6}$$

Further, we formulate our energy efficiency $\eta(\mathbf{r})$ optimization problem:

$$\max_{\{r_i\}_{i=1}^K} \eta(\mathbf{r}) = \max_{\{r_i\}_{i=1}^K} \frac{\sum_{i=1}^{K} r_i}{\sum_{i=1}^{K} p_i(r_i) + \sum_{i=1}^{K} p_i^c}, \text{ subject to:} \tag{7}$$

$$r = \sum_{i=1}^{K} r_i \geq r_0, \tag{8}$$

where r_0 is the minimum required bitrate. We also impose a reasonable constraint on the achievable data rate r_i (and hence, p_i), so that it cannot be negative:

$$r_i \geq 0, i = \overline{1, K}, \tag{9}$$

and, finally, account for the maximum allowed transmit power limit as:

$$p_i(r_i) \leq p_i^{max}, i = \overline{1, K}.$$

Note that since the function $p_i(r_i)$ is bijective, the above can be written as:

$$r_i \leq r_i^{max}, i = \overline{1, K}, \tag{10}$$

where $r_i^{max} = r_i(p_i^{max})$ can be calculated from (5).

For LTE-A system, the maximum throughput value can be expressed as:

$$\tilde{r}_i^{\max} = w_i \frac{T_{data}}{T_{total}} N_{bits}, \tag{11}$$

where N_{bits} is number of bits per symbol for the highest supported modulation.

Therefore, the maximum data rate should actually be calculated as $r_i^{max} = min(\tilde{r}_i^{\max}, r_i(p_i^{max}))$. We also take into account that $\eta(r_1, ..., r_K)$ is a non-zero function bounded on the interval $[\mathbf{0}, \infty)$, where $\mathbf{0}$ is a zero vector $(0, ..., 0)$. Now considering an equivalent form of (7) and rearranging the above inequalities (8), (9), as well as (10), our optimization problem can be summarized as:

$$\min_{\{r_i\}_{i=1}^K} U(\mathbf{r}) = \min_{\{r_i\}_{i=1}^K} \frac{1}{\eta(\mathbf{r})} = \min_{\{r_i\}_{i=1}^K} \frac{\sum_{i=1}^K p_i + \sum_{i=1}^K p_i^c}{\sum_{i=1}^K r_i} \tag{12}$$

subject to the constraints:

$$\phi(\mathbf{r}) = r_0 - \sum_{i=0}^{K} r_i \leq 0, \tag{13}$$

$$f_i(r_i) = -r_i \leq 0, i = \overline{1, K}, \tag{14}$$

$$g_i(r_i) = r_i - r_i^{max} \leq 0, i = \overline{1, K}. \tag{15}$$

2.3 General Way of Solving the Optimization Problem

The objective function given by (12)–(15) constitutes an inequality-constrained optimization problem. The general way of solving such optimization problem may be described by applying the Karush-Kuhn-Tucker (KKT) approach [12]. Accordingly, a system of equations and inequalities can be set up, known as regularity KKT conditions. For our optimization problem, the regularity KKT conditions are given as:

$$\frac{\partial U(\mathbf{r})}{\partial r_i} + \sum_{i=1}^{K} \lambda_i \frac{dg_i(r_i)}{dr_i} + \sum_{i=1}^{K} \mu_i \frac{df_i(r_i)}{dr_i} + \gamma \frac{d\phi(\mathbf{r})}{dr_i} = 0 \Leftrightarrow$$

$$\Leftrightarrow \frac{\frac{dp_i}{dr_i} \cdot r - (\sum_{i=1}^K p_i + \sum_{i=1}^K p_i^c)}{r^2} + \lambda_i - \mu_i - \gamma = 0, \quad i = \overline{1, K}$$

$$\lambda_i (r_i - r_i^{max}) = 0, r_i - r_i^{max} \leq 0, \lambda_i \geq 0, i = \overline{1, K}, \tag{16}$$

$$\gamma \left(\sum_{i=1}^{K} r_i - r_0 \right) = 0, \sum_{i=1}^{K} r_i - r_0 > 0, \gamma \geq 0,$$

$$\mu_i r_i = 0, r_i \geq 0, \mu_i \geq 0, i = \overline{1, K},$$

where λ_i, μ_i and γ are KKT multipliers.

Thus, in order to establish the optimal solution to the considered constrained optimization problem, the system of $3K + 1$ equations under $4K + 2$ inequalities has to be solved. We note that the search domain bounded by these inequalities has to be non-empty. Otherwise, the entire problem does not have a solution.

Noteworthy, the KKT conditions by themselves do not provide a method of finding the maximum/minimum points. Instead, they only determine the stationary points (where the gradient is zero) among which the minimum point can be found. In general, solving the system of equations and inequalities is known to be difficult. Therefore, instead of doing that, we apply a particular approach to solve the target optimization problem.

2.4 Particular Solution to the Optimization Problem

The solution to the system of equations:

$$\frac{\partial U(\mathbf{r})}{\partial r_i} = 0, i = \overline{1, K} \tag{17}$$

determines the optimum according to (12) without any inequality constraints. We begin with finding the stationary points for this unconstrained optimization problem in the domain $(-\infty, \mathbf{0}) \cup (\mathbf{0}, \infty)$. We then use this solution when dealing with the constrained optimization problem later.

Hence, we solve the target optimization problem in two steps

1. Solving the unconstrained optimization problem (12).
2. Updating the optimum by taking into account the constraints (13), (14), and (15) for each component of $\mathbf{r}$ step by step.

Optimal Solution without Constraints. In order to determine the stationary points of $U(\mathbf{r})$, we substitute the derivative $\frac{dp_i}{dr_i} = A_i B_i 2^{B_i r_i} \ln 2$ of function p_i (4) into (17). This results in:

$$\frac{\partial U(\mathbf{r})}{\partial r_i} = \frac{A_i B_i 2^{B_i r_i} \ln 2 \sum_{i=1}^{K} r_i - \left[\sum_{i=1}^{K} A_i (2^{B_i r_i} - 1) + p_c\right]}{\left(\sum_{i=1}^{K} r_i\right)^2} = 0, i = \overline{1, K}. \tag{18}$$

Therefore, we establish the following condition for the stationary points:

$$A_i B_i 2^{B_i r_i} \ln 2 \sum_{i=1}^{K} r_i - \sum_{i=1}^{K} A_i 2^{B_i r_i} + \left[\sum_{i=1}^{K} A_i - p_c\right] = 0, i = \overline{1, K}. \tag{19}$$

Rearranging (19) indicates that the term $A_i B_i 2^{B_i r_i} \ln 2$ does not depend on i:

$$A_i B_i 2^{B_i r_i} \ln 2 = \frac{\sum_{i=1}^{K} A_i 2^{B_i r_i} - \left[\sum_{i=1}^{K} A_i - p_c\right]}{\sum_{i=1}^{K} r_i}. \tag{20}$$

Further, we introduce the notation:

$$D = \frac{\sum_{i=1}^{K} A_i 2^{B_i r_i} - \left[\sum_{i=1}^{K} A_i - p_c\right]}{\sum_{i=1}^{K} r_i}. \tag{21}$$

Applying (20) and (21), the terms in (19) including the unknown values r_i can be expressed as:

$$A_i 2^{B_i r_i} = \frac{D}{B_i \ln 2}, \quad i = \overline{1, K}. \tag{22}$$

$$r_i = \frac{1}{B_i} \log_2 \frac{D}{A_i B_i \ln 2} = \frac{1}{B_i} [\log_2 D - \log_2 (A_i B_i \ln 2)], \quad i = \overline{1, K} \tag{23}$$

Applying (22) and (23) in (19) and rearranging leads to:

$$\frac{D}{\ln 2} \sum_{i=1}^{K} \frac{1}{B_i} - D \sum_{i=1}^{K} \frac{1}{B_i} \log_2 D + D \sum_{i=1}^{K} \frac{1}{B_i} \log_2 (A_i B_i \ln 2) = \left[\sum_{i=1}^{K} A_i - p_c \right].$$

We denote $\sum_{i=1}^{K} \frac{1}{B_i}$ and $\sum_{i=1}^{K} \frac{1}{B_i} \log_2 (A_i B_i \ln 2)$ as B and G respectively. Applying these notations, we obtain:

$$DB \frac{1}{\ln 2} - DB \log_2 D + DG = \left[\sum_{i=1}^{K} A_i - p_c \right]. \tag{24}$$

Rearranging (24) yields:

$$DB \left(\ln \left[D \cdot 2^{-\left(\frac{1}{\ln 2} + \frac{G}{B} \right)} \right] \right) = \ln 2 \left[p_c - \sum_{i=1}^{K} A_i \right].$$

Let us also denote $D \cdot 2^{-\left(\frac{1}{\ln 2} + \frac{G}{B} \right)}$ as X:

$$X (\ln X) = \ln 2 \frac{\left[p_c - \sum_{i=1}^{K} A_i \right]}{B} 2^{-\left(\frac{1}{\ln 2} + \frac{G}{B} \right)}. \tag{25}$$

From equation (25), we may obtain the value of X and, consequently, the expression for D:

$$D = 2^{\left(\frac{1}{\ln 2} + \frac{G}{B} \right)} \cdot \exp \left(W \left(\ln 2 \frac{\left[p_c - \sum_{i=1}^{K} A_i \right]}{B} 2^{-\left(\frac{1}{\ln 2} + \frac{G}{B} \right)} \right) \right), \tag{26}$$

where $W(x)$ is the Lambert's function [13].

Applying (26) together with the definitions of G and B in (23) leads to the stationary point:

$$r_i^* = \frac{1}{B_i} \left[\frac{1}{\ln 2} + \frac{\sum_{i=1}^{K} \frac{1}{B_i} \log_2 (A_i B_i \ln 2)}{\sum_{i=1}^{K} \frac{1}{B_i}} - \log_2 (A_i B_i \ln 2) \right] +$$

$$+ \frac{1}{B_i \ln 2} W \left(\ln 2 \frac{\left[p_c - \sum_{i=1}^{K} A_i \right]}{\sum_{i=1}^{K} \frac{1}{B_i}} 2^{-\left(\frac{1}{\ln 2} + \frac{\sum_{i=1}^{K} \frac{1}{B_i} \log_2 (A_i B_i \ln 2)}{\sum_{i=1}^{K} \frac{1}{B_i}} \right)} \right), \tag{27}$$

where A_i and B_i are given by (3). The power level to operate on a particular channel can finally be obtained by (4).

Optimization Under Constraints. Here we employ the solution $\mathbf{r}^* \in R^K$ of the unconstrained optimization problem (27) to the respective constrained problem in (12)–(15).

Several alternative cases are possible:

1. The argument of the function $W(x)$ in (26) is less than $-e^{-1}$, which means that the objective function of the unconstrained problem does not have a stationary point. Hence, we need to search the optimal point on the border of the domain by choosing the component r_i with the minimum contribution and setting this component to zero.
2. If $r_i^* \leq 0$ or $r_i^* \geq r_i^{max}$, then the stationary point lies outside the search domain and should be instead found on the respective plane $r_i^* = 0$ or $r_i^* = r_i^{max}$. If either of these conditions holds for several indexes, we need to choose the component r_i with the minimum contribution and set $r_i^* = 0$ or $r_i^* = r_i^{max}$.
3. If $\sum_{i=1}^{K} r_i^* < r_0$, then the optimal point lies on the plane $\sum_{i=1}^{K} r_i^* = r_0$ and we need to follow the above steps again.

Having fixed one of the components, we proceed by solving the respective optimization problem of dimension $K-1$. It can be done in a similar way as solving the original problem $\mathbf{r}^*$ (we omit the details here due to the space constraints). The above steps are to be repeated until the set of components $\mathbf{r}^*$ is obtained which satisfy the given constraints.

3 Simulation Methodology

3.1 Description of the Simulator

After our energy-efficient power control scheme has been introduced, we aim to augment our solution with simulation results derived from a detailed model of 3GPP LTE-A system. Below we shortly describe the considered link-level simulator (LLS). In particular, we seek to apply the LLS tool for calibrating the empirical coefficients α and β in (3). Our approach is realistic modeling of (i) all the necessary transmitter operations of LTE-A Release-10 uplink, (ii) the radio channel with additive white Gaussian noise (AWGN), and (iii) all the corresponding receiver operations. The general structure of the LLS is presented in Figure 2.

A fragment of data in the form of a transport block (TB) is generated of random bits. It is then fed to the input of the transmitter part and passed through the stages of turbo encoding, rate matching [14], scrambling, and QAM mapping [15]. Turbo encoding with a fixed code rate of 1/3 is performed according to the LTE-A standard. The rate matching stage performs bit puncturing or repetition coding to match the original code rate of 1/3 to an arbitrary code rate. Puncturing and repetition patterns are implemented in full compliance with the LTE-A specifications. The scrambler performs modulus two addition of rate-matched bits with a random sequence specified by the standard, whereas

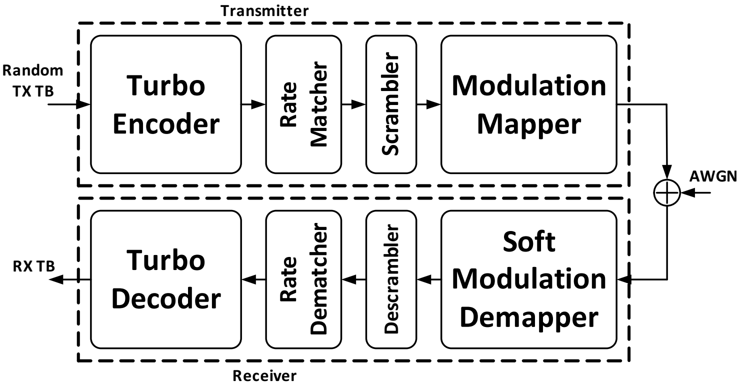


Fig. 2. Structure of the LTE-A link-level simulator

the mapper generates QPSK, 16-QAM, or 64-QAM constellations. An arbitrary modulation and coding scheme (MCS), specifying a pair of code rate and modulation type, is set as a simulation parameter and is fixed during a particular simulation run.

The AWGN channel performs addition of constellations (QAM symbols) at the output of the transmitter with randomly generated AWGN. The target is to reach the required signal-to-noise ratio (SNR) set as a simulation parameter.

Noisy constellations are fed to the input of the receiver part and pass through the soft demapping, descrambling, rate dematching, and turbo decoding stages. The soft demapper calculates logarithms of likelihood ratio using the max-log approximation. The descrambler performs the inverse transform as opposed to the scrambler, while rate dematcher performs addition of likelihood ratios corresponding to the repeated bits and/or taking the ratios corresponding to the punctured bits equal to zero. Turbo decoder realizes max-log maximum a posteriori decoding algorithm to derive a received TB with, possibly, some erroneous bits. Then, this TB is compared with the transmitted TB to detect errors and to determine the correctness of TB reception.

3.2 Experiment Plan and Results

The transmission of a particular TB is repeated multiple times for a fixed set of parameters to perform statistical averaging over the random realizations of a TB and AWGN. After the simulation is completed, the collected statistics shows the dependence of transport block error probability (BLER) versus SNR for a fixed MCS. Then, similar simulation is repeated for a set of available MCSs.

Each of the MCSs determines a possible data rate in the LTE-A system. It can be calculated as a transport block size (TBS) divided by the subframe length and multiplied by the probability of successful TB transmission:

$$r_k = (1 - BLER_{target})\,\frac{TBS_k}{T_{subframe}}, \quad SNR_k = SNR^{(k)}\,(BLER_{target})\,, \tag{28}$$

where r_k is the data rate corresponding to the k-th MCS, k is the MCS index, TBS_k is the TBS of the k-th MCS, $BLER_{target}$ is the target BLER, $T_{subframe}$ is the subframe length (equal to 1 ms in LTE-A), SNR_k is the SNR required for the k-th MCS, $SNR^{(k)}(BLER)$ is the inverted dependency of BLER on SNR for the k-th MCS.

As the result of our LTE-A simulations, in Figure 3 we demonstrate the dependency of the achievable data rate on the SNR level for different MCSs. We also compare the simulation values with the Shannon's capacity formula to conclude that it serves as a reasonable approximation of the practical data rate.

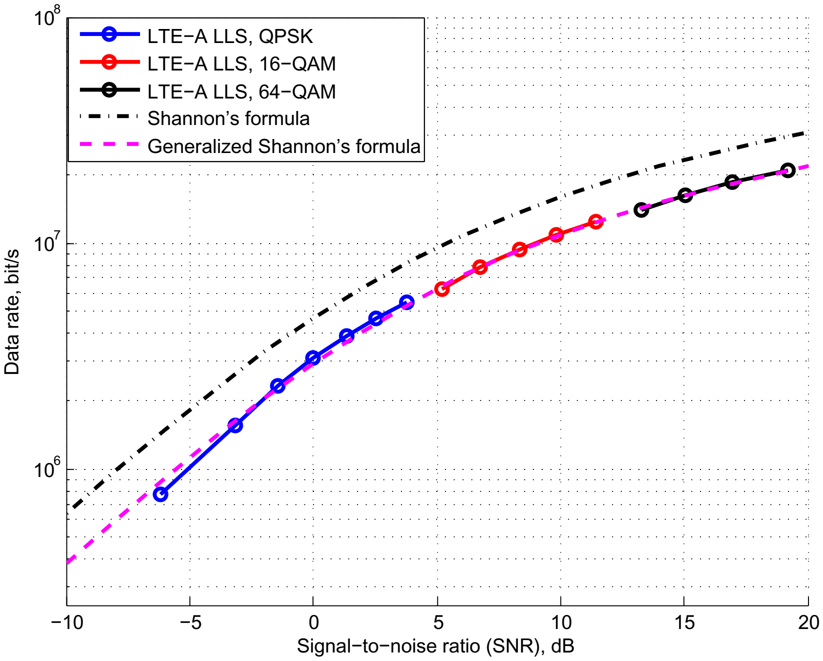


Fig. 3. Comparison of the Shannon's formula ($\alpha = 1$, $\beta = 1$) against simulation results for QPSK, 16-QAM, and 64-QAM

3.3 Calibrating the Generalized Shannon's Formula

Here we use our simulation results to calibrate the coefficients of the generalized Shannon's formula (4). Figure 3 provides dependencies corresponding to the conventional Shannon's formula ($\alpha = 1$ and $\beta = 1$), the generalized Shannon's formula (fitted coefficients $\alpha = 1.2456$ and $\beta = 1.3463$), and the simulated data rates for QPSK, 16-QAM, and 64-QAM modulation schemes. In order to properly adjust the coefficients α and β, the three sets of points corresponding to different modulations may first be converted into a single curve.

Our calibration procedure consists in the optimal selection (by e.g. least-squares technique) of the coefficients α and β in order to minimize the discrepancy between the available simulation results and the formula:

$$r_i = w \frac{T_{data}}{T_{total}} \frac{1}{\beta} \log_2 \left(1 + \frac{SNR_i}{\alpha}\right), \tag{29}$$

where $SNR = \frac{p_i g_i}{N_i}$.

4 Numerical Results

4.1 Setup Details and Parameters

In this section, we concentrate on an illustrative numerical example to evaluate the performance of the proposed power control scheme. We assume that there are $K = 2$ communication channels available to the mobile user. Correspondingly, the first recipient LPN is located at the point $x_1 = 0$, whereas the second one is at the point $x_2 = R$ (see Figure 4). The user is assumed to move all along the x-axis between the two LPNs, and its current coordinate is $x \in [0, R]$.

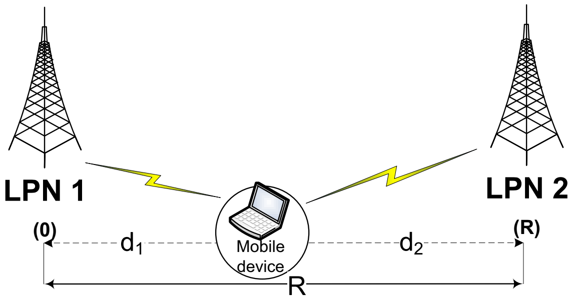


Fig. 4. Topology of our numerical setup

In order to compare our energy-efficient power control with alternative power management techniques, we introduce two simple and intuitive strategies.

1. The user transmits on both channels simultaneously by allocating a fixed amount of power to every channel.
2. The user transmits on one channel by selecting it basing on the channel quality and allocating a fixed amount of power to the best channel only.

In both cases, for the sake of simplicity, we assume that the allocated power level is equal the maximum allowed power (see Table 1). We also assume that the channels are symmetric, i.e. employ similar parameters, including the following propagation model [16]:

$$PL = 22.0 \log_{10} d + 28.0 + 20 \log_{10} f_c, 10 < d < d_{BP}, \tag{30}$$

$$PL = 40\log_{10} d + 7.8 - 18\log_{10} h'_{LPN} - 18\log_{10} h'_{user} + 2\log_{10} f_c, d_{BP} < d < 5000,$$

where PL is the path loss (dB), h'_{LPN} and h'_{user} are effective antenna heights (m), d and d_{BP} are the distance to the LPN and the break point distance (m) respectively, and f_c is the center frequency (GHz). See Table 1 for more details.

Table 1. Summary of simulation assumptions

Parameter	Value
Simulation approach	Link level simulations of LTE-A Release 10
Radio channel model	AWGN
Channel estimation and synchronization	Ideal
PUSCH bandwidth	5.4 MHz (30 resource blocks per slot)
Cyclic prefix	Normal
Turbo decoder	Max-log turbo decoder with 8 iterations
Target block error rate (BLER)	10%
Thermal noise power	-103 dB
Carrier frequency	2 GHz
User antenna height	1.5 m
LPN antenna height	10 m
Environment	Micro cell in urban area
Maximum transmit power	23 dBm
Circuit power	0.1 W
Idle power	0.01 W
Antenna gain	3 dB
Minimum data rate	1.19 Mbps
Number of bits per QAM symbol for the maximum modulation order	6

4.2 Discussion of the Results

In Figure 5, we overlay the results for the total achievable data rate, energy efficiency, and power consumption depending on the user location x. With our optimal power control, the user begins with transmitting on one channel (A). As it moves toward the center (B), the user adjusts its transmit power to compensate for the varying pathloss value [16]. The shape of the transmit power function here is determined solely by the pathloss alterations in (30).

Moving further, the user does not yet need to apply the maximum power to reach the highest energy efficiency until the point (C), when the constraint r_0 takes effect. Because of this bitrate constraint, the power rises dramatically up to the point (D), when the use of the second channel becomes reasonable. Then power gradually grows up to the maximum transmit power level (E) to stay there until (F). Further, this behavior mirrors symmetrically (as both channels are equivalent).

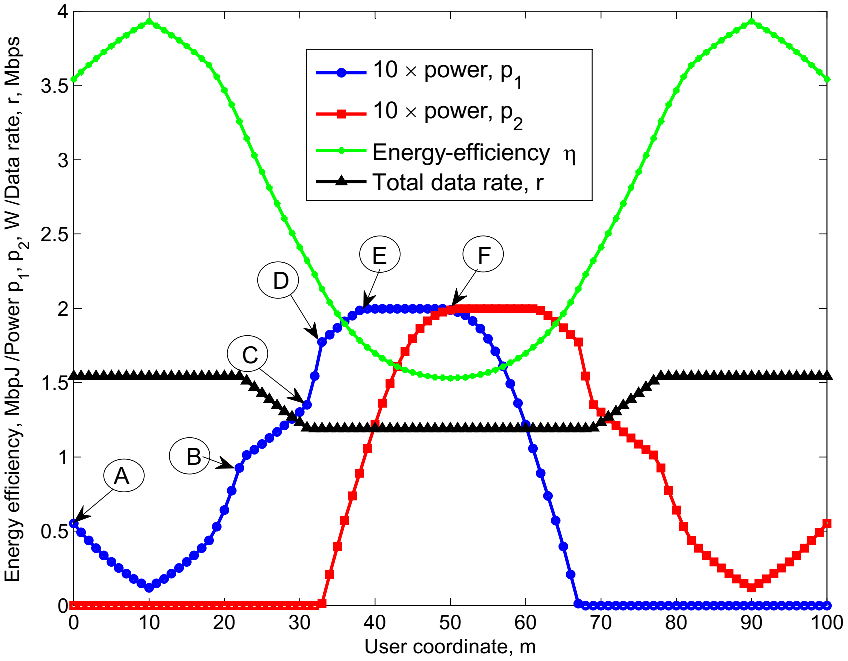


Fig. 5. Achievable data rate, transmit power, and energy efficiency of the mobile user

Table 2. Energy efficiency and total data rate comparison for base schemes, $K = 2$

Distance, m	r^*, Mbps	r_1, Mbps	r_2, Mbps	η^*, MbpJ	η_1, MbpJ	η_2, MbpJ	$\Delta\eta_1$, %	$\Delta\eta_2$, %
0	1.54	1.58	1.54	3.54	2.49	3.54	30	0
10	1.54	1.61	1.54	3.93	2.71	3.93	31	0
20	1.54	1.65	1.54	3.47	2.57	3.47	26	0
30	1.23	1.57	1.36	2.41	2.01	2.35	17	2
40	1.19	1.29	0.93	1.7	1.65	1.61	3	5
50	1.19	1.19	0.6	1.53	1.53	1.03	0	33
60	1.19	1.29	0.93	1.7	1.65	1.61	3	5
70	1.23	1.57	1.36	2.41	2.01	2.35	17	2
80	1.54	1.65	1.54	3.47	2.57	3.47	26	0
90	1.54	1.61	1.54	3.93	2.71	3.93	31	0
100	1.54	1.58	1.54	3.54	2.49	3.54	30	0

The data rate plot clearly demonstrates the upper and the lower regions of better and worse channel quality, respectively. Consequently, the lowest energy efficiency is reached in the central point, where the quality of both channels is the poorest, whereas the maximum is reached at the distance of obstruction, when channel quality is the best according to (30).

Note that our power control scheme can be implemented basing on a simple lookup table of the Lambert's function thus requiring negligible time/energy expenses when solving the discussed optimization problem.

In Table 2, the energy efficiency η^* and data rate r^* of our scheme are compared against those of two intuitive (heuristic) strategies (η_1, r_1 for the first and η_2, r_2 for the second, respectively). Clearly, the one-channel (second) strategy fails to satisfy the minimum bitrate requirement in the central region. Furthermore, our approach allows to reach the performance of the two-channel (first) strategy, but results in lower power consumption when the target bitrate r_0 is already met. The relative increase in energy efficiency with our power control ($\Delta\eta_1$, $\Delta\eta_2$) is also given in Table 2.

5 Conclusion

In this work, we have considered the problem of energy efficient power control when the mobile user may communicate on several uplink wireless channels at the same time. We have also calibrated our analytical solution with the detailed link-level LTE-A simulations. Our numerical example for two symmetric channels suggests the gain of up to 30 % when prefering our power management strategy to simpler heuristic mechanisms. Our current work is the comparison of the proposed energy efficient approach against more sophisticated power control strategies and the consideration of more practical device power models.

Acknowledgment. This work is supported by the Internet of Things program of Tivit, funded by Tekes, and Hungarian Government through the TÁMOP-4.2.2C-11/1/KONV-2012-0001 and the OTKA K101150 projects.

References

1. Ericsson. More than 50 billion connected devices (2011)
2. Cisco. Global Mobile Data Traffic Forecast Update, 2011-2016 (2012)
3. Akyildiz, I., Gutierrez-Estevez, D., Reyes, E.: The evolution to 4G cellular systems: LTE-Advanced. Physical Communication 3, 217–244 (2010)
4. Pentikousis, K.: In search of energy-efficient mobile networking. IEEE Communications Magazine 48, 95–103 (2010)
5. Tombaz, S., Vastberg, A., Zander, J.: Energy- and cost-efficient ultra-high capacity wireless access. IEEE Wireless Communications 18, 18–24 (2011)
6. Raychaudhuri, D., Mandayam, N.: Frontiers of wireless and mobile communications. Proceedings of the IEEE 100, 824–840 (2012)
7. Hu, R., Talwar, S., Zong, P.: Cooperative, Green and Mobile Heterogeneous Wireless Networks. IEEE 802.16x (2011)
8. R2-130478. Mobility in HetNet scenarios. Nokia Siemens Networks (2013)
9. Miao, G., Himayat, N., Li, G.: Energy-efficient link adaptation in frequency-selective channels. IEEE Transactions on Communications 58, 545–554 (2010)
10. Rappaport, T.: Wireless Communications: Principles and Practice. Pearson Education (2009)

11. 3GPP LTE Release 10 & beyond (LTE-Advanced)
12. Kuhn, H., Tucker, A.: Nonlinear programming. In: Proc. of 2nd Berkeley Symposium (1951)
13. Corless, R., Gonnet, G., Hare, D., Jeffrey, D.: On Lambert's *W* function. Technical report, University of Waterloo (1993)
14. 3GPP TS 36.212. Multiplexing and channel coding (2012)
15. 3GPP TS 36.211. Physical Channels and Modulation (2012)
16. 3GPP TR 36.814. Further advancements for E-UTRA physical layer aspects (2010)

Publication 3

© 2013 Springer Science + Business Media. Reprinted, with permission, from

O. Galinina, A. Anisimov, S. Andreev, and Y. Koucheryavy, "Performance Analysis of Uplink Coordinated Multi-Point Reception in Heterogeneous LTE Deployment", in *Proc. of the 11th International Conference on Wired/Wireless Internet Communications (WWIC)*, 2013.

Performance Analysis of Uplink Coordinated Multi-Point Reception in Heterogeneous LTE Deployment

Olga Galinina[1], Alexey Anisimov[2], Sergey Andreev[1], and Yevgeni Koucheryavy[1]

[1] Tampere University of Technology, Finland
{olga.galinina,sergey.andreev}@tut.fi, yk@cs.tut.fi
[2] Nokia Siemens Networks, MBB LTE, Russia
alexey.anisimov@nsn.com

Abstract. In this work, we study a heterogeneous 3GPP LTE network where macro base station deployment is coupled with underlay low power (pico) nodes to augment system capacity. However, at the cell edges, a macro-associated user may still suffer from poor performance due to low uplink channel quality. This is when the neighboring low power nodes can help by independently trying to receive data packets from the macro user and share the result with the base station if successful. Known as coordinated multi-point (CoMP) reception, this scheme is expected to dramatically improve uplink cell-edge performance. To predict the actual gains, we conduct our analysis of a typical CoMP setup for dynamic traffic load and depending on the user proximity to the serving base station.

1 Introduction and Background

Recent advances in wireless communications introduce fundamental changes to mobile Internet access, as well as challenge the researchers with increasingly demanding problems. As long as the proportion of mobile traffic is expected to grow [1], the currently deployed cellular technologies are very likely to face dramatic overloads resulting in shortage of available capacity and general degradation of the user service experience. Reacting to this pressing demand, the fourth generation (4G) broadband communication standard [2] offers decisive improvements to the levels of achievable spectral and energy efficiency as well as quality of service. However, user performance may still remain unsatisfactory at the cell edges, where the connection to the serving base station is weak and the transmission is further limited by interference from the neighboring cells.

Conventionally, user service uniformity has been achieved with appropriate network planning, when specific frequency reuse patterns were employed to combat the inter-cell interference. This, however, often resulted in low spatial reuse factors and poor resource utilization [3]. A more advanced solution may be to enable collaborative transmission or reception by multiple network entities. Such approach

V. Tsaoussidis et al. (Eds.): WWIC 2013, LNCS 7889, pp. 1–14, 2013.
© Springer-Verlag Berlin Heidelberg 2013

is expected to naturally leverage the diversity gains between geographically separated points. In 3GPP Long Term Evolution (LTE) cellular technology, this technique is known as Coordinated Multi-Point (CoMP) and is believed to boost the system performance dramatically, especially at the edges of a cell.

The performance of CoMP in both uplink (UL) and downlink (DL) of conventional macro deployment has been thoroughly studied in the literature. Noteworthy, UL CoMP schemes tend to receive more research attention as they have less impact on the LTE specification [4]. Therefore, already in [5] the cell-edge benefits promised by UL CoMP have been quantified with system-level simulations. Whereas the first evaluation attempts focused on static full-buffer environments, more recent works [6], [7] employ dynamic processes and rigorously analyze the network design aspects of CoMP in terms of the required density of the serving base stations. In particular, the research in [6] suggests the use of *selection combining* CoMP scheme, when upon a reception failure the serving base station chooses the decoding outcome with the highest channel quality among the available alternative receivers of a particular data packet. Since only successful outcomes are exchanged, this approach is attractive due to moderate amounts of data transferred between the collaborating points.

Whereas much work has been dedicated to evaluating UL CoMP in the neighborhood of (macro) base stations of the same type and power class, the selection combining technique is expected to yield even higher gains across heterogeneous deployments. Heterogeneous networks are characterized by a mixture of macro base stations and low power (pico) nodes, which may generate excessive intercell interference. As interference coordination in such harsh environments can be complex, UL CoMP may prove to be very useful in the metropolitan areas with dense network deployment [8].

Offering a general classification of CoMP techniques, the work in [4] suggests an important use case when a macro-cell collaborates with several low power receive points within its coverage to better serve UL traffic by a macro user. By concentrating on its design principles and choices, the research in [8] also confirms that UL CoMP may render user experience more uniform in a similar heterogeneous scenario. Moreover, the outdoor measurements reported in [3] showcase the attractiveness of CoMP-based approaches while at the same time indicating several technical challenges, such as the need for high-capacity and low-latency interface (backhaul connection) between the serving points.

In particular, the performance of selection combining technique in a heterogeneous deployment has been investigated by [9] and [10] with full-buffer system-level simulations to exploit the pico-node proximity to the macro-associated users. Further, [11] considers a similar CoMP setup and takes advantage of the asymmetry between the UL and the DL with appropriate cooperation-aware power control to mitigate the near-far effect. Complementing prior simulation data, [12] reports field trial results of selection combining CoMP in dense heterogeneous networks to leverage the macro diversity gain around the cell edge.

By taking the idea of UL/DL asymmetry of CoMP further on, it is also possible to tailor the conventional handover procedures specifically to heterogeneous networks. With a suitable scheduling discipline, macro-cell traffic may be dynamically offloaded onto small cells [13] to provide seamless handover-like user experience. More broadly, handover procedure in cellular networks is an important aspect and may actually be improved by properly accounting for CoMP [14]. Last but not least, energy efficiency is becoming increasingly important for small-scale battery-powered mobile devices. Consequently, catering for the best trade-off between spectral- and energy-efficiency (such as bits-per-Joule capacity) is a crucial and timely problem [15]. The related analysis implies that UL CoMP results in higher cell-edge energy efficiency than a non-cooperative system. Finally, the practical aspects of DL CoMP design, including the impact of imperfect backhaul connections, have recently been addressed in [16].

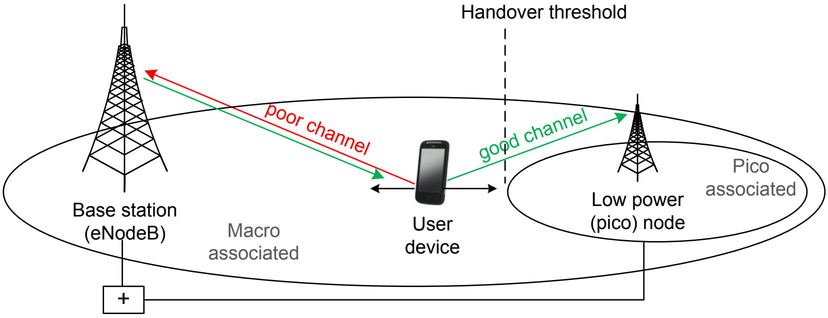


Fig. 1. Selection combining UL CoMP in a heterogeneous LTE network

Summarizing, the various aspects of the UL CoMP operation have indeed been addressed by the existing work, but the majority of the findings are disjoint due to the difference in adopted system models, assumptions, and methodologies. Furthermore, most papers evaluate CoMP performance with simulation, while analytical attempts are singular and only loosely connected with respective simulation assumptions. Therefore, we are motivated to propose a comprehensive CoMP-centric evaluation methodology by coupling both analytical and simulation components. While remaining simple, our closed-form analysis captures many important CoMP features, such as heterogeneous environment, user mobility, impact of power control and handover decisions, energy efficiency, as well as imperfect UL channel to the serving base station (see Figure 1). In particular, we evaluate the selection combining UL CoMP scheme, where the cooperating low power nodes are independently trying to receive data packets from a macro user and share their successful outcomes with the serving base station.

The rest of this text is organized as follows. Section 2 details our system model and the main assumptions. In Section 3, we introduce our analytical approach to calculate the key performance metrics, such as data packet service time (9), mean packet delay (14), packet drop probability (15), and the corresponding energy consumption (19). Section 4 contains some important numerical results, while Section 5 concludes the paper.

2 System Model

In this section, we summarize the assumptions of our system model. We consider a heterogeneous 3GPP LTE deployment consisting of one macro base station (BS) and $N-1$ neighboring low power nodes (LPNs). The user equipment (UE) is assumed to be constantly associated with the macro BS. Technically, it measures the quality of the DL signal from the serving base station and sets its transmit power as given by e.g. [9]:

$$P_{TX} = \min\{P_{max}, P_{TX,0} + 10\log_{10} N_r + \alpha L\}, \tag{1}$$

where P_{max} is the maximum transmit power, $P_{TX,0}$ is the target receive power, N_r is the number of resources assigned to the UE, L is the pathloss between the UE and its serving BS, and $0 \leq \alpha \leq 1$ is the pathloss compensation factor.

It is also assumed that the user is mobile, that is, it may change its location with respect to the serving BS. Whenever approaching any of the LPNs, the UE is supposed to make a handover decision. Accordingly, the effective macro-cell border is determined as:

$$\Phi_{BS} - L_{BS} = \Phi_{LPN} - L_{LPN}, \tag{2}$$

where $\Phi_{[\cdot]}$ is the transmit power of the BS/LPN (on the logarithmic scale), $L_{[\cdot]}$ is the pathloss between the BS/LPN and the UE.

Hence, we are interested in the UE performance at the macro cell-edges and before the macro-associated user has actually made its LPN handover decision. In these areas, the user channel quality to the BS is typically poor (see Figure 1). We thus analyze the benefits of UL CoMP selection combining scheme when the UE data packets are independently received by the proximal LPNs. All the collaborating LPNs are synchronized and connected to the macro BS via a high-capacity and low-latency interface (assumed to be instantaneous and error-free without the loss of generality), so that the successful packet reception outcomes by the LPNs can be immediately shared with the BS.

Further, we consider a Single-Carrier Frequency Division Multiple Access (SC-FDMA) system with Multi-level Quadrature Amplitude Modulation (M-QAM) in the presence of Additive White Gaussian Noise (AWGN). In the UL of every subframe, a certain number of resource blocks is provided for the target user by the BS where the user may transmit its equal-size packets.

The numbers of new data packets arriving at the UE during the consecutive subframes are i.i.d. random variables. For simplicity of further analysis, we assume Poisson arrival flow. Hence, the UE generates new data packets with the average arrival rate of λ packets per slot.

We further assume that the UE is equipped with K independent (virtual) buffers according to the LTE specification [2]. Upon its arrival at the UE, a packet is being placed into one of the K buffers with the constant probability $q = 1/K$. The data packets in a particular buffer i may only be attempted for transmission (served) in every ith subframe (slot) and the transmission of every

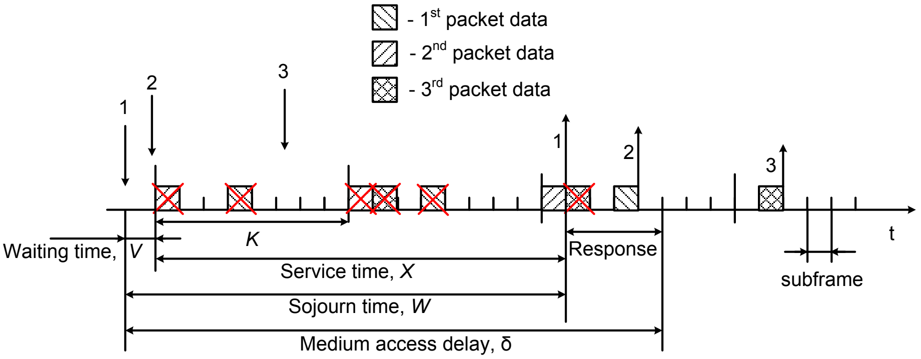


Fig. 2. Example time diagram of UL CoMP

packet takes exactly one slot (see Figure 2). All the buffers are assumed to have unlimited size.

A data packet may be received by the macro BS or an LPN with some constant probability. This probability depends on the UL Signal-to-Noise Ratio (SNR) and thus varies for different receive points. A data packet transmission is considered successful if the BS or at least one of the LPNs receives this packet. The chance of success therefore depends on the corresponding events at BS/LPNs.

Following Hybrid Automatic Repeat Request (HARQ) procedure, the BS forwards the per-packet positive (ACK) and negative (NACK) acknowledgments to the user after the fixed delay of τ subframes. In case of failure, the UE may retransmit its packet after exactly K subframes. The maximum number of allowed transmission attempts per packet equals n_{max}. If the last transmission attempt has been unsuccessful, the packet is dropped (discarded) by the user.

As we are also interested in the user energy consumption, we differentiate between the following UE power states (see Figure 3):

- Idle state. In this state, the UE's buffer is empty and the minimum power P_0 is consumed.
- Active state. The device is active and has at least one packet in any of the buffers. However, it does not transmit in the current subframe and the power P_1 is spent.
- Transmit (Tx) state. The device is transmitting its data with the power of $P_2 = P_{TX}$ as defined by (1). In this state, the maximum power $P_1 + P_2$ is consumed.

In what follows, we concentrate on the analytical modeling of the above system in order to investigate its primary performance metrics, such as packet success and drop probabilities, energy efficiency, and the expected packet delay.

3 Performance Evaluation

Given a particular BS/LPN deployment (e.g., by [17]), the system parameters of interest are primarily determined by the probability of successful data packet

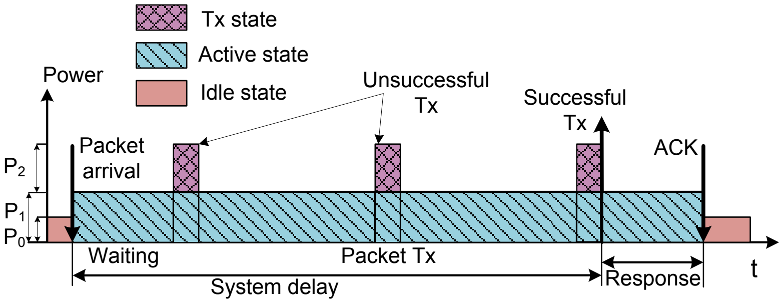


Fig. 3. Example UE power consumption diagram

transmission. With UL CoMP based on selection combining, this probability depends on respective individual probabilities to receive this data packet at the macro BS or a particular LPN. Below we detail our approach to the calculation of the sought probability basing on the UL SNR values at the receiving points.

3.1 Probability of Success

The Case of One Receiver. Consider the system with a single receiver (i.e., macro BS) of UE data. The corresponding value of Bit Error Rate (BER) can be determined from the UL SNR value, which depends on the transmit power and radio conditions at the BS. We employ the approximation from [18] for uncoded M-QAM and replace the Q-function by elementary functions, which makes our interpretation more tractable analytically.

Hence, for a particular UE-BS distance, the actual modulation order in M-QAM may vary to mimic the process of user rate adaptation. Assuming an ideal coherent phase detection in the AWGN channel n, the value of BER is well approximated by (see [18]):

$$p_n = \frac{4}{\log_2 M} Q\left(\sqrt{\frac{3\gamma_n}{M-1}}\right), \tag{3}$$

where γ_n is the SNR in the channel n, $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{1}{2}t^2} dt = 1 - \Phi(x)$, $\Phi(x)$ is an error function which is the integral of the standard normal distribution [19].

For the sake of analytical tractability, we propose to modify the above approximation by using solely elementary functions. We note that $Q(x) = \frac{1}{2}\mathrm{erfc}\left(\frac{1}{\sqrt{2}x}\right)$, $x \geq 0$. Here, erfc $= 1-$erf is the complementary error function and erf is the error function. Further, we borrow the approximation for the error function from [20]:

$$\mathrm{erf}(x) \approx \mathrm{sgn}(x)\sqrt{1 - \exp\left(-x^2 \frac{4/\pi + ax^2}{1 + ax^2}\right)}. \tag{4}$$

where $a = \frac{8(\pi-3)}{3\pi(4-\pi)} \approx 0.14$.

Eliminating the unnecessary components by simple transformation, we arrive at the following:

$$\mathrm{erf}(x) \approx \mathrm{sgn}(x)\sqrt{1-\exp\left(-\frac{x^2}{2}\right)}. \tag{5}$$

Finally, in Figure 4 we compare our proposed BER approximation against its alternatives from [18] (based on Q-function and other elementary functions) and intend to use it in what follows.

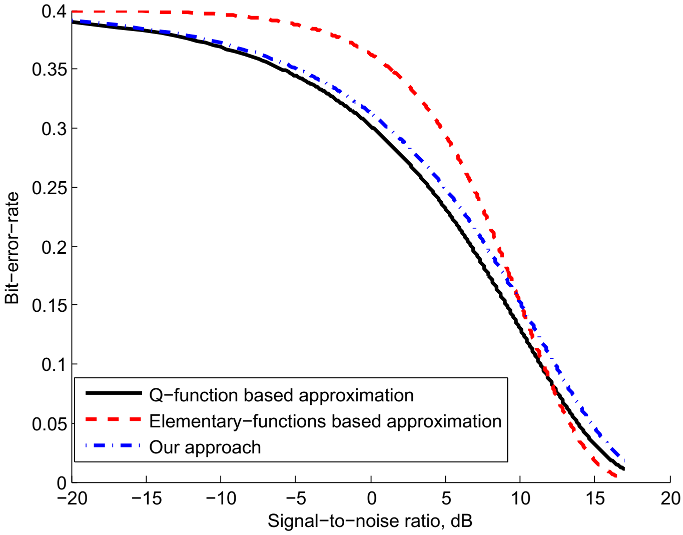


Fig. 4. Example BER approximation for 32-QAM

Therefore, the probability of successful packet reception at the BS follows as:

$$p_n = \frac{2}{\log_2(M)}\left[1-(1-e^{-\frac{3\gamma_n}{2(M-1)}})^{0.5}\right]. \tag{6}$$

The Case of Multiple Receivers. Consider the system with several receivers (i.e., macro BS and LPNs) of UE data. The total probability of successful reception by at least one receiving point can be easily calculated accounting for the independence of individual success events with the probabilities p_n, $n = \overline{1, N}$:

$$p = Pr\{\text{at least one receiver succeeds}\} =$$

$$= \prod_{n=1}^{N} p_n - \sum_{i=1}^{N}\prod_{n=1,n\neq i}^{N} p_n + \sum_{i=1}^{N}\sum_{j=1}^{N}\prod_{n=1,n\neq i,j}^{N} p_n + ... - (-1)^N \sum_{i=1}^{N} p_n. \tag{7}$$

Hence, we derive an expression for the packet success probability p, agnostic to a particular BS/LPN deployment, which scales with the actual number of neighboring LPNs. We proceed further with our queuing model.

3.2 Proposed Queuing Model

We note that for the adopted system model (see Section 2), all K virtual buffers are served independently. Therefore, we decouple system operation into K independent First-Come-First-Served (FCFS) queues. The arrival process at a particular queue i is the result of splitting (thinning) the initial Poisson process. Hence, it also constitutes a Poisson process with the arrival rate of $\lambda_i = \lambda/K$.

Consequently, we formulate an M/G/1 model (see Figure 5), i.e. consider a stochastic process $N_i(t)$ on the state space $\{0, 1, 2, ...\}$, where $N_i(t)$ is the number of packets in the queue i at the end of a subframe. Arrivals happen according to a Poisson process with the arrival rate $\lambda_i = \lambda/K$. Service times are i.i.d. random variables which will be addressed below. The queue size is unbounded.

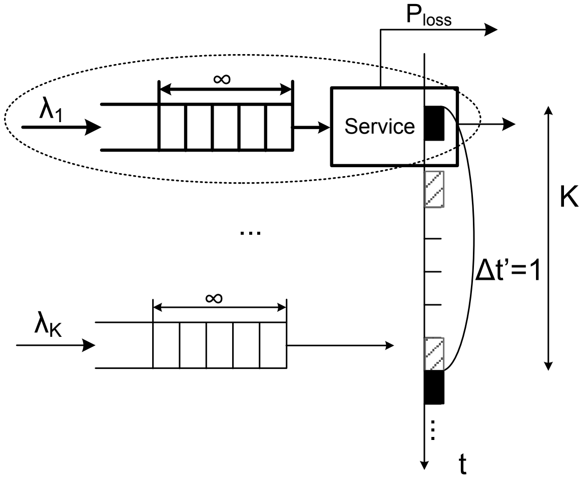


Fig. 5. Aggregated queuing system

We denote the random variable representing the packet service time as X_i, which is the time interval between the first attempt to serve this packet and the packet departure time (success or drop). Further, the total time interval between the arrival of a packet and the time its service begins is denoted as $W_i = V_i + X_i$, where V_i is the waiting time in the queue. Importantly, V_i includes (i) the time of waiting between the arrival and the moment when the queue i obtains opportunity to be served ($V_i^{(1)}$) and (ii) waiting time as long as the preceding packets in the queue i are being served ($V_i^{(2)}$).

Finally, we denote the queue load as $\rho_i = \lambda_i E[X_i]$ and assume that $\rho_i < 1$, so that the steady-state distribution exists. Below, we concentrate on the expected time $E[W_i]$ that a packet spends in our system. For the sake of simplicity, we scale our system up and aggregate K subframes into one time unit $\Delta t' = 1$ to calculate some auxiliary variables with respect to the aggregate time.

3.3 Service Time Distribution and Packet Drop Probability

The random variable representing service time X_i is distributed according to the truncated geometric distribution:

$$\Pr\{X_i = 1\} = p,$$

$$\Pr\{X_i = 2\} = p(1-p), ...,$$

$$\Pr\{X_i = n_{\max}\} = p(1-p)^{n_{\max}-1} + (1-p)^{n_{\max}},$$

where p is the (integral) probability of successful packet transmission, i.e. when at least one receiver acquires the packet, see (7). We note that here we have already taken into account the effect of time aggregation.

The expected service time is defined as:

$$E[X_i] = \sum_{i=1}^{\infty} i \Pr\{X = i\} = \sum_{i=1}^{n_{max}} ip(1-p)^{i-1} + n_{\max}(1-p)^{n_{\max}}.$$

Therefore, the expectation $E[X_i]$ may readily be calculated as:

$$E[X_i] = \frac{1}{p}\left[1 - (1-p)^{n_{\max}}\left(n_{\max}p + 1\right)\right] + n_{\max}(1-p)^{n_{\max}}. \tag{8}$$

Due to the fact that the queues are identical and independent, the expected service time $E[X] = E[X_i] + \tau$ can also be derived in terms of the original system time as:

$$E[X] = \frac{K}{p}\left[1 - (1-p)^{n_{\max}}\left(n_{\max}p + 1\right) + n_{\max}(1-p)^{n_{\max}}\right] - (K-1) + \tau, \tag{9}$$

where τ is the additional time of waiting for the BS response.

For our further calculations, we need to obtain the respective coefficient of variation, which is defined as the ratio between the standard deviation and the mean. Substituting expressions (8) and (11), we derive the sought formula as:

$$c_i = \frac{\sqrt{D[X_i]}}{E[X_i]} = \frac{\sqrt{E[X_i^2] - (E[X_i^2])^2}}{E[X_i]}, \tag{10}$$

where $E[X_i^2]$ is the second moment of the random variable X_i calculated as:

$$E[X_i^2] = \sum_{i=1}^{\infty} i^2 \Pr\{X = i\} = \sum_{i=1}^{n_{max}} i^2 p(1-p)^{i-1} + n_{\max}^2(1-p)^{n_{\max}}.$$

By the same calculations as above, it can be easily established that:

$$E[X_i^2] = \frac{1}{p}\left[1 - (1-p)^{n_{\max}}\left(n_{\max}p + 1\right)\right] + n_{\max}^2(1-p)^{n_{\max}}. \tag{11}$$

3.4 System Delay

The packet delay component of while the preceding packets in a particular queue i are being served (that is, waiting time in our M/G/1 system) can be given by the Pollaczek-Khinchine formula [21] as follows:

$$V_i^{(2)} = \frac{\rho_i E[X_i] \left(1 + c_i^2\right)}{2(1 - \rho_i)}, \tag{12}$$

where c_i and $E[X_i]$ are given above, and $\rho_i = \lambda_i E[X_i]$ is the load of the queue i.

When a packet arrives, it can join the queue which is the next one to be served (with the probability q), as well as any other queue uniformly. Hence, the period of waiting between the arrival and the moment when the queue i obtains opportunity to be served is:

$$V_i^{(1)} = q\left((K-1) + (K-2) + ... + (0)\right) = \frac{K(K-1)}{2K} = \frac{K-1}{2}. \tag{13}$$

Then, the total data packet delay δ in terms of the original system time may be given as the sum of the sojourn time W and the feedback time τ:

$$\delta = E[W_i] + \tau = E[V_i^{(1)}] + E[V_i^{(1)}] + E[X_i] + \tau =$$

$$= \frac{K-1}{2} + \frac{\rho E[X_i] \left(1 + c_i^2\right)}{2(1-\rho)} + KE[X_i] - (K-1) + \tau, \tag{14}$$

where $E[X_i]$ and c_i are given by expressions (9) and (10) respectively, while $\rho = K\lambda_i E[X_i] = \lambda E[X_i]$ is the system load and $E[X_i]$ is given by (8).

From the distribution of the service time, we derive the packet drop probability as the probability that a packet exhausts the maximum number of transmission attempts and would be discarded by the user:

$$P_{loss} = \Pr\{\text{packet is dropped} \mid \text{packet has been attempted}\} = (1-p)^{n_{\max}}. \tag{15}$$

3.5 User Energy Efficiency

The user energy consumption per subframe may be established as the sum of the fractions of time spent in every UE power state weighted by the actual power consumption in the respective state. We note that with respect to the original system time, the proportion of time that the UE spends in the transmit state exactly equals the system load [21]:

$$q_2 = \lambda E[X_i]. \tag{16}$$

Some additional energy expenditures come from the fact that the UE has to wait for the BS response/feedback for τ subframes after its successful transmission. We note that if the system is empty, that is, all the queues have no packets at the same subframe, then the UE would spend the idle power. However, due to the period τ of waiting after the successful transmission, the UE has to change its idle power level to active. The expected number of subframes per packet, when the UE thus spends P_1 instead of P_0 may be obtained as:

$$E[\tau_0] = \sum_{i=1}^{\tau-1} ip_0^i(1-p_0) + \tau {p_0}^{\tau} = \left[\frac{p_0}{1-p_0}\left(1-\tau p_0^{\tau-1}+\tau p_0^{\tau}-p_0^{\tau}\right)+\tau p_0^{\tau}\right],$$

where $p_0 = (1-\rho_i)^K$ is the probability that all the queues are empty at the same time and the individual queue load is given by:

$$\rho_i = \rho\left[1-\lambda_i a\left(\frac{a(1-(K-2)a^{K-3}+(K-3)a^{K-2})}{1-a}+(K-1)a^{K-1}-1\right)\right], \quad (17)$$

where $a = e^{-\lambda_i}$. Hence, the proportion of the idle subframes, which have switched to the active subframes, can be calculated as:

$$q_{0\to1} = \lambda\left[\frac{p_0}{1-p_0}\left(1-\tau p_0^{\tau-1}+\tau p_0^{\tau}-p_0^{\tau}\right)+\tau p_0^{\tau}\right]. \quad (18)$$

Further, we evaluate the remaining time fractions that the user spends in other states:

$$q_0 = p_0 - \lambda\left[\frac{p_0}{1-p_0}\left(1-\tau p_0^{\tau-1}+\tau p_0^{\tau}-p_0^{\tau}\right)+\tau p_0^{\tau}\right],$$

where p_0 is the probability that all the queues are empty. Obviously,

$$q_1 = 1-\left[(1-\rho_i)^K+\rho\right]+\lambda\left[\frac{p_0}{1-p_0}\left(1-\tau p_0^{\tau-1}+\tau p_0^{\tau}-p_0^{\tau}\right)+\tau p_0^{\tau}\right].$$

Accounting for the parameters q_0, q_1, q_2, we may obtain the exact value of the mean user energy expenditure as:

$$\epsilon = P_0 q_0 + P_1 q_1 + (P_2+P_1)\,q_2 =$$

$$= P_0 p_0 + P_1\left[1-p_0-\rho\right] + (P_2+P_1)\,\lambda E[X_i] + (P_1-P_0)\,q_{0\to1}, \quad (19)$$

where $p_0 = (1-\rho_i)^K$, and $q_{0\to1}$ and $E[X_i]$ are given by expressions (18) and (8) respectively. Finally, the mean user energy efficiency, which is defined as the effective stable arrival rate λ normalized by the corresponding energy consumption ϵ, follows from:

$$\phi = \frac{\lambda}{\epsilon}. \quad (20)$$

4 Numerical Results

In this section, we validate our analytical model with some simulation results. As the baseline case, we consider one LPN and a single UE moving between the macro BS and this LPN (see Figure 1). Note that the proposed approach can technically be applied to an arbitrary LPN deployment, as only the packet success probability is affected. Table 1 summarizes the main system parameters.

In Figure 6, energy efficiency as a function of the distance between the macro BS and the LPN is given. The UE cell-edge mobility is limited by the handover

Table 1. Primary system parameters

Notation	Parameter description	Value
–	Inter-cell distance	500 m [9]
Δt	Subframe size	1 ms
λ	Mean arrival rate of packets at the UE	0.5
N	Number of CoMP points per cell	2
$P_{\max}$	Maximum UE transmit power	23 dBm
$P_{Tx,0}$	Target transmit power	-82 dBm [9]
α	Pathloss compensation factor	0.8 [9]
Φ_{BS}	Macro BS transmit power	46 dBm [9]
Φ_{LPN}	LPN transmit power	30 dBm [9]
P_1	UE active power	100 mW
P_0	UE idle power	10 mW
K	Number of buffers at the UE	8
τ	BS response/feedback time	4 ms
$n_{\max}$	Max. number of packet transmissions	4
D	Distance between macro BS and LPN	250 m
N_0	Thermal noise power	-103 dB

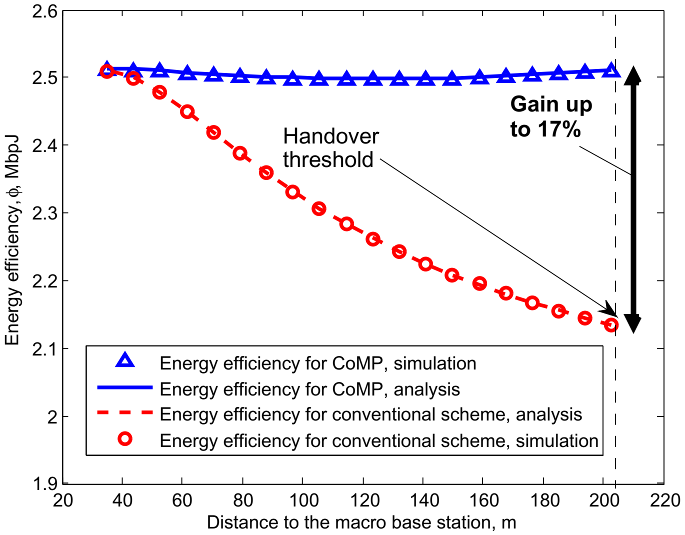


Fig. 6. User energy efficiency w.r.t. the distance between macro BS and LPN

threshold as defined in Section 2. Two alternatives are compared for the user in the UL: (i) its transmission to the serving macro BS exclusively (red curves) and (ii) improved macro given by CoMP selection combining scheme (blue curves). As expected, UL CoMP demonstrates consistent energy gains, which increase as long as the UE is moving toward the cell-edge. Hence, we conclude that already with one LPN the system may recover up to 17% of user energy efficiency.

We also investigate the behavior of user data packet delay. Figure 7 demonstrates the mean UE delay again depending on the distance between the macro BS and the LPN. Clearly, the use of UL CoMP enables significant delay reductions which become more pronounced at the edge of the cell. This is due to the improved values of packet success probability.

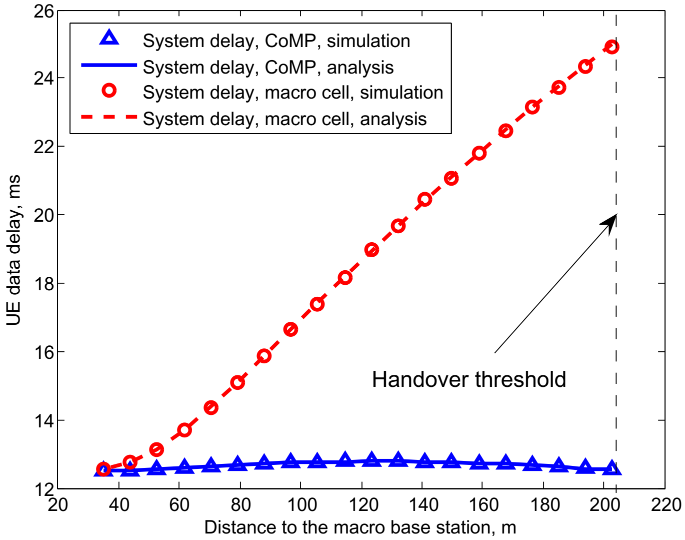


Fig. 7. User data packet delay w.r.t. the distance between macro BS and LPN

5 Conclusion

In this paper, we consider a heterogeneous 3GPP LTE deployment where neighboring LPNs may assist the macro-associated UE by independently receiving its UL data packets and forwarding the successful outcomes to the serving BS. Such CoMP scheme is known as selection combining and is believed to considerably improve user cell-edge performance. With our evaluation methodology, we combine analysis and simulations to account for the UE mobility, power control, and dynamic traffic load. We confirm that the expected energy efficiency and packet delay gains remain significant and consistent even for the low number of available LPNs.

Acknowledgments. This work is supported by Graduate School in Electronics, Telecommunications and Automation (GETA) as well as Internet of Things program of Tivit, funded by Tekes.

References

1. Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011-2016 (February 2012)
2. 3GPP LTE Release 10 & beyond (LTE-Advanced)
3. Irmer, R., Droste, H., Marsch, P., Grieger, M., Fettweis, G., Brueck, S., Mayer, H.-P., Thiele, L., Jungnickel, V.: Coordinated multipoint: Concepts, performance, and field trial results. IEEE Communications Magazine 49(2), 102–111 (2011)

4. Lee, D., Seo, H., Clerckx, B., Hardouin, E., Mazzarese, D., Nagata, S., Sayana, K.: Coordinated multipoint transmission and reception in LTE-Advanced: Deployment scenarios and operational challenges. IEEE Communications Magazine 50(2), 148–155 (2012)
5. Zheng, N., Boussif, M., Rosa, C., Kovacs, I., Pedersen, K., Wigard, J., Mogensen, P.: Uplink coordinated multi-point for LTE-A in the form of macro-scopic combining. In: Proc. of the IEEE Vehicular Technology (2010)
6. Choi, K., Kim, D.: Outage probability analysis of macro-diversity combining in Poisson field of access points. IEEE Communications Letters 16(8), 1208–1211 (2012)
7. Banani, S., Adve, R.: Required Base Station Density in Coordinated Multi-Point Uplink with Rate Constraints (2013), http://arxiv.org/abs/1302.1592
8. Lee, J., Kim, Y., Lee, H., Ng, B., Mazzarese, D., Liu, J., Xiao, W., Zhou, Y.: Coordinated multipoint transmission and reception in LTE-Advanced systems. IEEE Communications Magazine 50(11), 44–50 (2012)
9. Falconetti, L., Landstrom, S.: Uplink coordinated multi-point reception in LTE heterogeneous networks. In: Proc. of the 8th International Symposium on Wireless Communication Systems, pp. 764–768 (2011)
10. Huiyu, Y., Naizheng, Z., Yuyu, Y., Skov, P.: Performance evaluation of coordinated multipoint reception in CRAN under LTE-Advanced uplink. In: Proc. of the 7th International ICST Conference on Communications and Networking in China, pp. 778–783 (2012)
11. Zhang, J., Soldati, P., Liang, Y., Zhang, L., Chen, K.: Pathloss determination of uplink power control for UL CoMP in heterogeneous network. In: Proc. of the International Workshop on Cloud Base-Station and Large-Scale Cooperative Communications, pp. 250–254 (2012)
12. Simonsson, A., Andersson, T.: LTE uplink CoMP trial in a HetNet deployment. In: Proc. of the IEEE Vehicular Technology Conference (2012)
13. Mazzarese, D., Zhou, Y., Ren, X., Sun, J., Xia, L., Zhang, J.: An efficient feedback and scheduling scheme for cooperative multiple-point transmission in heterogeneous networks. In: Proc. of the 23rd Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 112–117 (2012)
14. Lin, C.-C., Sandrasegaran, K., Zhu, X., Xu, Z.: Limited CoMP handover algorithm for LTE-Advanced. Journal of Engineering 2013, 1–9 (2013)
15. Onireti, O., Heliot, F., Imran, M.: On the energy efficiency-spectral efficiency tradeoff in the uplink of CoMP system. IEEE Transactions on Wireless Communications 11(2), 556–561 (2012)
16. Yang, C., Han, S., Hou, X., Molisch, A.: How do we design CoMP to achieve its promised potential? IEEE Wireless Communications 20(1), 67–74 (2013)
17. 3GPP TR 36.819. Coordinated multi-point operation for LTE physical layer aspects (Release 11) (December 2011)
18. Goldsmith, A.: Wireless Communications. Cambridge University Press (2005)
19. Feller, W.: An introduction to probability theory and its applications, vol. 1. Wiley, New York (1968)
20. Winitzki, S.: A handy approximation for the error function and its inverse (2008)
21. Kleinrock, L.: Queueing Systems, vol. 1. Wiley Interscience, New York (1975)

Publication 4

© 2013 IEEE. Reprinted, with permission, from

O. Galinina, A. Turlikov, S. Andreev, and Y. Koucheryavy, ”Stabilizing Multi-Channel Slotted Aloha for Machine-Type Communications,” in *Proc. of the IEEE International Symposium on Information Theory (ISIT)*, 2013.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Tampere University of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

Stabilizing Multi-Channel Slotted Aloha for Machine-Type Communications

Olga Galinina*, Andrey Turlikov†, Sergey Andreev* and Yevgeni Koucheryavy*
*Department of Electronics and Communications Engineering
Tampere University of Technology, Korkeakoulunkatu 1, FI-33720, Tampere, Finland
Email: {olga.galinina, sergey.andreev}@tut.fi, and yk@cs.tut.fi
†Department of Information and Communication Systems
State University of Aerospace Instrumentation, Bolshaya Morskaya 67, 190000, St. Petersburg, Russia
Email: turlikov@vu.spb.ru

Abstract—**THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD. In this paper, we consider a wireless cellular system with an unbounded population of machine-type users. The system provides a number of non-interfering slotted-time channels which users contend for when sending their uplink data packets. We propose a provably stable control procedure for the channel access probability in the sense that it maintains a finite number of unserviced users in the system. We also compare the proposed algorithm against the optimal multi-channel slotted Aloha to conclude that our solution demonstrates near-optimum performance.**

I. Introduction

In wireless communications, Medium Access Control (MAC) algorithms are applied to arbitrate access of the user population to the shared channel. Currently, Aloha [1], slotted Aloha [2], and their numerous variations are the most deployed and studied solutions. The main idea of Aloha-based channel access is to defer the packet retransmission probabilistically whenever two or more users collide.

However, it is commonly known that the original (uncontrolled) slotted Aloha protocol may be unstable [3] in some scenarios. Addressing such instability, several dynamic control procedures have been proposed by [4] to prevent channel saturation. One such heuristic retransmission control algorithm, named later the Binary Exponential Backoff (BEB), allows a user to adjust its channel access probability basing on history of its own transmission attempts. As such, the use of BEB has been increasingly attractive due to little feedback required from the channel. Not surprisingly, over the years the BEB algorithm has become a part of most wireless systems.

Currently, BEB is widely used in cellular networks (such as 3GPP LTE and IEEE 802.16), when user requests resources from the network, as well as in local area networks (such as IEEE 802.11), for actual data transmissions. Whereas these systems were performing well in practice, the community has been cautioned about the poor asymptotic behavior of the BEB algorithm. For instance, in [5] the BEB algorithm was shown to be unstable in the infinite population model. By contrast, in [6] the BEB algorithm was demonstrated to be stable for sufficiently small arrival rates and finite population model. This seeming contradiction has been due to different assumptions and definitions of stability.

We reiterate the fact that infinite population model (where every new data packet is typically treated as a separate contending user) is known to highlight the limiting performance metrics of an algorithm, whereas finite population model (typically, with unbounded queues for newly arrived packets) provides insight into the practical applicability of an algorithm. Since in conventional human-oriented networks the number of users remained relatively small, the results for infinite population model have often been treated by most practitioners as a merely theoretical exercise.

This has been true until recently, when a huge number of unattended devices became integrated into the wireless infrastructure. Industry predicts that up to 30 thousand devices can connect to a single cellular base station [7] while infrequently transmitting very small data packets. Consequently, cellular technologies are now developing enhancements to support emerging machine-type communications (MTC) [8].

This evolved vision suggests a decisive step back toward the infinite population model. In such extreme conditions, the conventional BEB-based control schemes become overloaded and fail to guarantee an acceptable quality of service [9]. Hence, the incentive is growing high to seek for alternative stabilizing solutions by e.g. looking back to the classical works, such as [10], [11], or [12].

Another characteristic feature of MTC is that the system (e.g., 3GPP LTE) naturally provides a number of non-interfering channels (in time, frequency, or code) for MTC users to send their small data packets [13]. Whereas the original slotted Aloha is unstable for multiple channels [14], there are some practical works [15], [16] that propose feasible modifications. In particular, [16] and its extended version [17] build upon the control procedure from [11]. However, as no strict stability proof for infinite population model has been provided, it remains unclear whether these and many other algorithms will be stable for growing user population in MTC.

To the best of our knowledge, there has been no multi-channel algorithm that is provably stable for infinite population model in the sense that it can maintain a finite number of unserviced users in the system. We, therefore, propose our own variation of controlled slotted Aloha basing on the technique from [12] and rigorously prove its stability.

In the remainder of this paper, we first summarize the assumptions of our system model. Further, we formulate an optimal channel access algorithm, which however cannot be implemented in practical systems, and then our proposed algorithm. We show that our algorithm is stable for infinite user population and that its performance is close to the optimum.

II. System Model

We consider a centralized wireless system, where an unbounded number of users contend when sending their uplink data packets to a common base station. In our infinite population slotted-time model, we assume that $A(t)$ new users arrive in slot t (e.g., according to a homogeneous Poisson point process). The random variables $A(t)$ are independent and identically distributed, the arrival rate is $E[A(t)] = \lambda$, and $E[A^2(t)] < \infty$.

A newly arrived user acquires exactly one packet to be transmitted and competes with other backlogged users retransmitting their packets. Once a user has transmitted its packet successfully, it permanently leaves the system. Therefore, we may employ the terms user and packet interchangeably.

The system comprises K synchronous and non-interfering channels. Once a user decides to transmit, it randomly chooses one such channel following the uniform distribution. Packet transmissions may begin only at the start of a fixed-length slot and every user is synchronized to the slotted time. Data packets from all users have equal size and last the entire slot.

When a user has a packet to send, it may attempt to transmit it in any slot. Typically, a newly arrived user will send its packet in the next available slot, whereas backlogged users retransmit probabilistically. However, following e.g. [18] we assume that the newly arrived packets are treated identically to the backlogged ones.

At the end of a slot, the base station reliably determines the event in the channel, i.e. whether the channel was idle in this time slot or that it had successful or failed transmission. The transmission is considered successful if and only if exactly one user transmits on a given channel. If two or more users attempt to transmit on the same channel, all such transmissions are considered failed and the users remain backlogged. The maximum number of retransmission attempts is unbounded. The total number of unserviced users at the slot t forms the system backlog $N(t)$.

Every backlogged user decides whether to transmit in a particular slot with a given probability. We assume that such channel access probability is identical for all the backlogged users in the current slot and that it has been made available by the base station instantaneously and reliably at the end of the previous slot. The value of the channel access probability is determined by the base station basing on the history of the system. Various dynamic control procedures to manage this probability are considered in the rest of this paper.

III. Channel Access Algorithms

A. Optimal Control Procedure

Below we formulate the optimal procedure to control the channel access probability in our system.

We denote the channel access probability as $p(\lambda, w)$, where λ is the arrival rate and w denotes the complete history of the system up to the slot t (including information about individual transmission history, channel events, new user arrivals, system backlog, etc.) We are interested in formulating the optimal control procedure $p^*(\lambda, w)$ that minimizes the mean number of backlogged users in the system $n(t) = E[N(t)]$:

$$p^*(\lambda, w) = p^*(\lambda, N(t)) = \frac{K}{N(t)}, \qquad (1)$$

where K is the number of channels and $N(t)$ is the system backlog at the time slot t.

In [19], the optimality of a similar control procedure in a single-channel system (that is, for $K = 1$) has been rigorously proved. Here, we generalize the result from [19] for $K > 1$ and formulate it as a theorem. In existing literature (see e.g., [17]), this optimal procedure is mentioned as folklore knowledge, but the authors of this work are unaware of any strict proof of this important fact.

Theorem 1. *For any arrival flow such that the random variables $A(t)$ are i.i.d., $E[A(t)] = \lambda$, and $E[A^2(t)] < \infty$, the control procedure $p^* = \frac{K}{N(t)}$ maintains the minimum number of backlogged users in the system, i.e., $E[N(t; \lambda, p^*)] \leq E[N(t; \lambda, p)]$ for all $t = 0, 1, 2, ...$ and any $p \in S$, where S is the set of all possible control procedures.*

Proof is given in the Appendix. □

In practical systems, however, the implementation of such optimal control procedure is not possible. This is due to the fact that the number of backlogged users in the system $N(t)$ cannot be known even at the base station without prohibitive levels of additional coordination between the users. The question we seek to answer in the rest of this text is whether there is a feasible control procedure based solely on the history of channel events.

B. Proposed Control Procedure

Since the optimal control of channel access probability is impractical, we propose an alternative control procedure $p(\lambda, w)$ based only on the channel history and not requiring any information about the system backlog. We would like it to have the highest value of the maximum stable throughput λ_c, which is defined as the least upper bound of arrival rates λ for which the system is stable [20].

In [12], an adaptive control procedure for the single-channel slotted Aloha has been formulated and its stability conditions have been established. Motivated by [12], we propose our own control procedure suitable for a multi-channel system.

We introduce the following random process $Z(t)$, $Z(1) = 1$ with the evolution given as:

$$Z(t+1) = \max\{1, Z(t) + \Delta Z(t)\}, \qquad (2)$$

$$\Delta Z(t) = \sum_{i=1}^{K} \left(aI\{v_i(t)=0\} + bI\{v_i(t)=1\} + cI\{v_i(t) \geq 2\}\right),$$

where $v_i(t)$ is the number of users attempting to access the channel i at the slot t; $I\{v_i(t) = 0\}$, $I\{v_i(t) = 1\}$, $I\{v_i(t) \geq 2\}$ are the indicator functions of channel events, when 0, 1, or more users attempt to access the channel, respectively; and a, b, and c are some constant values which we discuss in more detail in the next section.

Summarizing, we propose the following channel access probability control procedure for multi-channel slotted Aloha, which is based on the channel events in the slot $t-1$:

$$p(\lambda, w) = p(\lambda, Z(t)) = \frac{K}{Z(t)}, \tag{3}$$

where the process $Z(t)$ has been described above.

IV. Proof of Stability

In this section, we establish stability conditions for the proposed control procedure. There is a variety of alternative definitions of stability with one of the first strict interpretations given by [21]. Most of the time, it is equivalent to stating that a stable algorithm can maintain a finite number of backlogged unserviced users with probability one [22].

We note that a random process $X(t) = (N(t), Z(t))$ constitutes a two-dimensional Markov chain. Therefore, the ergodicity of the process $X(t)$ would ensure the stability of our algorithm in terms of [22] and [21]. As such, below we determine a combination of coefficients a, b, and c of the process $Z(t)$, such that the process $X(t)$ is ergodic and the corresponding maximum stable throughput λ_c is the highest.

In order to establish the ergodicity conditions, we will further use the stability theorems for two-dimensional process formulated by [12]. We begin by obtaining the average drifts for the two-dimensional process $X(t)$ and then calculate the boundary vectors for the average drifts of $X(t)$.

A. Average Drifts

The average drifts for the process $X(t)$ are given as:

$$E[N(t+1) - N(t)|N(t) = n, Z(t) = z], \tag{4}$$

$$E[Z(t+1) - Z(t)|N(t) = n, Z(t) = z]. \tag{5}$$

Proposition 1. *Let $Y(t)$ represent the number of users leaving the system in the slot t; $N(t) = n$, $Z(t) = z$. Then, for a multi-channel slotted Aloha algorithm with K channels it holds:*

$$E[Y(t)|n, z] = \frac{nK}{z}\left(1 - \frac{1}{z}\right)^{n-1}.$$

Proof. By definition,

$$E[Y(t)|n, z] = E\left[\sum_{i=1}^{K} Y_i(t)|n, z\right] = \sum_{i=1}^{K} E[Y_i(t)|n, z],$$

where $Y_i(t)$ denotes the number of users that select the channel number i and successfully leave the system in the slot t. We note that $E[Y(t)|n, z] = K \cdot E[Y_i(t)|n, z], i = \overline{1, K}$, where $E[Y_i(t)|n, z]$ can easily be calculated as:

$$E[Y_i(t)|n, z] = \frac{np(t)}{K}\left(1 - \frac{p(t)}{K}\right)^{n-1} = \frac{n}{z}\left(1 - \frac{1}{z}\right)^{n-1},$$

Therefore,

$$E[Y(t)|n, z] = \frac{nK}{z}\left(1 - \frac{1}{z}\right)^{n-1}.$$

□

Lemma 1. *The drift (4) of the process $X(t)$ is established as:*

$$E[N(t+1) - N(t)|N(t) = n, Z(t) = z] = \lambda - \frac{nK}{z}\left(1 - \frac{1}{z}\right)^{n-1}. \tag{6}$$

Proof. We denote by $A(t)$ the number of newly arrived users and by $Y(t)$ the number of users that successfully leave the system in the slot t.

$$E[N(t+1) - N(t)|n, z] = E[A(t) - Y(t)|n, z] =$$

$$= E[A(t)] - E[Y(t)|n, z] = \lambda - E[Y(t)|n, z]. \tag{7}$$

We substitute the results of Proposition 1 into (7) to obtain:

$$E[N(t+1) - N(t)|N(t) = n, Z(t) = z] = \lambda - n\frac{K}{z}\left(1 - \frac{1}{z}\right)^{n-1}.$$

□

Lemma 2. *The drift (5) of the process $X(t)$ is established as:*

$$E[Z(t+1) - Z(t)|N(t) = n, Z(t) = z] =$$

$$= Kc + K(a-c)\left(1 - \frac{1}{z}\right)^{n} + K(b-c)\frac{n}{z}\left(1 - \frac{1}{z}\right)^{n-1}. \tag{8}$$

Proof. We consider the average increment of the process $Z(t)$:

$$E[\Delta Z(t)] = \sum_{i=1}^{K} E[aI\{v_i(t) = 0\} + bI\{v_i(t) = 1\} + cI\{v_i(t) \geq 2\}] =$$

$$= \sum_{i=1}^{K} (aE[I\{v_i(t) = 0\}] + bE[I\{v_i(t) = 1\}] + cE[I\{v_i(t) \geq 2\}]),$$

and find the sought expectation by using Proposition 1. (9)

Further, we consider a particular channel. The probability that it has been chosen (i) by zero users is $\pi_0 = (1 - p(t)\frac{1}{K})^n$, (ii) by exactly one user is $\pi_1 = np(t)\frac{1}{K}(1 - p(t)\frac{1}{K})^{n-1}$, and (iii) by two or more users is $\pi_2 = 1 - \pi_0 - \pi_1$, where $p(t) = \frac{K}{z}$. Then, we can establish the expressions for the components of the equation (9) as:

$$E\left[\sum_{i=1}^{K} I\{v_i(t) = 0\}\right] = \sum_{k=1}^{K} k \cdot \binom{K}{k}\pi_0^k(1 - \pi_0)^{K-k} =$$

$$= K\pi_0 = K(1 - \frac{1}{z})^n,$$

$$E\left[\sum_{i=1}^{K} I\{v_i(t) = 1\}\right] = \sum_{k=1}^{K} k \cdot \binom{K}{k}\pi_1^k(1 - \pi_1)^{K-k} =$$

$$= K\pi_1 = Kn\frac{1}{z}(1 - \frac{1}{z})^{n-1},$$

$$E\left[\sum_{i=1}^{K} I\{v_i(t) \geq 2\}\right] = \sum_{k=1}^{K} k \cdot \binom{K}{k}\pi_2^k(1 - \pi_2)^{K-k} =$$

$$= K\pi_2 = K\left(1 - (1 - \frac{1}{z})^n - n\frac{1}{z}(1 - \frac{1}{z})^{n-1}\right),$$

We now substitute the obtained results into the expression (9):

$$E\left[Z(t+1) - Z(t)|N(t) = n, Z(t) = z\right] = K\times$$

$$\times\left[a\left(1 - \frac{1}{z}\right)^n + b\frac{n}{z}\left(1 - \frac{1}{z}\right)^{n-1} + c\left(1 - \left(1 - \frac{1}{z}\right)^n - \frac{n}{z}\left(1 - \frac{1}{z}\right)^{n-1}\right)\right] =$$

$$= Kc + K(a-c)\left(1 - \frac{1}{z}\right)^n + K(b-c)\frac{n}{z}\left(1 - \frac{1}{z}\right)^{n-1}.$$

□

B. Boundary Vectors and Stability

In what follows, we continue by establishing the boundary vectors of the average drift. By a limit argument $\sqrt{n^2 + z^2} \to \infty, n/z = k$, we obtain the following boundary vector function $\mu = (\mu_n(k), \mu_z(k))$ of the process $X(t)$:

$$\mu_z(k) = \lim_{\sqrt{n^2+z^2}\to\infty, n/z=k} E[Z(t+1) - Z(t)|N(t) = n, Z(t) = z],$$

$$\mu_z(k) = Kc + K(a-c)e^{-k} + K(b-c)ke^{-k}, \tag{10}$$

$$\mu_n(k) = \lim_{\sqrt{n^2+z^2}\to\infty, n/z=k} E[N(t+1) - N(t)|N(t) = n, Z(t) = z],$$

$$\mu_n(k) = \lambda - Kke^{-k}. \tag{11}$$

We now investigate the stability by solving the vector equation $\mu(k)||\mathbf{k}$, which is equivalent to an equation:

$$\mu_n(k) = k\mu_z(k). \tag{12}$$

Substituting the expressions (10) and (11) into (12), we derive:

$$\lambda - Kke^{-k} = K\dot{k}(c+(a-c)e^{-k}+(b-c)ke^{-k}), k > 0. \quad (13)$$

Here, we refer to the stability theorems in [12] for two-dimensional process and claim that the considered Markov chain is ergodic if $\mu_n(k) < 0$ for all roots of (13). Next, we establish the maximum stable throughput λ_c.

Lemma 3. *For a multi-channel slotted Aloha algorithm with K channels, the maximum stable throughput $\lambda_c \leq Ke^{-1}$.*

Proof. Here, we refer to the results of [12] and [19]. The considered Markov chain $X(t)$ is ergodic only if the condition $\mu_n(k) < 0$ holds. Hence, $\lambda - Kke^{-k} < 0$. We also note that $ke^{-k} < e^{-1}$. Therefore, if $\lambda > Ke^{-1}$ then the chain is not ergodic, and, consequently, any stable throughput is always bounded by Ke^{-1}. As such, $\lambda_c \leq Ke^{-1}$. □

We finally establish the parameters of the process $Z(t)$, such that λ_c is the highest, that is, $\lambda_c = Ke^{-1}$.

Theorem 2. *If for the coefficients a, b, and c of the process $Z(t)$, where $c > 0$ and $a < 0$, it holds the following:*

$$c \cdot (e-2) + a + b = 0, \quad (14)$$

then the Markov chain $X(t)$ is ergodic and $\lambda_c = Ke^{-1}$.

Proof. We seek to find such values of a, b, and c that $\lambda_c = Ke^{-1}$. We also remind that for $\lambda < \lambda_c$ it holds that $\mu_n(k) < 0$. Firstly, we consider the maximum stable throughput $\lambda_c = Ke^{-1} > \lambda$. Therefore,

$$K(-k_i)e^{-k_i} = -\frac{K}{e}, k_i = 1, \quad (15)$$

We now substitute (15) into (13):

$$0 = Kk\left(c + (a-c)e^{-k} + (b-c)ke^{-k}\right),$$

$$c + (a-c)\frac{1}{e} + (b-c)\frac{1}{e} = 0.$$

Hence, satisfying the following equation provides ergodicity for any $\lambda < \lambda_c = Ke^{-1}$: $c \cdot (e-2) + a + b = 0$. It can also be proved that the Markov chain $X(t)$ is not recurrent if $a > 0$ or $c < 0$. For that reason, we only consider such sets of coefficients, when $a < 0$ and $c > 0$. □

Theorem 3. *If the coefficients a, b, and c satisfy the conditions of Theorem 2, then the proposed control procedure converges to the optimal control procedure in the sense that $n(t) \to n^*(t)$ for $K \to \infty$, where $n(t)$ is the system backlog under the proposed control and $n^*(t)$ is the respective value under the optimal control.*

Proof of this theorem employs the same technique as the proof of Theorem 1. But it is also using an auxiliary proposition that the random process $N(t)$ converges to $Z(t)$ for $K \to \infty$, and, particularly, $\delta(t) = (1/N(t) - 1/Z(t)) \to_{K\to\infty} 0$. This proposition can be proved analyzing the system behavior for K and $K+1$, accounting for the definition of a limit, and at the same time studying the convergence rate. We omit the proof here due to space limitations. □

V. Numerical Results

In this section, we compare the proposed multi-channel control procedure against the optimal multi-channel control. To give more insight, we also consider a single-channel procedure in the equivalent conditions, that is, when a newly arrived user is uniformly assigned a particular channel to transmit on and no channel reselection is allowed by users.

Such independent use of channels allows accounting for only one channel i with the optimal channel access probability $p_i(t) = 1/N_i(t)$, where $N_i(t)$ is the number of users sharing this channel. The performance metrics for the single-channel case may be obtained numerically according to [19]. For other procedures, we present simulation data of 10^6 slots. For the proposed control procedure, we set the parameters of the process $Z(t)$ (see eq. (3)) as $a = -1$, $b = -1$, and $c = \frac{1}{e-2}$. These values satisfy the conditions of Theorem 2.

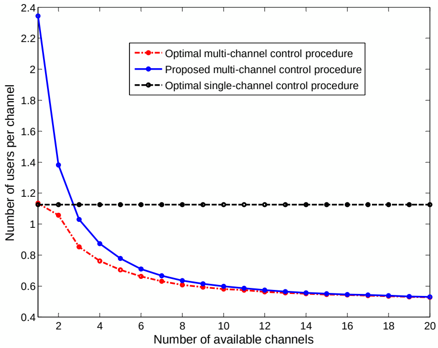


Fig. 1. System backlog per channel for arrival rate $\lambda = 0.3 \cdot K$.

In Figure 1, we fix the arrival rate per channel to be sufficiently high (i.e., $\lambda = 0.3 \cdot K$) and study the average number of users in the system for all the three procedures by varying the number of available channels. From the figure, we learn that the proposed multi-channel procedure rapidly approaches optimum with the increase in K, whereas the single-channel procedure maintains independent constant performance.

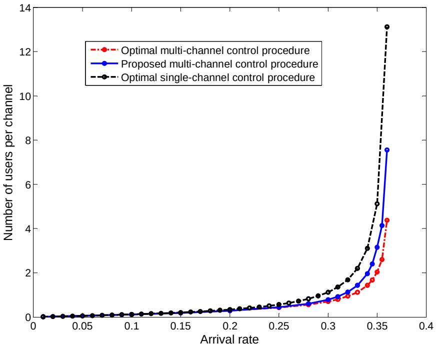


Fig. 2. System backlog per channel for $K = 5$ channels.

Contemporary practical single-channel multiple access schemes (typically, BEB-based) are, therefore, expected to perform not better than the optimal single-channel procedure. Therefore, already for $K = 3$ we conclude that the proposed procedure benefits over existing solutions owing to the joint use of channels.

In Figure 2, we fix the number of available channels to be $K = 5$ and again study the average system backlog by varying the arrival rate per channel. The use of the proposed control procedure results in the lower system backlog than that for the optimal control over a single channel, and remains close to the multi-channel optimum in the entire range of feasible arrival rates.

VI. Conclusion

This work considers a wireless cellular system with an unbounded population of contending users. These users share a number of non-interfering uplink channels subject to a common channel access probability advertised by the base station. Whereas we demonstrate that the optimal control of such probability is not feasible, we also detail a practical adaptive procedure that provably maintains a finite number of unserviced users in the system. With the increasing number of channels, the proposed procedure quickly converges to the optimal solution. We, therefore, conclude that our stabilized multi-channel slotted Aloha algorithm is naturally suitable for future machine-type systems with large user population.

Appendix

In this appendix, we prove Theorem 1 by application of the mathematical induction.

Proof. The basis of the mathematical induction follows from the fact that the number of users in the first slot does not depend on the control procedure:

$$E[N(0;\lambda,p^*)] = E[N(0;\lambda,p)].$$

We note that $N(t+1) = N(t) + A(t) - Y(t)$, where $A(t)$ is the number of newly arrived users in the slot t and $Y(t)$ is the number of users which successfully leave the system.

Let the induction hypothesis state for some $t > 0$ that $E[N(t;\lambda,p)] \geq E[N^*(t;\lambda,p^*)]$. Therefore,

$$E[N(t+1;\lambda,p)] - E[N^*(t+1;\lambda,p^*)] =$$
$$= (E[N(t;\lambda,p)] - E[N^*(t;\lambda,p^*)]) - (E[Y(t;\lambda,p)] - E[Y(t;\lambda,p^*)) \geq$$
$$\geq E[Y(t;\lambda,p^*)] - E[Y(t;\lambda,p)] =$$
$$= E\left[\sum_{i=1}^{K} Y_0^i(t;\lambda,p^*)\right] - E\left[\sum_{i=1}^{K} Y_0^i(t;\lambda,p)\right],$$

where $Y_0^i(t;\lambda,p^*)$ is the number of users which leave the system while using the channel i. We note that $Y_0^i(t;\lambda,p^*)$ does not depend on the number of the channel. Hence,

$$E[N(t+1;\lambda,p)] - E[N^*(t+1;\lambda,p^*)] \geq$$
$$\geq K \cdot E[Y_0(t;\lambda,p^*)] - K \cdot E[Y_0(t;\lambda,p)].$$

Let $\xi(t)$ denote the complete (system) history of the contention process up to the moment t and Ω is the space of a random variable $\xi(t)$. $\Omega_i(t)$ is the set of $\omega \in \Omega$, such that $n(t) = i$. Decomposing the expectation of $Y_0(t;\lambda,p)$ and then the probability $Pr\{Y_0(t;\lambda,p) = j\}$, we obtain:

$$E[Y_0(t;\lambda,p)] = \sum_{j=0}^{\infty} jPr\{Y_0(t;\lambda,p) = j\} =$$
$$= \sum_{j=0}^{\infty} j \sum_{i=0}^{\infty} \sum_{\omega\in\Omega_i(t)} Pr\{Y_0(t;\lambda,p) = j|\xi(t;\lambda,p) = \omega\} \times$$
$$\times Pr\{\xi(t;\lambda,p) = \omega\} =$$
$$= \sum_{i=0}^{\infty} \sum_{\omega\in\Omega_i(t)} Pr\{Y_0(t;\lambda,p) = 1|\xi(t;\lambda,p) = \omega\} \times$$
$$\times Pr\{\xi(t;\lambda,f) = \omega\} =$$
$$= \sum_{i=0}^{\infty} \sum_{\omega\in\Omega_i(t)} \frac{ip(\lambda,\omega)}{K} \left(1 - \frac{p(\lambda,\omega)}{K}\right)^{i-1} Pr\{\xi(t;\lambda,p) = \omega\}.$$

Since the function $ix(1-x)^{i-1}$ achieves its maximum at the point $x = \frac{1}{i}$, i.e., $p(\lambda,\omega) = \frac{K}{i}$, we establish the following:

$$E[Y_0(t;\lambda,p)] \geq \sum_{i=0}^{\infty} \sum_{\omega\in\Omega_i(t)} (1 - \frac{1}{i})^{i-1} Pr\{\xi(t;\lambda,p) = \omega\} =$$
$$= \sum_{i=0}^{\infty} \left(1 - \frac{1}{i}\right)^{i-1} \sum_{\omega\in\Omega_i(t)} Pr\{\xi(t;\lambda,p) = \omega\} =$$
$$= \sum_{i=0}^{\infty} \left(1 - \frac{1}{i}\right)^{i-1} Pr\{N(t;\lambda,p) = i\}.$$

Finally, we derive:

$$E[N(t+1);\lambda,p] - E[N^*(t+1);\lambda,p^*] \geq$$
$$\geq K \sum_{i=0}^{\infty} \left(1 - \frac{1}{i}\right)^{i-1} Pr\{N(t;\lambda,f) = i\} - K \cdot E[Y_0(t;\lambda,p)] = 0.$$

□

References

[1] N. Abramson, "The Aloha system – another alternative for computer communications," in *Proc. AFIPS*, pp. 281–285, 1970.

[2] L. Roberts, "ALOHA packet system with and without slots and capture," in *Proc. ACM SIGCOMM*, vol. 5, pp. 28–42, 1975.

[3] L. Kleinrock and S. Lam, "Packet switching in a multiaccess broadcast channel: Performance evaluation," *IEEE Trans. Commun.*, vol. COM-23, pp. 410–423, 1975.

[4] S. Lam and L. Kleinrock, "Packet switching in a multiaccess broadcast channel: Dynamic control procedures," *IEEE Trans. Commun.*, vol. COM-23, pp. 891–904, 1975.

[5] D. Aldous, "Ultimate instability of exponential back-off protocol for acknowledgment based transmission control of random access communication channels," *IEEE Trans. Inf. Theory*, vol. 33, pp. 219–223, 1987.

[6] J. Goodman, A. Greenberg, N. Madras, and P. March, "Stability of binary exponential backoff," *Journ. of the ACM*, vol. 35, pp. 579–602, 1988.

[7] M. Cheng, G. Lin, H. Wei, and A. Hsu, "Overload control for machine-type-communications in LTE-Advanced system," *IEEE Commun. Mag.*, vol. 50, pp. 38–45, 2012.

[8] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. Johnson, "M2M: from mobile to embedded Internet," *IEEE Commun. Mag.*, vol. 49, pp. 36–43, 2011.

[9] M. Gerasimenko, V. Petrov, O. Galinina, S. Andreev, and Y. Koucheryavy, "Impact of MTC on energy and delay performance of random-access channel in LTE-Advanced," *Trans. on Emerging Telecom. Techn.*, vol. t.b.d., p. submitted, 2013.

[10] B. Hajek and T. Van Loon, "Decentralized dynamic control of a multiaccess broadcast channel," *IEEE Trans. Autom. Control*, vol. 27, pp. 559–569, 1982.

[11] R. Rivest, "Network control by Bayessian broadcast," *IEEE Trans. Inf. Theory*, vol. 33, pp. 323–328, 1987.

[12] V. Mikhailov, "Geometrical analysis of the stability of Markov chains in Rn+ and its application to throughput evaluation of the adaptive random multiple access algorithm," *Probl. Inform. Transm.*, vol. 24, pp. 47–56, 1988.

[13] S. Andreev, A. Larmo, M. Gerasimenko, V. Petrov, O. Galinina, T. Tirronen, J. Torsner, and Y. Koucheryavy, "Efficient small data access for machine-type communications in LTE," in *Proc. IEEE ICC*, 2013.

[14] I. Pountourakis and E. Sykas, "Analysis, stability and optimization of Aloha-type protocols for multichannel networks," *Computer Comm.*, vol. 15, pp. 619–629, 1992.

[15] W. Yue and Y. Matsumoto, "Output and delay of multi-channel slotted ALOHA systems for integrated voice and data transmission," *Telecom. Systems*, vol. 13, pp. 147–165, 2000.

[16] D. Shen and V. Li, "Stabilized multi-channel ALOHA for wireless OFDM networks," in *Proc. IEEE GLOBECOM*, vol. 1, pp. 701–705, 2002.

[17] D. Shen and V. Li, "Performance analysis for a stabilized multi-channel slotted ALOHA algorithm," in *Proc. IEEE PIMRC*, pp. 249–253, 2003.

[18] A. MacKenzie and S. Wicker, "Stability of multipacket slotted Aloha with selfish users and perfect information," in *Proc. IEEE INFOCOM*, pp. 1583–1590, 2003.

[19] G. Falin, "Performance evaluation for a class of algorithms of random multiple access to a radio-communication channel," *Probl. Peredachi Inf.*, vol. 18, pp. 85–90, (in Russian) 1982.

[20] D. Bertsekas and R. Gallager, *Data Networks*. Prentice Hall, 1992.

[21] B. Tsybakov and V. Mikhailov, "Random multiple packet access: Part-and-try algorithm," *Probl. Inform. Transm.*, vol. 16, pp. 305–317, 1980.

[22] D. Chan and T. Berger, "Upper bound for the capacity of multiple access protocols on multipacket reception channels," in *Proc. IEEE ISIT*, pp. 1603–1607, 2012.

Publication 5

© 2014 IEEE. Reprinted, with permission, from

O. Galinina, S. Andreev, M. Gerasimenko, Y. Koucheryavy, N. Himayat, S. Yeh, and S. Talwar, "Capturing Spatial Randomness of Heterogeneous Cellular/WLAN Deployments With Dynamic Traffic", *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 1083-1099, 2014.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Tampere University of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to `http://www.ieee.org/publications_standards/publications/rights/rights_link.html` to learn how to obtain a License from RightsLink.

Capturing Spatial Randomness of Heterogeneous Cellular/WLAN Deployments with Dynamic Traffic

Olga Galinina[†], Sergey Andreev, Mikhail Gerasimenko, Yevgeni Koucheryavy,
Nageen Himayat, Shu-ping Yeh, and Shilpa Talwar

***Abstract*—As fourth generation communications technology is already being deployed, the attention of the recent research efforts is shifting to what comes beyond the state-of-the-art wireless systems. Driven by the anticipated acceleration in mobile traffic demand, wireless industry is specifically focused on improving capacity and coverage of current networks through aggressive reuse of cellular spectrum. Together with deploying an increasingly dense overlay tier of smaller cells, mobile network operators are beginning to rely on unlicensed-band WLAN technologies to leverage additional spectrum and relieve congestion on their networks. Consequently, the emerging vision of *heterogeneous networks* exploits the potential of a diverse range of devices requiring connectivity at different scales to augment available system capacity and improve user connectivity experience.**

In this paper, we seek to meet this important trend with our novel integrated methodology for *assisted* (managed) radio network selection capturing spatial randomness of converged cellular/WLAN deployments together with *dynamic* uplink traffic from their users. To this end, we employ tools coming from stochastic geometry to characterize performance of macro and pico cellular networks, as well as WLAN, mindful of user experience and targeting intelligent network selection/assignment. We complement our analysis with system-level simulations providing deeper insights into the behavior of future heterogeneous deployments.

I. Introduction and Motivation

WHEREAS decisive improvements in many aspects of wireless system design have indeed been offered by the recently completed fourth generation mobile broadband standards [1], it is widely believed that current technology will still be unable to face the projected growth of traffic demand [2]. According to the recent predictions in [3], the offered volumes of mobile data will increase at least 13-fold over the following 5 years, aggravated by the rapid proliferation in types and numbers of wireless devices. As a consequence, the available network *capacity* and *coverage* may become insufficient [4] thus severely degrading the resulting quality of service (QoS) for end users [5].

With a historical 10-year cycle for every existing generation, it is expected that novel fifth generation (5G) wireless systems will be deployed sometime around 2020 [6]. While there is currently no complete definition of what comes after the state-of-the-art networking technologies, many agree in that the only comprehensive solution to mitigate the increasing disproportion between the user QoS and the available wireless resources is by deploying the higher density of femto and pico cells in current cellular architecture. Due to shorter radio links, network densification generally promises higher bit rates and reduced energy for uplink transmission, especially in urban cellular environments [7].

However, licensed spectrum continues to be scarce and expensive, whereas the traditional methods to improve its efficient use approach their theoretical limits [8]. Therefore, it is expected that the majority of near-term capacity gains will come from advanced architectures and protocols that would leverage the unlicensed spectrum and take advantage of the intricate interactions between the device and the network, as well as between the devices themselves, across the converged *heterogeneous* deployments. Consequently, the incentive to efficiently coordinate between the alternative radio access technologies (RATs) is growing stronger [9]. Here, the distributed unlicensed-band networks (e.g., WLAN technologies) may take advantage of the *centralized* control function residing in the cellular network to effectively perform dynamic multi-RAT network association.

In summary, as cell-sizes shrink, the footprints of cellular, local, and personal area networks are increasingly overlapping. This creates an opportunity to simultaneously utilize multiple RATs for improved capacity and connectivity [10], [11]. However, very limited research attention has been dedicated to the *assisted* joint use of multiple networks, whereas much effort has been invested into optimizing the performance of individual radio technologies. We firmly believe that intelligent coupling between multiple RATs may leverage several dimensions of diversity (including spatial, temporal, frequency, interference, load, etc.) and that both short- and long-range technologies may need to work cooperatively [12] to realize the desired improvements in capacity and service experience.

II. General Background and Our Contributions

A. *Current Trends in Heterogeneous Networking*

Over the past few years, tighter interworking between various RATs has been receiving more momentum [13]. While previously cellular and WLAN technologies were developing largely independently, today WiFi is becoming an integral part of an operator's cellular network [14]. This is on the one hand due to the fact that contemporary consumer devices massively support WiFi together with other RATs. On the other hand,

O. Galinina, S. Andreev, M. Gerasimenko, and Y. Koucheryavy are with the Department of Electronics and Communications Engineering, Tampere University of Technology, FI-33720 Tampere, Finland.

N. Himayat, S.-p. Yeh, and S. Talwar are with Intel Corporation, Santa Clara, CA, USA.

[†]O. Galinina is the contact author: P.O. Box 553, FI-33101 Tampere, Finland; e-mail: olga.galinina@tut.fi

Manuscript received December 06, 2013; revised April 06, 2014.

mobile network operators increasingly rely on WLAN-based offloading to relieve congestion on their cellular networks [15] and hence desire more control of how WLAN is utilized and managed.

In light of the above, Heterogeneous Networks (HetNets) have recently emerged as advanced networking architecture (see Figure 1) enabling aggressive capacity and coverage improvements towards future 5G networks [16]. This architecture comprises hierarchical deployment of wide-area macro cells for basic connectivity and coverage augmented with (densely deployed [17]) small cells of various footprints and by different RATs (femto and pico cells, WiFi access points, relay nodes, integrated WiFi-LTE small cells, etc.) to boost capacity [18]. In particular, when WLAN is managed as part of an operator's cellular network, more advanced levels of interworking between cellular and WLAN RATs become available.

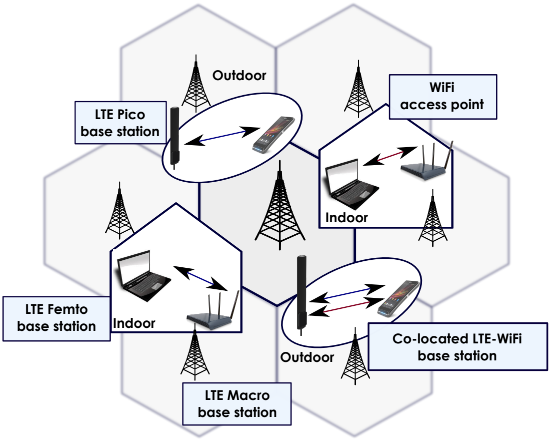


Fig. 1. Our vision of a heterogeneous network.

Not surprisingly, recent literature has been very rich in addressing the important aspects of load balancing and access network selection for multi-RAT HetNets [19]. The existing publications range from considering simpler user-centric network selection strategies (known as vertical handover) to full multi-tier and multi-radio cooperation [20], [21], e.g., where WiFi becomes a "virtual carrier" anchored on the cellular network. However, the focus has been mostly on centrally-managed systems with full control at the base station or totally distributed solutions, but not so much on network-assisted schemes. Most recently, the concept of LTE-unlicensed has attracted interest of industry and academia alike with the goal of allowing LTE systems to utilize bandwidth-rich unlicensed spectrum around 5 GHz band to augment their capacity [22]. Another emerging industry trend considered in latest publications is multi-radio small cells with *co-located* cellular and WLAN interfaces able to reduce deployment costs and leverage common infrastructure across heterogeneous cells [23].

Reacting to this recent interest, the Third Generation Partnership Project (3GPP) is becoming increasingly active in developing new interworking solutions between 3GPP cellular technologies, such as UMTS or LTE, and WiFi (IEEE 802.11 [24]) technology. However, given that co-located cellular/WLAN deployments are presently not common, current standardization efforts focus more on user-centric interworking architectures while only assuming limited degrees of cooperation/assistance across the HetNet [25]. The field of investigation spans across (i) schemes for trusted access to 3GPP services with WLAN devices, (ii) support for Access Network Discovery and Selection functions, and (iii) seamless mobility between cellular and WLAN technologies.

More recently, several new study/work items have been open targeting the interworking solutions that involve cooperation within the Radio Access Network (RAN) [26] by contrast to prior schemes that have loosely defined functions within the 3GPP core network (such as security and inter-RAT mobility) [27], [28], [29]. This shift is dictated by the need to support improved QoS on WLAN networks as prescribed by a consortium of network operators with their tighter requirements for carrier grade WiFi. The WLAN community has also responded with their new initiatives on Hot Spot 2.0, as well as an emerging "High Efficiency WLAN" effort by the IEEE 802.11 work group. Therefore, we expect the trend for tighter integration of cellular and WLAN technologies to continue by potentially encompassing other radio technologies beyond current WiFi and additional use cases beyond spectrum aggregation.

B. Characterizing Uplink Traffic Dynamics

In this work, we specifically focus on the important problem of network selection between multi-tier cellular and WLAN RATs [30] assuming that WLAN belongs to an operator deployed and *managed* multi-RAT HetNet. Further, we make a step ahead with respect to the current 3GPP efforts and consider intelligent *assistance* from the cellular network in the RAT selection process, when a new coordinating entity in the cellular RAN is made to receive relevant information from multi-radio devices (e.g., their position, QoS requirements, how much interference/load they sense on the nearby WLAN networks, etc.) and then advises the users on the attractive connectivity options. We intend to thoroughly investigate a particular assisted network admission and selection scheme, as well as demonstrate that it may provide considerable improvements in overall system performance while guaranteeing the required user QoS.

For consistency with current network deployments, we concentrate on distributed small cell overlay with standalone WiFi access points as well as pico cell base stations, assuming that there is no direct interface between the cellular and WLAN radio networks. However, our methodology may also characterize co-located cellular/WLAN deployments as well as more advanced technologies and scenarios to become appealing in the context of 5G networks [31]. In particular, the network selection mechanisms considered are operating at the RAN layer, which resides below the IP layer. More specifically, we focus on *uplink* performance as it has not been fully addressed in existing literature due to more challenging interference-related aspects [32].

To further advance the state-of-the-art research on HetNets primarily focusing so far on static (full-buffer) steady-state formulations, we target *flow-level* performance and consider stochastic traffic loads. In particular, new data flows representing, e.g., real-time data sessions with the minimum target bitrate are arriving randomly and leave the system after the service has been received [33]. Consequently, the number of active flows varies with time, which is often referred to as the flow-level dynamics. Analyzing dynamic setups is important to gain better understanding of real-world systems, but it also incurs extra complexity. Therefore, dynamic systems receive much less research attention than their static alternatives, that is, with a fixed set of backlogged users.

Every data flow in a dynamic network may generally represent a stream of packets corresponding to a new file transfer, web-page browsing, or real-time voice/video session. To mimic the flows produced by a large population of independent users, Poisson processes have extensively been applied in the past. Originally, flow-level frameworks were helpful investigating flexible bandwidth allocation mechanisms in the context of wired systems. Extending their applicability to wireless networks, it was concluded that the throughput experienced by a dynamic user population can substantially differ from that received by a fixed number of users [34]. As such, studying dynamic wireless systems is becoming increasingly important and we concentrate on characterizing HetNet dynamics in what follows.

C. Coupling with Spatial Randomness

Another important aspect of HetNets is that locations of the network users relative to each other highly impact the resulting system performance [35]. Indeed, given that users are not regularly spaced, there may be a high degree of spatial randomness which needs to be considered explicitly. We thus adopt an appropriate random spatial model where user locations are drawn from a particular realization of a random process. Coupling such topological randomness with system dynamics requires a fundamental difference in characterizing user signal power and interference. Fortunately, the field of stochastic geometry provides us with a rich set of powerful results and analytical tools that can capture the network-wide performance of a random user deployment [36].

The use of stochastic geometry (that is, statistical modeling of spatial relationships) has become increasingly popular over the last decades to analyze network performance averaged over multiple spatial realizations. As part of a more recent surge, it has also been useful in characterizing many important aspects of current cellular technology, from conventional macro cell deployments to hyper-dense heterogeneous and small cell networks [37]. The application of stochastic geometry typically features a particular spatial point process to statistically capture, e.g., user locations yielding insights on the impacts of user density, transmit power, path loss, and interference.

In the absence of prior information about user locations, the simplest statistical tool to model user deployment is a uniform distribution within a finite area in the two-dimensional plane. Such uniformity constitutes a direct result of applying a homogeneous Poisson point process, which in turn assumes that, conditioned on the number of points of the process lying in an arbitrary finite region, the points are independently and uniformly distributed over that region.

Other more realistic, but also significantly more complex point processes are binomial process spawning a fixed number of users in a given area and Poisson cluster process allowing users to cluster in certain locations. Finally, there is also hard core point process which is a thinning of the Poisson point process such that the users have a guaranteed minimum separation.

In summary, the **main contributions** of this paper are (i) a novel space-time *analytical* HetNet model that couples spatial randomness of user locations with their uplink flow-level traffic dynamics by contrast to full-buffer (saturated) considerations; (ii) a particular (heuristic) policy of *assisted* user admission and intelligent RAT selection/assignment, which results in an adequate compromise between fully-distributed (uncoordinated) and centrally-controlled solutions; (iii) a detailed *system-level* evaluation of the respective network and user performance used to verify our analysis as well as to provide deeper insights into the behavior of future HetNet deployments.

III. Proposed HetNet-Centric Methodology

In this section, we introduce our integrated system model comprising WLAN, macro, and pico cellular networks, which we hereinafter refer to as *tiers*. We summarize the core assumptions of the model below.

A. General Assumptions

We study one (typical) cell of a macro network with the radius R featuring a macro base station (BS) in its center together with several pico BSs and WLAN access points (APs), as depicted in Figure 2. In what follows, the macro cell is termed the *Macro tier*, while the pico cell and the WLAN are referred to as the *Pico tier* and the *WLAN tier*, respectively. All the BSs/APs are capable of serving uplink data from their wireless users concurrently. The considered traffic is characteristic of real-time sessions with the target bitrate of r_0.

Basing on the recent specifications [26], we further assume *non-overlapping* frequency bands for all three tiers. Therefore, user transmissions on one tier do not interfere with those on the other. However, all WLAN/Pico tier links share the frequency bands of their respective tiers and thus interfere, whereas the Macro tier is interference-free. Our general system model is illustrated in Figure 2 representing areas of the Macro, Pico, and WLAN tiers together with the corresponding users and infrastructure nodes.

We assume that the transmitting users (or, *sessions*) with some uplink data traffic demand arrive on the joint network according to one-dimensional Poisson process of rate λ in time. We thus associate a newly arrived user with its session and its location, which is assumed fixed throughout the session's lifetime. For the sake of tractability, we also assume that the duration of a user session is exponentially distributed with

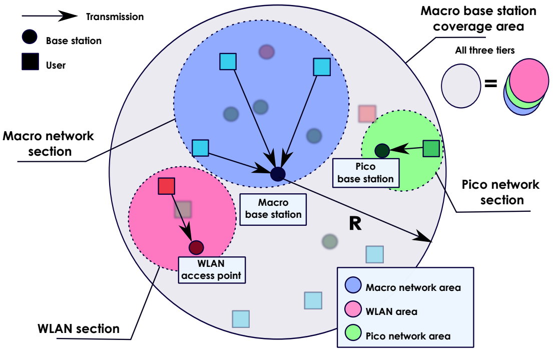


Fig. 2. System model of a HetNet with three tiers within macro BS coverage range of radius R: the *cuts* demonstrate different network tiers.

the mean of μ^{-1}, which may correspond to, e.g., a real-time voice/video call.

To explicitly model topological randomness in our network, we employ several stochastic processes and, to this end, formulate the following principal assumptions.

Assumption 1. ***Spatial distribution of infrastructure.*** *The locations of APs (on the WLAN tier) and BSs (on the Pico tier) are independent and spatially distributed according to a Poisson point process (PPP) on the two-dimensional plane with the rates of L_w and L_p, respectively.*

We note that the densities L_w and L_p may be thought of as the average numbers of APs/BSs per a unit of area, while their positions may be located outside the circle of macro coverage (potentially, serving other-cell users).

To further account for traffic dynamics, we adopt the following assumption on the user locations.

Assumption 2. ***Spatial distribution of users.*** *The locations of arriving users are distributed according to the PPP on the two-dimensional plane. The area of our interest is limited by the considered macro cell (e.g., circle $\mathbf{B}$ of radius R), which results in uniform distribution of users within $\mathbf{B}$.*

We note that in practice the restriction of deploying users within a particular area may be dictated by, e.g., maximum transmit power constraints and/or channel degradation factors. Moreover, uniform distribution is only assumed here as a baseline example. Generally, we may consider an arbitrary joint distribution $f(x, y)$ of user locations, which would somewhat complicate further analysis technically, but without significant impact on the derivation methodology.

The user arrival pattern together with the spatial distribution of their locations constitute a *space-time* process with the generalized rate function $\Lambda(x, t)$, where $x \in \mathbb{R}^2$ is the spatial component and $t \in \mathbb{R}^+$ is the time component, and

$$\Lambda(x, t) = \frac{\lambda}{S_R}, \text{ if } x \in \mathbf{B}, \tag{1}$$

where $S_R = \pi R^2$ is the area of the macro cell. The function $\Lambda(x, t)$ has the meaning of the frequency with which the events are expected to occur in the unit of $\mathbb{R}^2 \times \mathbb{R}^+$.

Assumption 3. ***Signal propagation.*** *The wireless channel gain $\gamma_{i,j}$ between the user U_i and the AP/BS j depends on the distance $d_{i,j}$ separating them. Therefore, following, e.g., [38], we assume that for the session i the channel gain (path gain) $\gamma_{i,j}$ is expressed as a power function of distance:*

$$\gamma_{i,j} = \frac{G}{{d_{i,j}}^{\kappa}}, \tag{2}$$

where $d_{i,j}$ is the distance between the AP/BS and the transmitting user, κ is the propagation exponent, and G is the propagation constant. The parameters κ and G are determined by the particular radio access technology and account for the corresponding channel model.

To clarify the above assumption, we emphasize that the models in [38] contain the line-of-sight (LOS) and non-line-of-sight (NLOS) components triggered by a Bernoulli-distributed random variable corresponding to the LOS case. Typically, those models may be easily fitted into our power model (2), where the constants G and κ would correspond to the target scenario- and technology-related parameters.

Since our model is primarily intended for characterizing the long-term average system operation without excessive user mobility, the impact of fast fading is averaged out by the use of forward error correction, and therefore has minimal (and, nearly constant) effect on performance. The effects of slow fading have most of their influence on the link budget, rather than the variability in the link quality, and in our methodology these are captured by the appropriate propagation model. Additional considerations on fading are given in Appendix A by introducing a random component into the link budget. Along these lines, to accommodate both types of fading in our analysis, we introduce the corresponding *fading margin* η, which is then added to the link budget.

In terms of the achievable accuracy, it might have been beneficial to include the fading effects into the model explicitly. On the other hand, this would require significant complication of the analysis at hand, primarily to accommodate such functions as HARQ, MCS selection, and closed-loop power control. All this would make strict analysis not feasible, leading to even more approximations and assumptions. Therefore, the channel gain γ remains the most determining factor in defining the relation between the transmit power and the achievable data rate for the considered link between the user and the AP/BS.

Without the loss of generality, we further assume that the data rate is continuous and that the power-rate mapping is defined by the Shannon's formula. This consideration has recently been shown in [39] to remain very accurate for current wireless networks.

Assumption 4. ***Power-rate mapping.*** *The transmit power p_i of a user and its corresponding data rate r_i (in [nats/s]) are coupled by the generalized Shannon's formula:*

$$r_i = B \log\left(1 + Ap_i\right), \tag{3}$$

where p_i is the output power of the radio frequency power amplifier, whereas A and B are the scaling coefficients dependent on the particular wireless technology used. For the *sake of an example, these are given as:*

$$A = \frac{\eta \gamma_{i,i}}{N_0 + I}, \quad B = w, \tag{4}$$

where $\gamma_{i,i}$ is the path gain between the user and the AP/BS for session i, η is the fading margin, w is the channel bandwidth, N_0 is the noise level, and I is the interference level at the receiver.

In order to constrain the growth of the data rate due to the fixed set of modulation and coding schemes, we assume that the rate remains constant r_{lim} when $d < d_0$. The latter effectively means that further increase in the SNR does not yield the unbounded data rate growth after a certain value of d_0. The parameter d_0 may be written as:

$$d_0 = \left[\frac{G \cdot p}{(N_0 + I)\left(e^{r_{lim}/w} - 1\right)} \right]^{1/\kappa}. \tag{5}$$

While random network topology is the primary focus of our model, we also investigate flow-level system dynamics. This involves an appropriate queuing model, where the session arrives and leaves the system after being served (the service time is determined by the random session length). When a new session arrives or a served session leaves the system, the centralized assisting entity in the cellular network performs *admission* and *power control* on all tiers by deciding whether the session would be admitted to a particular tier or not (admission control) and/or advising on the user's transmit power (power control).

Assumption 5. *Admission control.* *Every real-time session requires the target bitrate of r_0. Therefore, the system admits a newly arrived session if there is still sufficient resource to serve it. In other words, each ongoing session i has to occupy exactly r_0/r_i-fraction of the system time (where the overhead is accounted for later), while for all the active sessions it holds the following:*

$$\sum_{all\ sessions} \left(\frac{r_0}{r_i} \right) \leq \delta, \tag{6}$$

where δ is the available resource at a particular AP/BS (e.g., excluding resources allocated for fading compensation), $r_i \leq r_i^{max}$ is the instantaneous data rate depending on the distance between the user and the receiving AP/BS, and r_i^{max} is the maximum achievable data rate at the maximum power level.

Additionally, admission control policy may determine whether current interference exceeds a given threshold or not.

Assumption 6. *Interference assessment.* *Further, it is imposed that a tier with $n-1$ active users admits a new session/user n if for the set $\{U_i\}_{i=1}^n$ of all users the following condition holds at each BS/AP A_j:*

$$r_i \geq r_0 \text{ and } p_i \gamma_{i,j} \leq N_0, \quad \forall j, i \neq j, \tag{7}$$

where $\gamma_{i,j}$ is the path gain between the user and the AP/BS j and p_i is the corresponding allocated power.

By examining the expression $p_i \gamma_{i,j}$, we ensure that the user does not cause interference higher than the (modified) noise level for the considered radio technology (see [40] for further discussions). Rephrasing the above, admission control ensures that the required minimum bitrate can be achieved by a user, and that the interference at the AP/BS A_i produced by the user U_j does not exceed a given threshold. This threshold is highly dependent on the technology features and is discussed below separately for the WLAN and the Pico tiers, whereas no interference is assumed on the Macro tier.

Assumption 7. *Interference boundary.* *We also assume that the noise plus interference has the form of $N_0 + I = KN_0$, where the value of K is a scaling parameter fixed across the network, which has the meaning of the interference margin per AP/BS.*

In reality, the pico BSs attempt to limit the interference of their admitted users onto the neighboring BSs by means of various control mechanisms (dynamic scheduling, link adaptation, etc.). Therefore, the following engineering technique might be feasible. We may aggregate the individual interferences created by the proximate users of a particular tier into a cumulative background noise level, which in the practice of network planning is taken into account as a particular interference margin.

We note here that our interference and rate estimation has predictive character and assists the network in making a decision on whether a user should be admitted or not. Alternatively, if a session cannot be admitted on a particular tier, it is considered *blocked* with the probability $P_{block}^{(1/2/3)}$.

Further, we differentiate between two alternative joint strategies for *power control* and *scheduling*, which are termed the *maximum power* policy and the *round robin* policy.

Assumption 8. *Power control and scheduling.* *The considered policies are the following.*

1. The maximum power policy sets a fixed transmit power which is the allowable maximum for a particular radio technology. Then, admission control checks if the target bitrate can be achieved with this maximum power.

2. The round robin policy ensures that the system resource is always shared between the users equally and, therefore, employs another power and admission control. Each admitted session out of n running sessions is allocated an equal portion of the total time resource, i.e., $\frac{r_0}{r_i} = \frac{1}{n}$. Then, the users adjust their transmit power to match their required target bitrate as long as it does not exceed the maximum allowed power level.

B. System Operation and Metrics

In this work, we concentrate on a particular (heuristic) mechanism for user admission and network selection. The example of system operation is illustrated in Figure 3. We consider the following *cascade* service when a new session arrives into the system. First, the network selection assistance entity residing on the cellular network attempts to offload the newly arrived session onto the *nearest* WLAN AP by performing the *WLAN admission control* managed centrally. We note that the nearest AP may also be located outside of the circle R.

If the session is accepted on the WLAN tier, it is served there without interruption until when it successfully leaves the system. Otherwise, if this session cannot be admitted onto the WLAN, the pico network admission control is executed. Hence, either the session is accepted on the Pico tier and served by the *nearest* pico BS or the macro network itself attempts to serve this session. Eventually, if the session cannot be admitted onto the Macro tier either, it is considered *permanently blocked* and leaves the system unserved without any impact on the new arrivals.

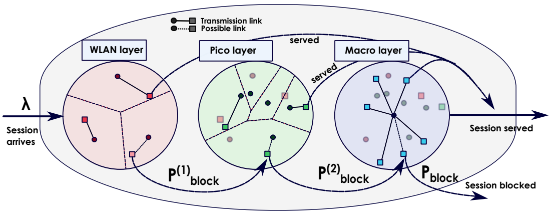


Fig. 3. Consecutive system operation: WLAN, Pico, and Macro tiers.

The proposed sequential *network-assisted* admission results in a good compromise between fully-distributed (uncoordinated) and centrally-controlled solutions. Whereas the former may be substantially sub-optimal, the latter may result in excessive computational complexity and associated signaling overheads. By contrast, our cascade model is capable of packing network capacity sequentially and thus becomes an attractive candidate for future HetNet deployments. We thus leave the consideration of other (network-optimal) policies out of scope of the paper, focusing in what follows only on some methodological aspects to evaluate such.

In addition, the choice of the cascade service in Figure 3 is desirable in terms of the associated user/operator costs. Due to the fact that WiFi connectivity is often available free of charge, it is always beneficial to offload as many sessions as possible onto the WLAN tier as long as user QoS remains acceptable (which in our model is ensured by the respective network-assisted admission control). When the WLAN tier is enough loaded and is not able to accept more users, the second preferable candidate is the Pico tier. This is due to the operator's desire to balance the load across small cells for users with low mobility. The conventional macro network service would then be left by the operators as a fall-back option for their users (or for highly-mobile users not considered here).

Whenever admitted, a transmitting user exploits a fraction of the system time resource and sets its power as commanded by the power control module to achieve its required data rate. The system makes a new decision on scheduling and transmission power allocations for all the active users at every new arrival or when an existing session is served and leaves the system.

For each tier, we introduce the corresponding blocking probability $P_{block}^{(i)}$ and acceptance probability $P_a^{(i)} = 1 - P_{block}^{(i)}$, where $i = 1, 2, 3$ are the indexes of the WLAN, Pico, and Macro tiers, respectively. Moreover, we remind that the session arrival rate on the (first) WLAN tier is $\lambda_w = \lambda$ (see Assumption 2).

Assumption 9. ***Decoupling assumption.*** *To preserve analytical tractability of our mathematical model, we assume that all three network tiers serve their users independently, which results in a random thinning of the arrival process with the corresponding acceptance probabilities.*

The above assumption is a natural methodological move to decompose the system into a set of tractable and well-defined components, which may be easily replaced and/or interchanged (e.g., should one ever decide to modify the priority of tiers in the admission control procedure).

Proposition 1. *Due to Poisson property of the thinned flow, the arrivals on the (second) Pico tier (those not accepted by the WLAN tier) follow a Poisson process of density* $\lambda_p = \lambda\left(1 - P_a^{(1)}\right)$*, where* $P_a^{(1)}$ *is the WLAN tier accept probability. Similarly, the arrivals on the (third) Macro tier adhere to a Poisson process of density* $\lambda_m = \lambda\left(1 - P_a^{(1)}\right)\left(1 - P_a^{(2)}\right)$*, where* $P_a^{(2)}$ *is the Pico tier accept probability.*

Abstracting away the locations of users for analytical tractability, we assume that the arrivals on the Pico and the Macro tiers are also placed uniformly within the circle of radius R. In contrast to Proposition 1, the latter consideration *does not* hold in reality. Instead, there is some pattern in which users are taken for service by the WLAN and the Pico tiers. However, our simulation results (as reported in Section V) reveal that the assumption of uniformity is surprisingly accurate. This makes the analysis of our system under the above mentioned assumptions to be an adequate approximation for the practical HetNet operation.

Consequently, the overall system *blocking* probability P_{block} may be established as follows:

$$P_{block} = 1 - \left[P_a^{(1)} + \left(1 - P_a^{(1)}\right)P_a^{(2)} + \left(1 - P_a^{(2)}\right)\left(1 - P_a^{(1)}\right)P_a^{(3)}\right], \tag{8}$$

where $P_a^{(3)}$ is the macro cell accept probability.

Below we detail the distinguishing features of each of the three tiers in the order of system operation: WLAN, Pico, and Macro tiers, respectively.

C. WLAN Tier Model

Here we consider the WLAN tier comprising a random number N_w of WiFi APs $A_j, j = \overline{1, N_w}$ placed according to Assumption 1, i.e., uniformly on the plane.

WLAN tier follows the maximum power policy as discussed in Assumption 8, which means that a user sends its data at the maximum allowed transmit power level. Due to the fact that the WLAN system is inherently interference-limited, we employ the Shannon's capacity theorem (3) as the power-rate mapping and take into account the current number of users associated with an AP. Then, the instantaneous data rate for the session i is determined by the maximum transmit power $p_{\max}$ as:

$$r_i = r_i^{max} = w \log\left(1 + \text{SINR}_i\right) = w \log\left(1 + \frac{\eta p_{\max}\gamma_{i,i}}{KN_0}\right),$$

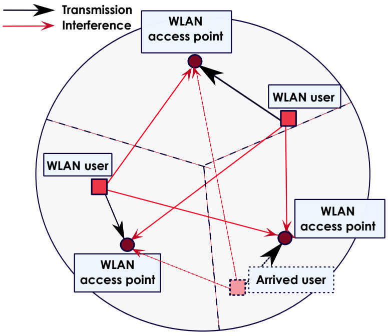


Fig. 4. WLAN tier operation.

where N_0 is the fixed noise power level and K is interference-related parameter as described above.

Therefore, the system admits a newly arrived user and associates it with a particular AP if the following condition holds:

$$\sum_{all\ AP\ sessions} \left(\frac{r_0}{r_i^{max}} \right) \leq \delta_w,$$

where δ_w is the actually available resource for the AP-centered cluster after removing the overheads of the random-access based MAC (associated with signaling and contention, see details in [41]) and r_{max} is the highest achievable data rate at the maximum power level.

In addition, and specifically for the WLAN interference control, we note that due to the Clear Channel Assessment (CCA) function we may refer to the noise level as to the CCA threshold. Therefore, the WLAN admission control policy examines whether the interference at each neighboring AP (produced by the extra power $p_{\max}$ from a newly arrived user) does not exceed the given noise threshold N_0, and the respective $K = 1$.

D. Pico Tier Model

Further, we consider the Pico tier comprising a random number N_p of pico BSs P_j, $j = \overline{1, N_p}$ placed uniformly according to Assumption 1.

We remind that all the pico cell connections operate on the same frequency, which is assumed here without loss of generality since a new frequency may be added to our system as an additional Pico tier. Therefore, similarly to the WLAN tier, we take into account the interference from admitted users at the neighboring pico BSs.

The transmit power p_i of a user and its data rate r_i are also coupled by the Shannon's capacity theorem for interference-limited environment:

$$r_i = w \log \left(1 + \mathrm{SINR}_i\right) = w \log \left(1 + \frac{\eta p_i \gamma_{i,i}}{K N_0}\right).$$

On the Pico tier, we mimic the operation of the Open Loop Power Control (OLPC) procedure (see [42], Section 5.4.1.3) by assuming that the centralized entity estimates the system parameters and advises on the appropriate transmit power values p_i accordingly. The estimation is based on both power and interference control as follows.

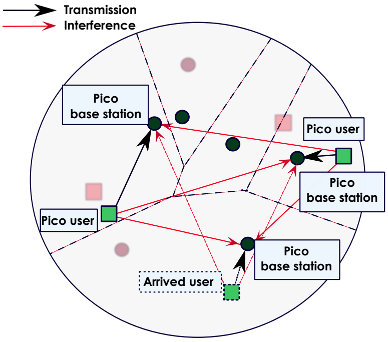


Fig. 5. Pico tier operation.

The Pico tier employs the following interference model. As an upper bound on the interference level at the pico BS, we assume the maximum average interference level of a highly-loaded pico BS, thus parameterizing our model. Being a function of the number of neighboring pico BSs, the interference level in question is considered to be independent of the number of users. This relaxes the interference check at the admission control stage and yields the worst-case interference estimate.

By contrast to the WLAN tier, the Pico tier employs the round robin power control/scheduling policy. Therefore, in order to admit a new session number n_0, the pico BS has to increase the power of the already running transmissions, such that they would still fit into their smaller time allocations. If it is not possible for at least one of the n_0 active sessions (including the new session) at any particular pico BS, that is, $r_i^{max} = w \log \left(1 + \frac{\gamma_i \eta p_{\max}}{K N_0}\right) < n_0 r_0$, then a newly arrived user cannot be admitted onto the Pico tier. Otherwise, the system time is re-allocated for n sessions and users employ other (higher) transmit power levels:

$$p_i = \frac{1}{\eta \gamma_i} \left(e^{n_0 r_0 / w} - 1\right) (K N_0),$$

where $p_i \leq p_{\max}$ and $K N_0$ is the power of noise plus interference from the neighboring pico BSs.

Therefore, the Pico tier only admits a newly arrived session if there are still sufficient resources to serve it. In other words, each ongoing session at a particular pico BS has to occupy exactly $1/n_0$-fraction of time frame duration, while for all the sessions the following holds:

$$\sum_{all\ BS\ sessions} \left(\frac{r_0}{r_i^{\max}} \right) = \delta_p,$$

where δ_p is the total available resource for the pico BS-centered cluster.

E. Macro Tier Model

Finally, we consider the Macro tier with a single serving BS which is exempt from the inter-cell interference (which is similar to treating the interference coming from the neighboring macro cells as the increased background noise). This formulation implies interference-free communication as long as uplink user transmissions are orthogonal by network design.

Hence, by contrast to the WLAN and the Pico tiers, Assumption 4 transforms into the following (simplified) power-rate mapping:

$$r_i = w\log\left(1+SNR_i\right) = w\log\left(1+\frac{\eta\gamma_i}{KN_0}p_i\right),$$

where SNR is the signal-to-noise ratio, γ_i is the channel gain between the user and the macro BS, and N_0 is the noise level.

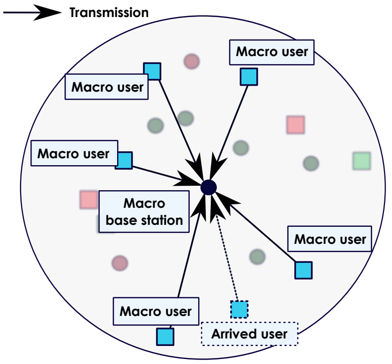


Fig. 6. Macro tier operation.

Following the operation of the OLPC scheme, the admission and power control on the Macro tier are performed similarly to those on the Pico tier where each admitted session is allocated an equal portion of the total system time:

$$\frac{r_0}{r_i} = \frac{\delta_m}{n}, r_i = r_0 n, \quad \forall i = \overline{1,n},$$

where δ_m is the actually available resource at the macro BS.

However, due to the absence of interference, the Macro tier admission control should disregard any interference-related considerations. Therefore, a newly arrived session can only be admitted by the Macro tier *iff*:

$$r_i^{max} = w\log\left(1+\frac{\eta\gamma_i p_{\max}}{KN_0}\right) \geq n\frac{r_0}{\delta_m}.$$

If the above condition holds, the system time is again re-allocated for n running sessions and the users' transmit powers are set as $p_i = \frac{KN_0}{\eta\gamma_i}\left(e^{nr_0/(w\delta_m)}-1\right) \geq p_{\max}$. Otherwise, the candidate session is considered to be permanently blocked and leaves the system unserved.

IV. Analysis of Random Dynamic HetNets

In this section, we provide a summary of our analytical efforts to evaluate the important HetNet performance metrics. Hereinafter, we consider three different tiers separately. We underline here that our system analysis is built on the *decoupling principle* as per Assumption 9. This technique is used widely and allows evaluating even very complex systems by regarding them as an integrated set of tractable problems.

A. Stochastic Model

Here, we outline our general stochastic model for the WLAN, Pico, and Macro tiers based on the assumptions of Section III. Assume that the arrivals on the three tiers follow a Poisson process with the rates $\lambda_w = \lambda$, λ_p, and λ_m, respectively. We observe the Macro tier at the particular moments t of session (user) arrivals/departures. Since the arrivals follow a Poisson process and the service (session length) is distributed exponentially, our system behavior may be represented as a stochastic Markov process $S(t)$, where the future process evolution is determined by the set of the ongoing sessions that are currently served on a given tier.

1) Macro Tier Analysis: For the Macro tier, the state of the process $S(t)$ is determined by the characteristics of the ongoing sessions on the macro cell. For convenience, we denote these abstract characteristics as ω and note that they depend on the location of the user. Therefore, the system state is represented by the vector $(\omega_1,...,\omega_n)$, where n is the number of sessions in service (see Figure 7).

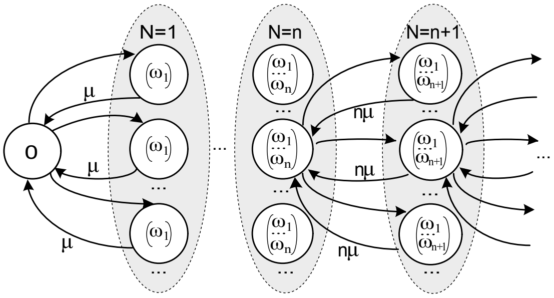


Fig. 7. State diagram for the Macro tier.

Let the Macro tier have n running sessions in the state s. We denote the probability of rejection at the state s for the newly arrived session as $Q_{n+1|s}$. Then, transitions from the state $s=(\omega_1,...,\omega_n)$ to the state $(\omega_1,...,\omega_n,\omega_{n+1})$ and backwards have the rates of $\lambda_m\left(1-Q_{n+1|s}\right)$ and $(n+1)\mu$, respectively.

2) WLAN and Pico Tiers Analysis: Since the WLAN and the Pico tiers are both interference-limited, we provide the following reasoning for the two of them jointly. This is because the respective stochastic processes have identical state-related properties.

The WLAN/Pico tier comprises several APs/BSs which are distributed on the plane with the densities of L_w and L_p. Therefore, the state of the stochastic Markov process $S(t)$ may be represented by the set of sessions with respect to the corresponding APs/BSs. Similarly, we adopt notation ω for the session characterization. Then, the state s of the WLAN/Pico tier is represented by:

$$(\omega_1,...,\omega_{n_1};\omega_{n_1+1},...,\omega_{n_1+n_2};...;\omega_{s_n+1},...,\omega_{s_n+n_{N_{w/f}}}),$$

where $s_n = \sum_{i=1}^{N_{w/f}-1} n_i$, as well as n_1, n_2, and $n_{N_{w/p}}$ are the numbers of users associated with the first, the second, and the last AP/BS, respectively. The random variable $N_{w/p}$ corresponds to the number of APs/BSs in a certain area and follows a Poisson distribution. The state diagram of the considered system is illustrated in Figure 8.

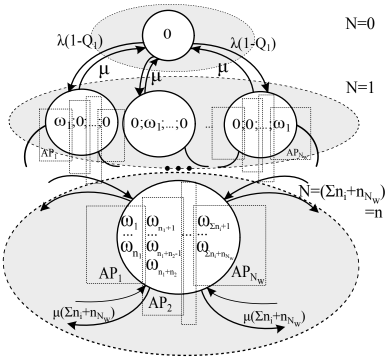


Fig. 8. State diagram for the WLAN/Pico tier.

We consider state s, where the WLAN/Pico tier is serving n ongoing sessions with a random number of N_w/N_p APs/BSs. We denote the probability of rejection for the newly arrived session as $Q_{n+1|s}$. Then, the transitions from the state s to the state of $n+1$ active sessions have the rate of $L_{w/p}\left(1-Q_{n+1|s}\right)$. The backward rate equals $(n+1)\mu$ since the service does not depend on the state, but rather on the number of simultaneously served sessions.

B. Steady-State Distribution

Due to the uncountable number of states in the considered system, it may be complicated to attack the steady-state distribution straightforwardly. However, the corresponding Markov process may be simplified by employing the *state aggregation* technique.

Assumption 10. ***State aggregation.***

1. For the Macro tier, we aggregate all the states containing n sessions into the unifying state n, regardless of the actual locations of users.

2. For the more complex WLAN and Pico tiers, we combine all possible states of the system (which contain n ongoing sessions) into the state n, regardless of the locations of the current users or their connections to a certain AP/BS. The described combining process is illustrated in Figure 8.

3. In order to keep our system memoryless, we adopt a simplification, where the sessions at the state n, while keeping all of their other properties, do not preserve their locations from state to state. For the sake of analytical tractability, these locations are assumed to be generated anew at every particular state n.

We note, however, that the system still keeps track of the previously admitted sessions owing to the probabilities Q_{n+1} to reject the session arrived at the state n conditioning on the fact that the current nth session satisfies the admission control criteria.

As the result of the state aggregation, we arrive at the birth-death processes for all the three tiers with the rates of $\lambda_{m/w/p}\left(1-Q_{n+1}\right)$ and $(n+1)\mu$. We further denote the arrival rate $\lambda_{m/w/p}$ into the system simply as λ and formulate the following proposition.

Proposition 2. *The steady-state distribution $\{\pi_i\}_{i=0}^{\infty}$ for the considered process $S(t)$ with the transitions $\lambda\left(1-Q_{n+1}\right)$ and $(n+1)\mu$ can be closely approximated by:*

$$\pi_n=\pi_0\frac{\lambda^n_{m/w/p}}{\mu^n}\frac{\prod_{i=1}^{n}\left(1-Q_n\right)}{n!},\tag{9}$$

where

$$\pi_0=\left(\sum_{i=0}^{\infty}\frac{\lambda^n_{m/w/p}}{\mu^n}\frac{\prod_{i=1}^{n}\left(1-Q_n\right)}{n!}\right)^{-1},$$

and Q_{n+1} is the reject probability on the transition from the state n to the state $n+1$.

Proof. With the above expressions, we refer to the steady-state distribution for the well-known $M/M/c$ system, the derivation of which may be found in the corresponding literature. ■

Basing on the steady-state distribution and assuming that it exists, our approach empowers us to estimate a wide class of stationary characteristics in the considered system, such as the expected number of ongoing sessions, the probability of session's permanent blocking, or even its energy consumption. However, the latter is left out of scope of this paper due to space limitations. The average number of active sessions and the system blocking probability are defined as:

$$E[N]=\sum_{n=0}^{\infty}n\pi_n,\quad P_{block}=\sum_{n=0}^{\infty}Q_{n+1}\pi_n.\tag{10}$$

In our analysis, we disregard the history of the system processes from the perspective of the ongoing sessions. We thus replace the initial stateful systems with memoryless processes for which we examine the arbitrary set of respective random variables at each point t. If the reject probabilities Q_{n+1} are known for all tiers, we easily obtain the steady-state distribution by using (9). Therefore, in what follows we concentrate on calculating the values of Q_{n+1}. In order to take into account the memory property that we have thus omitted, we will refer to the corresponding conditional probabilities further on.

C. Characterizing Transitions

We continue by evaluating the probabilities Q_{n+1} and the transition rates $\lambda_{m/w/p}\left(1-Q_{n+1}\right)$ necessary for estimating the steady-state distribution. This would enable us to calculate the important stationary characteristics, such as the overall session blocking probability and the average number of simultaneously running sessions. The latter metric may also be used as the system (area) capacity prediction for sufficiently high arrival rate.

The rest of the text is organized in the following order. First, we illustrate our approach on the simplest Macro tier operation. Then, we continue with the more complex (due to the presence of interference) WLAN tier. Finally, we discuss the Pico tier which inherits some properties of both Macro and WLAN tiers.

1) Macro Tier Transitions: We begin with the Macro tier and detail the calculations which are necessary for characterizing the round robin policy. Hence, the transitions from the state n to the state $n+1$ are defined by:

$$\lambda_m\left(1-Q_{n+1}\right)=$$
$$=\lambda_m\left(\Pr\left\{\frac{r_0}{r_i^{max}}\leq\frac{\delta_m}{n+1},\forall i=1,n+1|\frac{r_0}{r_i^{max}}\leq\frac{\delta_m}{n},\forall i=1,n\right\}\right), \tag{11}$$

where by the above conditional probability we account for the previous system history, while the new system evolution process (n-based) is memoryless. In other words, we estimate the probability to share the resource between $n+1$ random sessions if n other stochastically different sessions have already been admitted at the previous state.

Next, we calculate this important probability basing on the expression above:

$$\Pr\left\{\tfrac{r_0}{r_i^{\max}}\leq\tfrac{\delta_m}{n+1},\forall i=1,(n+1)|\tfrac{r_0}{r_i^{\max}}\leq\tfrac{\delta_m}{n},\forall i=1,n\right\}=$$
$$=\Pr\left\{r_i^{\max}\geq\tfrac{r_0}{\delta_m}(n+1),\forall i=1,(n+1)|r_i^{\max}\geq\tfrac{r_0}{\delta_m}n,\forall i=1,n\right\}.$$

We may further decompose this expression into parts by separating new and ongoing sessions:

$$\Pr\left\{r_{n+1}^{\max}\geq\tfrac{r_0(n+1)}{\delta_m}\right\}\prod_{i=1}^{n}\Pr\left\{r_i^{\max}\geq\tfrac{r_0(n+1)}{\delta_m}|r_i^{\max}\geq\tfrac{r_0 n}{\delta_m}\right\}=$$
$$=\Pr\left\{r_{n+1}^{\max}\geq\tfrac{r_0(n+1)}{\delta_m}\right\}\left(\frac{\Pr\left\{r_i^{\max}\geq\frac{r_0}{\delta_m}(n+1)\right\}}{\Pr\left\{r_i^{\max}\geq\frac{r_0 n}{\delta_m}\right\}}\right)^n. \tag{12}$$

The probabilities $\Pr\left\{r_i^{\max}\geq\frac{r_0}{\delta_m}(n+1)\right\}$, $\Pr\left\{r_i^{\max}\geq\frac{r_0}{\delta_m}n\right\}$ are based on the distribution $F_r(r)$ of the random variable $r_i^{\max}$ and are described in Appendix. Given these probabilities, we may easily calculate the transition rate and, therefore, the steady-state distribution (9) as well as the relevant stationary metrics (10).

2) WLAN Tier Transitions: We continue by considering the WLAN tier and detail the calculations which are necessary for characterizing the maximum power policy. We remind that for the WLAN tier, the system admits a new session if both conditions hold: the bitrate is not less than the target r_0 and the interference to other APs is not greater than the given threshold N_0. Therefore, the transitions from the state n to the state $n+1$ are defined by:

$$\lambda_w(1-Q_{n+1})=\lambda_w\Pr\left\{A_j^{(n+1)},j=\overline{1,n+1}|A_j^{(n)},j=\overline{1,n}\right\}, \tag{13}$$

where event $A_j^{(n)}$ is given as:

$$A_j^{(n)}=\left\{\frac{r_0}{r_j^{\max}}\leq\delta_w-\sigma_n \text{ and } \gamma_{j,k}p_{\max}\leq N_0,\forall k\neq j\right\},$$

where δ_w is a share of the available resource at the AP (without the signaling overhead and collisions) and σ_n is a part of the resource given to other sessions at the same AP in the current state. We further denote $r_0/(\delta_w-\sigma_n)$ as $\tilde{r}_{0,n}$. The calculation of σ_n is given in Appendix separately for the WLAN and the Pico tiers and is based on the following assumption.

Assumption 11. ***AP link abstraction.*** *Here, to abstract away the session-AP details at the state n, we assume that upon its arrival into the system, a session observes the average (typical) number of users at the nearest AP (see Theorem 1). This average number depends on the number of ongoing sessions, i.e., on the state index n as well as on the parameter* $\tilde{r}_0$.

We also emphasize that the APs are distributed on the plane according to the PPP. Therefore, the number k of other than the nearest APs can be any large. For any session, we enumerate the APs in the order of increasing distance, so that $k=1$ denotes the closest one. We remind that the rate condition has to hold for the nearest AP and the interference condition should hold for the others. Hence, the transition rates may be calculated as:

$$(1-Q_{n+1})=\frac{\Pr\left\{A_j^{(n+1)},j=\overline{1,n+1}\right\}}{\Pr\left\{A_j^{(n)},j=\overline{1,n}\right\}},$$

where we denote $\left\{A_j^{(n)},j=\overline{1,n}\right\}$ as $\{\mathbf{A}_n\}$.

Therefore, the probability $\Pr\{\mathbf{A}_n\}$ may be expressed as:

$$\Pr\left\{r_i^{\max}\geq\tilde{r}_{0,n} \text{ and } \gamma_{i,k}p_{\max}\leq N_0, i=\overline{1,n},k>1\right\}.$$

We also assume that the values of $\gamma_{i,j}$ are independent. Then, we may continue as:

$$\Pr\{\mathbf{A}_n\}=\prod_{i=1}^{n}\Pr\left\{r_i^{\max}\geq r_{0,n}\right\}\prod_{i=1}^{n}\Pr\{\gamma_{i,k}p_{\max}\leq N_0,k>1\}=$$
$$=\prod_{i=1}^{n}\Pr\left\{\gamma_{i,i}\geq\tfrac{KN_0}{\eta p_{\max}}\left(e^{\frac{\tilde{r}_{0,n}}{w}}-1\right)\right\}\prod_{i=1}^{n}\Pr\left\{\gamma_{i,k}p_{\max}\leq N_0,k>1\right\}.$$

Further, we denote the corresponding $\gamma_{i,k}$ as γ_k, such that γ_1 is the path gain to the nearest AP over the distance d_1. We assume an identical distribution of γ_1 for all sessions i, as well as any γ_k. Hence, we establish:

$$\Pr\{\mathbf{A}_n\}=[\Pr\{d_1\leq d_{r,n}\}]^n\left[\Pr\left\{\gamma_k\leq\frac{N_0}{p_{\max}},k>1\right\}\right]^n,$$

where the constant value $d_{r,n}$ is $\left(\frac{p_{\max}\eta G}{KN_0}\right)^{\frac{1}{\kappa}}\left(e^{\frac{\tilde{r}_{0,n}}{w}}-1\right)^{-\frac{1}{\kappa}}$.

We note that the sought probability can be obtained via the relationship between the path loss γ and the distance d which is:

$$d=(G/\gamma)^{1/\kappa}, \tag{14}$$

and, therefore, we have the following:

$$\Pr\left\{\gamma_k\leq\frac{N_0}{p_{\max}},k>1\right\}=\Pr\{d_k\geq d_{thr},k>1\}=\Pr\{d_2\geq d_{thr}\},$$

where $d_{thr}=\left[\frac{Gp_{\max}}{N_0}\right]^{\frac{1}{\kappa}}$. Then, after substitution, we arrive at:

$$\Pr\{\mathbf{A}_n\}=\left[F_{d_1}\left(d_{r,n}\right)\right]^n\left[\Pr\left\{d_2<d_{thr}\right\}\right]^n.$$

We emphasize that the expression $d_2>d_{thr}$ holds *iff* zero or one AP is located within the circle of radius d_{thr} around the tagged user. Then, the probability $\Pr\{d_2\geq d_{thr}\}$ can be established through the Poisson distribution. Therefore, basing on the distributions of d_1 and d_2 (see (24), (25) in Appendix for details), we obtain the following:

$$\Pr\{\mathbf{A}_n\}=\left[1-e^{-\pi L_w d_r^2}\right]^n\left[L_w\pi d_{thr}^2 e^{-L_w\pi d_{thr}^2}+e^{-L_w\pi d_{thr}^2}\right]^n.$$

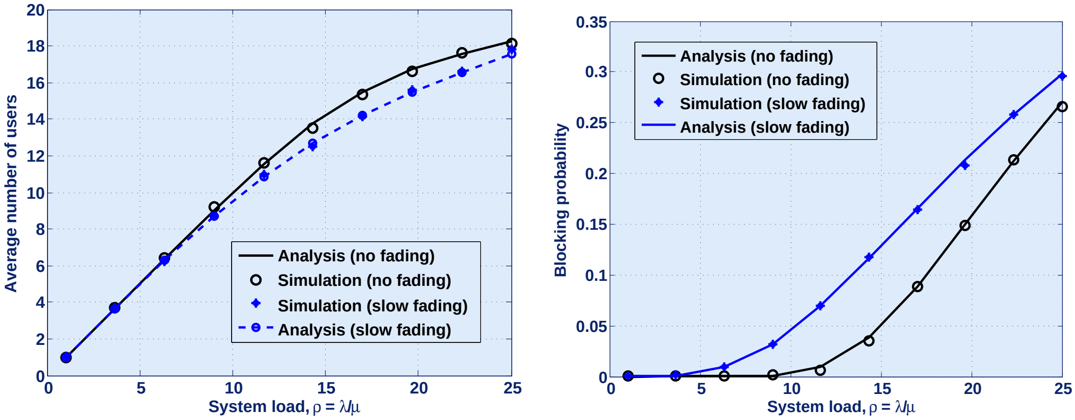


Fig. 9. Comparing simulation and analysis for the Macro tier with fading: expected number of sessions (left) and blocking probabilities (right).

Finally, we calculate rate transitions (13) as the ratio $\Pr\left\{\mathbf{A}_{n+1}\right\}/\Pr\left\{\mathbf{A}_{n}\right\}$ and hence:

$$1-Q_{n+1}=\frac{\left(1-e^{-\pi L_w d_{r,n+1}^2}\right)^{n+1}}{\left(1-e^{-\pi L_w d_{r,n}^2}\right)^{n}}\left(L_w \pi d_{thr}^2 e^{-L_w \pi d_{thr}^2}+e^{-L_w \pi d_{thr}^2}\right). \tag{15}$$

In summary, by introducing $d_{r,n}$, we emphasize that it depends on $\tilde{r}_{0,n}=r_0/(\delta_w-\sigma_n)$, which in turn is a function of the number of sessions on the WLAN tier via the occupied resource σ_n. The calculation of σ_n (33) as well as other necessary parameters is given in Appendix. The expression (15) enables us to derive the key metrics of interest, such as the expected number of ongoing sessions and overall blocking probability (10).

3) Pico Tier Transitions: We conclude by considering the Pico tier and providing the necessary calculations. Similarly to the Macro tier with the round robin policy, the transitions from the state n to the state $n+1$ on the Pico tier are given by:

$$\lambda_p(1-Q_{n+1})=\lambda_p \Pr\left\{A_j^{(n+1)}, j=\overline{1,n+1} | A_j^{(n)}, j=\overline{1,n}\right\}. \tag{16}$$

The event $A_j^{(n)}$ is a combination of two events, i.e., the rate condition has to hold for the nearest BS and the interference condition has to be satisfied for all others. Therefore, if the pico BSs are enumerated in the order of increasing distance to the transmitting user j, then:

$$A_j^{(n)}=\left\{r_j^{\max}\geq\frac{r_0 n_0}{\delta_p} \text{ and } \gamma_{j,k}p_j\leq N_0, k>1\right\},$$

where n_0 is the number of users associated with the nearest BS, δ_p is a share of the available resource at the BS, and σ_n is a part of the resource given to other sessions at the same BS in the current state (as above). The allocated power p_j may be calculated as:

$$p_j=\frac{KN_0}{\eta\gamma_{j,j}}\left(e^{\frac{r_0 n_0}{w\delta_p}}-1\right). \tag{17}$$

Similarly to what happens on the WLAN tier, the BSs are distributed on the plane according to the PPP and the number k of other than the nearest BSs can be any large for $k=\overline{1,\infty}$.

Denoting $\left\{A_j^{(n)}, j=\overline{1,n}\right\}$ as $\{\mathbf{A}_n\}$, we establish:

$$\Pr\left\{\mathbf{A}_n\right\}=\Pr\left\{r_i^{\max}\geq\frac{r_0}{\delta_p}n_0, \gamma_{i,k}p_i\leq N_0, i=\overline{1,n}, k>1\right\}.$$

We assume that the values of $\gamma_{i,j}$ are independent. Then, we may write:

$$\Pr\left\{\mathbf{A}_n\right\}=\prod_{i=1}^{n}\Pr\left\{r_i^{\max}\geq\tfrac{r_0}{\delta_p}n_0\right\}\prod_{i=1}^{n}\Pr\left\{\gamma_{i,k}p_i\leq N_0, k>1\right\}=$$
$$=\prod_{i=1}^{n}\Pr\left\{\gamma_{i,i}\geq\tfrac{KN_0}{\eta p_{\max}}\left(e^{\frac{r_0 n_0}{w\delta_p}}-1\right)\right\}\prod_{i=1}^{n}\Pr\{\gamma_{i,k}p_i\leq N_0, k>1\}.$$

Again, we denote the corresponding $\gamma_{i,k}$ as γ_k, such that γ_1 is the path gain to the nearest BS over the distance d_1. Hence, we obtain:

$$\Pr\left\{\mathbf{A}_n\right\}=\left[\Pr\left\{d_1\leq d_{r,n}\right\}\right]^n\left[\Pr\left\{\gamma_k\leq\tfrac{N_0}{p_i}, \forall k>1\right\}\right]^n,$$

where $d_{r,n}=\left(\frac{p_{\max}\eta G}{KN_0}\right)^{\frac{1}{\kappa}}\left(e^{\frac{r_0 n_0}{w}}-1\right)^{-\frac{1}{\kappa}}$. Finally, we have:

$$\Pr\left\{\mathbf{A}_n\right\}=\left[F_{d_1}\left(d_{r,n}\right)\right]^n\left[\Pr\left\{\gamma_2\leq\frac{\eta\gamma_1}{\left(e^{\frac{r_0 n_0}{w\delta_p}}-1\right)}\right\}\right]^n.$$

The details regarding the components of the expression above are given in Appendix. Based on them, the transition rates may be obtained as follows:

$$\lambda_p\left(1-Q_{n+1}\right)=\lambda_p\frac{\Pr\left\{\mathbf{A}_{n+1},\mathbf{A}_n\right\}}{\Pr\left\{\mathbf{A}_n\right\}}=\lambda_p\frac{\Pr\left\{\mathbf{A}_{n+1}\right\}}{\Pr\left\{\mathbf{A}_n\right\}}.$$

Given these, we may now easily establish the stationary distribution (9) and, hence, the expected number of sessions and the system blocking probability (10).

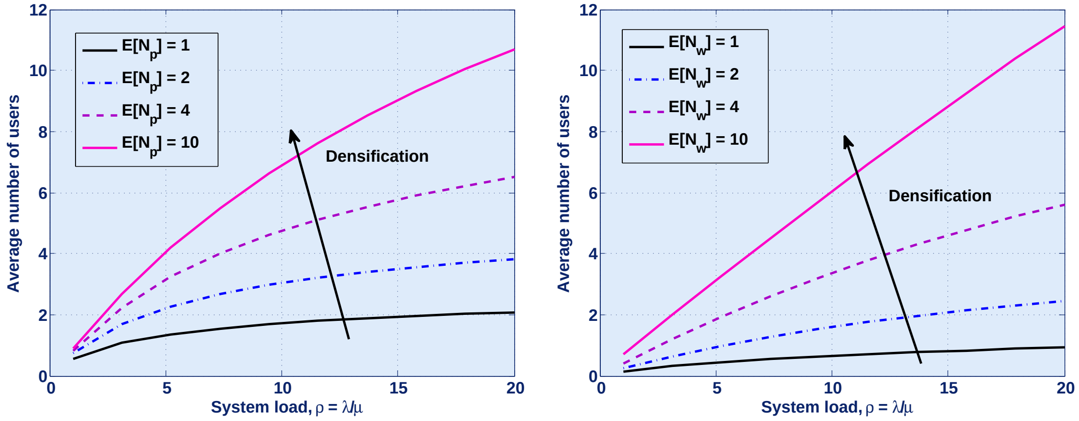


Fig. 10. Dependence on the density of BSs/APs for the Pico (left) and the WLAN (right) tiers: expected number of users/sessions.

V. Example Numerical Results and Conclusions

In order to verify the accuracy of our analytical model and its main assumptions, we construct a series of test scenarios. Our simulation implements explicitly all of the stochastic processes considered in Section III, inclusive of the associated system memory, whereas analysis corresponds to the proposed mathematical approximations. In particular, we recreate the urban environment defined by ITU/3GPP and combine that with varying pico BS and WLAN AP densities (as per [43]). We are specifically interested in dense deployments to analyze conditions where intelligent network selection would be most needed, i.e., when the cellular network would have difficulty supporting the offered traffic load on its own.

A. Scenario-related Considerations and Fading Effects

For the sake of an illustrative example, we consider the cell radius R of 288 meters (following the recommendations in [38]) with the varying user arrival rate λ. Every user upon its arrival generates a session of random duration with the mean μ^{-1} equal to 3 seconds and the target bitrate r_0 of 500 kbps. Naturally, the available spectral bandwidth for the WLAN, Pico, and Macro tiers is 20, 10, and 10 MHz, respectively. Further, the maximum allowed transmit power for a user is limited by 23 dBm on the WLAN and the Macro tiers, as well as by 20 dBm on the Pico tier. The remaining parameters are set in accordance with the available specifications and other standardization documents.

Our first example in Figure 9 compares macro cell performance with and without the effects of slow fading, which may introduce similar degradation on either of three tiers. Slow fading is typically modeled as the log-normal distribution with parameters $\mu = 0$ and $\sigma = 6$ dB as corresponds to macro urban scenario from [38]. The respective auxiliary distributions may be derived as explained in Appendix A. However, we leave this technical exercise out of scope of this paper. We generally observe that despite several simplifying assumptions our analysis remains exceptionally accurate across the entire range of session arrival rates.

B. Analyzing Metrics of Interest

Further, in Figure 10 we investigate the dependence of the expected number of sessions on the number of BSs/APs on the Pico/WLAN tier, respectively. We confirm that with the growing number of infrastructure nodes (i.e., network densification), the performance of both tiers improves dramatically. Finally, in Figure 11 we deeper detail the respective blocking probabilities for the integrated HetNet as well as for the three tiers individually: Macro, Pico, and WLAN. Our observation is that with two additional overlay tiers, the HetNet performance improves significantly over what can be achieved in the macro-only networks (cellular baseline). Remarkably, we actually witness visible performance improvement even with only a few additional infrastructure nodes, such as 4 WLAN APs and 4 Pico BSs in this example.

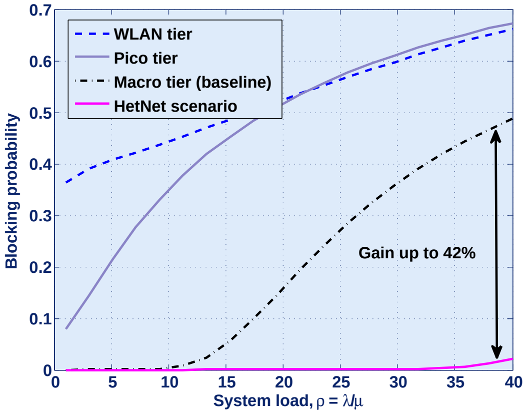


Fig. 11. Integrated HetNet with three tiers: blocking probabilities.

Therefore, we believe that multiple radio access technologies and the associated network selection intelligence for their efficient use will become a fundamental characteristic of future heterogeneous networks. In particular, we expect that the joint use of multiple RATs can leverage the rich multi-dimensional diversity across multiple radio networks to provide beyond-additive gains in network capacity and user connectivity experience.

Appendix

Here we provide details on calculation of several distributions necessary for the derivations described in the body of this paper.

A. Important distributions for the Macro tier

Using the transformation for $r^{\max}$, we can obtain the probability density function for the maximum instantaneous rate. For the simplification, we denote $r_i^{\max}$ as r and further study the random variable r.

Additionally, we note that the solution of the overall problem is heavily based on the distribution of the rate r and implicitly on the distribution of the channel gain γ. In case of different spatial point process the only part to change is both distributions due to different $f(x,y)$, which may become a good technical exercise to solve or to find an appropriate approximation.

Also we would like to emphasize that fading can be accounted for explicitly by introducing the attenuation random variable h, such that $\tilde{\gamma} = hGd^{-\kappa}$. In case of slow or shadow fading, h may result in log-normal distribution with the average of 1, whereas for the fast fading we may consider Rayleigh or Ricean distributions. In particular, new distribution of $\tilde{\gamma}$ may be found straightforwardly. That is, since fading process and location process are independent, we have:

$$f_{\tilde{\gamma}}(y) = \int\limits_{\gamma_{\min}}^{\gamma_{lim}} \frac{1}{x} f_{\gamma}(x) f_h(\frac{y}{x}) dx, \tag{18}$$

where f_γ is a baseline channel gain distribution due to spatialness of users and it is addressed below. In order to address the complex problems with dependency of the processes, one may use a joint probability density function.

Given the uniform distribution of locations within the circle, the distribution of distances between the user and the macro BS is $f_d(d) = 2d/R^2$, $0 \leq d \leq R$. Then, basing on the relation between the path loss γ and the distance d given by (14), we obtain the distribution for the channel gain γ:

$$f_\gamma(\gamma) = \frac{1}{\kappa}\left[\frac{G}{\gamma}\right]^{\frac{1-\kappa}{\kappa}} \cdot \frac{2G^{\frac{1}{\kappa}}}{\gamma^{\frac{1}{\kappa}}R^2} = \frac{2}{\kappa R^2}\left[\frac{G}{\gamma}\right]^{\frac{2}{\kappa}-1}, \gamma_R \leq \gamma \leq \gamma_{lim}, \tag{19}$$

where γ_R and γ_{lim} correspond to the lowest signal at the edge $d = R$ and the maximum level at $d = d_0$.

In order to obtain distribution of the random variable $r(d)$, we exploit the following expression for d:

$$d = \left[\frac{\eta G p_{\max}}{KN_0}\right]^{1/\kappa} \left(e^{r/w} - 1\right)^{-1/\kappa}, \tag{20}$$

and, hence, we establish the distribution for the random variable r without any constraints:

$$F_r(r) = 1 - \frac{1}{R^2}\left[\frac{\eta G p_{\max}}{KN_0}\right]^{2/\kappa} \left(e^{r/w} - 1\right)^{-2/\kappa}, r \geq r_R, \tag{21}$$

where r_R is the maximum possible rate at the border R (the lower border for possible values of maximum rate):

$$r_R = \min\left\{r_{lim}, w \cdot \log\left(1 + \frac{\eta G p_{\max}}{R^\kappa KN_0}\right)\right\} \geq r_i, \forall i. \tag{22}$$

Then, applying restrictions r_{lim} above the formula (21), we arrive at the following cumulative distribution function:

$$F_r(r) = 1 - \tfrac{1}{R^2}\left[\tfrac{\eta G p}{N_0}\right]^{2/\kappa} \left(e^{r/w} - 1\right)^{-2/\kappa}, r_R \leq r < r_{lim},$$
$$F_r(r_{lim}) = 1.$$

Therefore, the necessary probabilities in (12) may be easily obtained as:

$$\Pr\left\{r \geq \tfrac{r_0 n}{\delta_m}\right\} = 1 - \Pr\left\{r < \tfrac{r_0 n}{\delta_m}\right\} = 1 - F_r\left(\tfrac{r_0 n}{\delta_m}\right), \tag{23}$$

and $Pr\left\{r \geq \frac{r_0}{\delta_m}(n+1)\right\}$ is calculated similarly. The latter completes the formula (11) and delivers the steady-state distribution (9).

B. Important distributions for the WLAN tier

Let us establish auxiliary probabilities and distributions for the WLAN tier. First, we derive the distribution of user distances d_1 to the nearest AP:

$$F_{d_1}(d) = 1 - \Pr\{d_1 \geq d\} = 1 - \Pr\{N_w(d) = 0\}, d \geq 0,$$

where $N_w(d)$ is the random number of APs in the circle of radius d around the user. By analogy, we derive the distribution of distances to the second nearest AP:

$$F_{d_2}(d) = 1 - \Pr\{d_2 \geq d_{thr}\} = 1 - \Pr\{0 \leq N_w(d_{thr}) \leq 1\},$$

where $\Pr\{0 \leq N_w(d_{thr}) \leq 1\}$ is probability to have zero or one WLAN AP in the circle of radius d around the given user. Accounting for the fact that APs are distributed according to the Poisson process, we obtain for distances $d \geq 0$:

$$F_{d_1}(d) = 1 - e^{-\pi L_w d^2}, \quad f_{d_1}(d) = 2\pi L_w d e^{-\pi L_w d^2}. \tag{24}$$

Then, we find the probability for the second distance:

$$\begin{aligned}\Pr\{d_2 \geq d_{thr}\} = \Pr\{N_w(d_{thr}) \leq 1\} = \\ = L_w \pi d_{thr}^2 e^{-\pi d_{thr}^2} + e^{-L_w \pi d_{thr}^2}.\end{aligned} \tag{25}$$

For simplicity of notation, when accounting for the upper rate limit, we assume that d-distribution is strictly coupled with channel gain γ- and maximum rate r-distributions. Let us now find the distribution of random variable:

$$y = \frac{r_0}{r_i^{\max}} = \frac{r_0}{w}\left[\log\left(1 + p_{\max}\frac{\eta G}{KN_0} d^{-\kappa}\right)\right]^{-1}, d \geq d_0. \tag{26}$$

Expressing the distance from the equation above, we arrive at:

$$d = \left(\frac{p_{\max}\eta G}{KN_0}\right)^{\frac{1}{\kappa}} \left(e^{\frac{r_0}{wy}} - 1\right)^{-\frac{1}{\kappa}}, y \geq y_0, \tag{27}$$

where $y_0 = y(d_0)$, and the derivative of $d(y)$:

$$d'_y = \frac{1}{\kappa}\left(\frac{p_{\max}\eta G}{KN_0}\right)^{\frac{1}{\kappa}} \left(e^{\frac{r_0}{wy}} - 1\right)^{-\frac{1}{\kappa}-1} e^{\frac{r_0}{wy}} \frac{r_0}{wy^2}. \tag{28}$$

We may write the following distribution functions:

$$F_y(y)=1-e^{-\pi L_w\left(\frac{p_{\max}\eta G}{KN_0}\right)^{\frac{2}{\kappa}}\left(e^{\frac{r_0}{wy}}-1\right)^{-\frac{2}{\kappa}}}, y \geq y_0. \tag{29}$$

Hence, the probability distribution function is given as:

$$f_y(y) = 2\pi L_w d(y) d'(y) e^{-\pi L_w\left(\frac{p_{\max}\eta G}{KN_0}\right)^{\frac{2}{\kappa}}\left(e^{\frac{r_0}{wy}}-1\right)^{-\frac{2}{\kappa}}}. \tag{30}$$

Therefore, the expected value of random variable $E[y|y \leq \delta_w]$ may be found as:

$$\begin{aligned} E[y|y \leq \delta_w] &= \int_{y_0}^{\delta_w} y f_y(y|y \leq \delta_w) dy = \\ &= \frac{2\pi L_w}{C_3}\int_{y_0}^{\delta_w} y d(y) d'(y) e^{-\pi L_w\left(\frac{p_{\max}\eta G}{KN_0}\right)^{\frac{2}{\kappa}}\left(e^{\frac{r_0}{wy}}-1\right)^{-\frac{2}{\kappa}}} dy + \\ &+ \frac{2\pi L_w}{C_3} y_0 \int_{0}^{y_0} d(y) d'(y) e^{-\pi L_w\left(\frac{p_{\max}\eta G}{KN_0}\right)^{\frac{2}{\kappa}}\left(e^{\frac{r_0}{wy}}-1\right)^{-\frac{2}{\kappa}}} dy, \end{aligned} \tag{31}$$

where $C_3 = \Pr\{y \leq \delta_w\} = F_y(\delta_w)$ and y_0 is assumed to be less than δ_w.

Let us now consider the state of the system when there are n users in service. The density of users equals $\lambda_N = n/\left(\pi R^2\right)$. We continue filling the plane up to this density so that every AP may receive sessions to serve.

Theorem 1. *The average number of sessions per AP n_0 tends to $\frac{n}{L_w(\pi R^2)} = \frac{n}{E[N_w]}$ for large areas, where $E[N_w]$ is the expected number of APs within the circle R.*

Proof. For a large area $S \gg \pi R^2$, we estimate the average number of users per AP $N_n(S)/N_w(S)$, where N_n and N_w are random numbers of users and APs in the area S. The expected value of $N_n(S)/N_w(S)$ may be found by definition as follows:

$$\begin{aligned} E\left[\frac{N_n(S)}{N_w(S)}\right] &= \sum_{n=1}^{\infty}\sum_{k=0}^{\infty}\frac{k}{n}\Pr\{N_n(S)=k, N_w(S)=n\}] = \\ &= \sum_{n=1}^{\infty}\sum_{k=0}^{\infty}\frac{k}{n}\frac{(L_w S)^n}{n!}e^{-L_w S}\frac{(\lambda_N S)^k}{k!}e^{-\lambda_N S} \\ = \lambda_N S \sum_{n=1}^{\infty}\frac{1}{n e^{L_w S}}\frac{(L_w S)^n}{n!} &\to_{S\to\infty} \frac{\lambda_N}{L_w} = \frac{n}{L_w\left(\pi R^2\right)}, \end{aligned} \tag{32}$$

where $\lambda_N = \frac{n}{(\pi R^2)}$ is the density of users (per a unit of area), L_w is the system parameter of the AP density on the plane, and S is the area of interest.

Reformulating the above, the average number of users per AP in the marginal cases tends to the ratio between the changing density of users and fixed density of APs. ■

Note that Theorem 1 above is similar in its flavor to the research findings obtained previously in [44]. Then, basing on these results, we may reformulate the following as stated in Assumption 11. A newly arrived session observes the system where, on average, every AP already serves $n_0 = \frac{n}{L_w(\pi R^2)}$ sessions.

In that case, σ is representing the average part of the resource exploited at the state n and is given by:

$$\sigma = E\left[\frac{r_0}{r_i^{\max}}\middle|\frac{r_0}{r_i^{\max}} \leq \delta_w\right]\frac{n}{E[N_w]} = E[y|y \leq \delta_w]\frac{n}{E[N_w]}, \tag{33}$$

where $E[y|y \leq \delta_w]$ is obtained via the numerical integral (31).

C. Important distributions for the Pico tier

The distribution of user distances to the nearest BS may be obtained similarly to that on the WLAN tier:

$$F_{d_1}(d) = 1 - e^{-\pi L_p d^2}, \quad f_{d_1}(d) = 2\pi L_p d e^{-\pi L_p d^2}, d \geq 0.$$

Then, we find the sought probability using the expression (17) and following the same reasoning as for the WLAN tier:

$$\begin{aligned} &\Pr\left\{\gamma_2 \leq \frac{N_0}{p_i}\right\} = \Pr\left\{\gamma_2 \leq \frac{\eta\gamma_1 N_0}{KN_0\left(e^{\frac{r_0 n_0}{w\delta_p}}-1\right)}\right\} = \\ &= \Pr\{\gamma_2 + y \leq 0\} = \Pr\{z \leq 0\}, \end{aligned} \tag{34}$$

where we replace the following variables $y = -\frac{\eta\gamma_1}{K\left(e^{\frac{r_0 n_0}{w\delta_p}}-1\right)} = -\frac{\eta G}{K\left(e^{\frac{r_0 n_0}{w\delta_p}}-1\right)} d_1^{-\kappa}$ and $z = \gamma_2 + y$.

In order to calculate the distribution of z, we firstly aim to find the distribution of random variable y. For that purpose, we exploit the following transform:

$$d_1 = \left(\frac{\eta G}{K\left(e^{\frac{r_0 n_0}{w\delta_p}}-1\right)}\right)^{\frac{1}{\kappa}}(-y)^{-\frac{1}{\kappa}}, \tag{35}$$

and its first derivative:

$$d_1' = \frac{1}{\kappa}\left(\frac{\eta G}{K\left(e^{\frac{r_0 n_0}{w\delta_p}}-1\right)}\right)^{\frac{1}{\kappa}}(-y)^{-\frac{1}{\kappa}-1}. \tag{36}$$

Using the expressions above, we can easily obtain the distribution function:

$$F_y(y) = 1 - e^{-\pi L_p d(y)^2}, y_{min} < y < 0, \tag{37}$$

where y_{min} corresponds to the rate limit and d_0. Hence, the sought function has the following form:

$$f_y(y) = 2\pi L_p d(y) d'(y) e^{-\pi L_p d^2(y)}, y_{min} \leq y < 0. \tag{38}$$

Basing on the relation between path loss γ and distance d (14), we continue by calculating the distribution of random variable $0 < \gamma_2 \leq \gamma_{\max}$, which is one of the components in z:

$$\begin{aligned} &F_{\gamma_2}(\gamma) = \Pr\{\gamma_2 < \gamma\} = \Pr\{d_2 > d(\gamma)\} = \\ &= \Pr\{N_p(d(\gamma)) \leq 1\} = L_p\pi d^2(\gamma)e^{-L_p\pi d^2(\gamma)} + e^{-L_p\pi d^2(\gamma)}, \end{aligned} \tag{39}$$

where $d(\gamma) = G^{\frac{1}{\kappa}}\gamma^{-\frac{1}{\kappa}}$, $d'(\gamma) = -\frac{1}{\kappa}G^{\frac{1}{\kappa}}\gamma^{-\frac{1}{\kappa}-1}$, and $\gamma_{\max}$ corresponds to the distance d_0.

Given the expression $F_{d_2}(d) = 1 - \Pr\{d_2 \geq d_2(\gamma)\}$, we obtain the probability density function for the distance to the second near BS d_2 as:

$$f_{d_2}(d) = 2{L_p}^2\pi^2 d^3 e^{-L_p\pi d^2}, 0 \leq d < \infty.$$

For the probability density function of channel gain γ_2 (if it is bounded by d_0) we may write:

$$\begin{aligned} f_{\gamma_2}(\gamma) &= \tfrac{2}{\kappa}{L_p}^2\pi^2 G^{\frac{4}{\kappa}}\gamma^{-\frac{4}{\kappa}-1}e^{-L_p\pi d^2(\gamma)}, 0 < \gamma < \gamma_{\max} \\ f_{\gamma_2}(\gamma_{\max}) &= 1 - F_{\gamma_2}(\gamma_{\max}). \end{aligned} \tag{40}$$

Taking into account the limits $y_{min} \leq y < 0$ and $0 < \gamma \leq \gamma_{\max}$, we may derive the expression for $F_z(z)$:

$$\begin{aligned} F_z(z) = \int\limits_{0}^{\gamma_{\max}} f_\gamma(\gamma) \int\limits_{y_{min}}^{\min(0,z-\gamma)} f_y(y) dy d\gamma = \\ = \int\limits_{0}^{\gamma_{\max}} f_\gamma(\gamma) F_y\left(\min(0, z-\gamma)\right) d\gamma, \end{aligned} \tag{41}$$

where $\gamma_{\max}$ corresponds to d_0.

Therefore, we may substitute $z = 0$ into the expression above and find the unknown probability $\Pr\{z < 0\}$:

$$\begin{aligned} \Pr\{z < 0\} = F_z(0) = \int\limits_{0}^{\gamma_{\max}} f_\gamma(\gamma) F_y(-\gamma) d\gamma = \\ = \tfrac{2}{\kappa} {L_p}^2 \pi^2 G^{\frac{4}{\kappa}} \int\limits_{0}^{\gamma_{\max}} \gamma^{-\frac{4}{\kappa}-1} e^{-L_p \pi G^{\frac{2}{\kappa}} \gamma^{-\frac{2}{\kappa}}} \left(1 - e^{C_1 \gamma^{-\frac{2}{\kappa}}}\right) d\gamma, \end{aligned} \tag{42}$$

where

$$C_1 = -\pi L_p \left(\frac{\eta G}{K\left(e^{\frac{r_0 n_0}{w \delta_p}} - 1\right)} \right)^{\frac{2}{\kappa}}.$$

Hence, the probability to fulfill the interference condition $\Pr\{\gamma_2 p_i \leq N_0\} = F_z(0)$ equals:

$$\begin{aligned} 1 - \tfrac{2}{\kappa} {L_f}^2 \pi^2 G^{\frac{4}{\kappa}} \int\limits_{0}^{\gamma_{\max}} \gamma^{-\frac{4}{\kappa}-1} e^{-L_p \pi G^{\frac{2}{\kappa}} \gamma^{-\frac{2}{\kappa}}} e^{C_1 \gamma^{-\frac{2}{\kappa}}} d\gamma = \\ = 1 - \tfrac{2}{\kappa} {L_p}^2 \pi^2 G^{\frac{4}{\kappa}} \int\limits_{0}^{\gamma_{\max}} \gamma^{-\frac{4}{\kappa}-1} e^{\left(-L_p \pi G^{\frac{2}{\kappa}} + C_1\right) \gamma^{-\frac{2}{\kappa}}} d\gamma. \end{aligned} \tag{43}$$

Here, let us calculate the integral $\int\limits_{0}^{\gamma_{\max}} \gamma^{-\frac{4}{\kappa}-1} e^{-C\gamma^{-\frac{2}{\kappa}}} d\gamma$ for $C > 0$ by substituting $u = \gamma^{-\frac{2}{\kappa}}$, $\gamma = u^{-\frac{\kappa}{2}}$, $d\gamma = -\frac{\kappa}{2} u^{-\frac{\kappa}{2}-1} du$:

$$\begin{aligned} -\int\limits_{0}^{\gamma_{\max}} \tfrac{\kappa}{2} u^{-\frac{\kappa}{2}-1} u^{-\frac{\kappa}{2}(-\frac{4}{\kappa}-1)} e^{-Cu} d\gamma = -\tfrac{\kappa}{2} \int\limits_{0}^{\gamma_{\max}} u e^{-Cu} du = \\ = -\tfrac{\kappa}{2} \left(-\tfrac{1}{C} u e^{-Cu} \Big|_0^{\gamma_{\max}} + \tfrac{1}{C} \int\limits_{0}^{\gamma_{\max}} e^{-Cu} du \right) = \\ = \tfrac{\kappa \gamma_{\max}}{2C} e^{-C\gamma_{\max}} + \tfrac{\kappa}{2C^2} e^{-Cu} \Big|_0^{\gamma_{\max}} = \tfrac{\kappa \gamma_{\max}}{2C} e^{-C\gamma_{\max}} + \tfrac{\kappa}{2C^2}. \end{aligned} \tag{44}$$

Then, we may continue by:

$$\Pr\{\gamma_2 p_i \leq N_0\} = 1 - {L_p}^2 \pi^2 G^{\frac{4}{\kappa}} \left(\tfrac{\gamma_{\max}}{C} e^{-C\gamma_{\max}} - \tfrac{1}{C^2} \right), \tag{45}$$

where $C = L_p \pi G^{\frac{2}{\kappa}} - C_1$. We note that the above is fair for the case when multiplier $b = \frac{\eta}{K\left(e^{\frac{r_0 n_0}{w}} - 1\right)} < 1$. Otherwise, if $b = 1$, then in (34) $\gamma_2 < \gamma_1$ by definition and increasing $b \geq 1$ leads to $\Pr\{\gamma_2 p_i \leq N_0\} = 1$.

Let us now consider the state of the system when there are n users in service. The density of users equals $\lambda_N = n/\left(\pi R^2\right)$ similarly to the WLAN tier. The same density is kept all over the plane.

Theorem 2. *The average number of users per pico BS* n_0 *tends to* $\frac{n}{L_p(\pi R^2)} = \frac{n}{E[N_p]}$ *for large areas, where* $E[N_p]$ *is the expected number of pico BSs per cell.*

Proof. The proof is similar to the one for the WLAN tier. ■

With $n_0 = \max\left(1, \frac{n}{E[N_p]}\right)$ known for each state n, we may estimate the expression (45) and, therefore, establish the transition rates for the Pico tier.

Acknowledgment

This work is supported by Intel Corporation, GETA, and the Internet of Things program of DIGILE, funded by Tekes. The work of the second author is supported with a Postdoctoral Researcher grant by the Academy of Finland. The authors are grateful to Alexander Pyattaev (Tampere University of Technology, Tampere, Finland) and Prof. Andrey Turlikov (State University of Aerospace Instrumentation, St. Petersburg, Russia) for insightful discussions, as well as to the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of this paper.

References

[1] "3GPP LTE Release 10 & beyond (LTE-Advanced)."

[2] S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, and Y. Koucheryavy, "Cellular traffic offloading onto network-assisted device-to-device connections," *IEEE Communications Magazine*, vol. 52, pp. 20–31, 2014.

[3] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast, Update, 2012-2017," 2013.

[4] B. A. Bjerke, "LTE-Advanced and the evolution of LTE deployments," *IEEE Wireless Communications*, vol. 18, pp. 4–5, 2011.

[5] P. Marsch, B. Raaf, A. Szufarska, P. Mogensen, H. Guan, M. Farber, S. Redana, K. Pedersen, and T. Kolding, "Future Mobile Communication Networks: Challenges in the Design and Operation," *IEEE Vehicular Technology Magazine*, vol. 7, pp. 16–23, 2012.

[6] D. Raychaudhuri and N. B. Mandayam, "Frontiers of Wireless and Mobile Communications," *Proceedings of the IEEE*, vol. 100, pp. 824–840, 2012.

[7] S. Andreev, P. Gonchukov, N. Himayat, Y. Koucheryavy, and A. Turlikov, "Energy efficient communications for future broadband cellular networks," *Computer Communications Journal (COMCOM)*, vol. 35, pp. 1662–1671, 2012.

[8] L. Al-Kanj, Z. Dawy, and E. Yaacoub, "Energy-Aware Cooperative Content Distribution over Wireless Networks: Design Alternatives and Implementation Aspects," *IEEE Communications Surveys & Tutorials*, vol. 15, pp. 1736–1760, 2013.

[9] K. Andersson and C. Ahlund, "Optimized Access Network Selection in a Combined WLAN/LTE Environment," *Wireless Personal Communications*, vol. 61, pp. 739–751, 2011.

[10] M. Gerasimenko, N. Himayat, S.-p. Yeh, S. Talwar, S. Andreev, and Y. Koucheryavy, "Characterizing Performance of Load-Aware Network Selection in Multi-Radio (WiFi/LTE) Heterogeneous Networks," in *GLOBECOM Workshops (GC Wkshps)*, 2013.

[11] A. Y. Panah, S.-P. Yeh, N. Himayat, and S. Talwar, "Utility-based radio link assignment in multi-radio heterogeneous networks," in *Proc. of International Workshop on Emerging Technologies for LTE-Advanced and Beyond-4G on IEEE Globecom*, 2012.

[12] A. Ozgur, O. Leveque, and D. Tse, "Hierarchical Cooperation Achieves Optimal Capacity Scaling in Ad Hoc Networks," *IEEE Transactions on Information Theory*, vol. 53, pp. 3549–3572, 2007.

[13] S.-P. Yeh, S. Talwar, G. Wu, N. Himayat, and K. Johnsson, "Capacity and coverage enhancement in heterogeneous networks," *IEEE Wireless Communications*, vol. 18, pp. 32–38, 2011.

[14] M. Bennis, M. Simsek, W. Saad, S. Valentin, and M. Debbah, "When cellular meets WiFi in wireless small cell networks," *IEEE Communications Magazine*, vol. 51, pp. 44–50, 2013.

[15] C. Sankaran, "Data offloading techniques in 3GPP Rel-10 networks: A tutorial," *IEEE Communications Magazine*, vol. 50, pp. 46–53, 2012.

[16] S.-P. Yeh, A. Y. Panah, N. Himayat, and S. Talwar, "QoS aware scheduling and cross-RAT coordination in multi-radio heterogeneous networks," in *IEEE VTC*, 2013.

[17] S. Tombaz, A. Vastberg, and J. Zander, "Energy- and cost-efficient ultra-high-capacity wireless access," *IEEE Wireless Communications*, vol. 18, pp. 18–24, 2011.

[18] A. Prasad, O. Tirkkonen, P. Lundén, O. N. C. Yilmaz, L. Dalsgaard, and C. Wijting, "Energy-efficient inter-frequency small cell discovery techniques for LTE-Advanced heterogeneous network deployments," *IEEE Communications Magazine*, vol. 51, pp. 72–81, 2013.

[19] S. Singh, H. Dhillon, and J. Andrews, "Offloading in Heterogeneous networks: Modeling, analysis, and design insights," *IEEE Transactions on Wireless Communications*, vol. 12, pp. 2484–2497, 2013.

[20] T. Novlan, H. Dhillon, and J. Andrews, “Analytical modeling of uplink cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 12, pp. 2669–2679, 2013.

[21] S. Singh and J. Andrews, “Joint resource partitioning and offloading in Heterogeneous cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 13, pp. 888–901, 2014.

[22] *U-LTE: Unlicensed Spectrum Utilization of LTE, Huawei white paper, 2014.*

[23] N. Himayat, S.-P. Yeh, A. Y. Panah, S. Talwar, M. Gerasimenko, S. Andreev, and Y. Koucheryavy, “Multi-Radio Heterogeneous Networks: Architectures and Performance,” in *Proc. of IEEE International Conference on Computing, Networking and Communications (ICNC)*, 2014.

[24] “IEEE 802.11-2012, Part 11: Local and metropolitan area networks,” 2012.

[25] “Performance benefits of RAN level enhancements for WLAN/3GPP,” *3GPP R2-133604*, 2013.

[26] “Study on WLAN/3GPP Radio Interworking,” *3GPP TR 37.834*, 2013.

[27] “Access to the 3GPP Evolved Packet Core (EPC) via non-3GPP access networks,” *3GPP TS 24.302*, 2013.

[28] “Architecture enhancements for non-3GPP accesses,” *3GPP TS 23.402*, 2013.

[29] “Network based IP flow mobility,” *3GPP TR 23.861*, 2012.

[30] L. Wang and G. S. Kuo, “Mathematical Modeling for Network Selection in Heterogeneous Wireless Networks – A Tutorial,” *IEEE Communications Surveys & Tutorials*, vol. 15, pp. 271–292, 2013.

[31] M. Peng, D. Liang, Y. Wei, J. Li, and H.-H. Chen, “Self-configuration and self-optimization in LTE-Advanced heterogeneous networks,” *IEEE Communications Magazine*, vol. 51, pp. 36–45, 2013.

[32] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, “Modeling and analysis of K-tier downlink heterogeneous cellular networks,” *IEEE Journal on Selected Areas in Communications*, vol. 30, pp. 550–560, 2012.

[33] H. Kim and G. de Veciana, “Leveraging dynamic spare capacity in wireless systems to conserve mobile terminals energy,” *IEEE/ACM Transactions on Networking*, vol. 18, pp. 802–815, 2010.

[34] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, “RAT selection games in HetNets,” in *Proc. of IEEE INFOCOM*, 2013.

[35] J. G. Andrews, R. Ganti, M. Haenggi, and N. Jindal, “A primer on spatial modeling and analysis in wireless networks,” *IEEE Communications Magazine*, vol. 48, pp. 156–163, 2010.

[36] H. ElSawy, E. Hossain, and M. Haenggi, “Stochastic Geometry for Modeling, Analysis, and Design of Multi-Tier and Cognitive Cellular Wireless Networks: A Survey,” *IEEE Communications Surveys & Tutorials*, vol. 15, pp. 996–1019, 2013.

[37] J. G. Andrews, “Seven ways that HetNets are a cellular paradigm shift,” *IEEE Communications Magazine*, vol. 51, pp. 136–144, 2013.

[38] “Guidelines for evaluation of radio interface technologies for IMT-Advanced,” *ITU*, 2009.

[39] O. Galinina, A. Trushanin, V. Shumilov, R. Maslennikov, Z. Saffer, S. Andreev, and Y. Koucheryavy, “Energy-Efficient Operation of a Mobile User in a Multi-Tier Cellular Network,” in *20th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA’13)*, 2013.

[40] J. Xu, J. Zhang, and J. Andrews, “On the accuracy of the Wyner model in cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 10, pp. 3098–3109, 2011.

[41] A. Ometov, S. Andreev, A. Turlikov, and Y. Koucheryavy, “Characterizing the Effect of Packet Losses in Current WLAN Deployments,” in *Proc. of the International Conference Intelligent Transportation Systems Telecommunications (ITST)*, 2013.

[42] *3GPP TS 34.122. Terminal conformance specification; Radio transmission and reception (TDD)*, March 2014.

[43] *3GPP TR 36.819. Coordinated multi-point operation for LTE physical layer aspects*, September 2013.

[44] S. Foss and S. Zuyev, “On a Voronoi Aggregative Process related to a Bivariate Poisson Process,” *Advances in Applied Probability*, vol. 28, pp. 965–981, 1996.

Authors’ Biographies

Olga Galinina is a PhD Candidate in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. She received her B.Sc. and M.Sc. degree in Applied Mathematics from Department of Applied Mathematics, Faculty of Mechanics and Physics, Saint-Petersburg State Polytechnical University, Russia. She has publications on mathematical problems in the novel telecommunication protocols in internationally recognized journals and high-level peer-reviewed conferences. Her research interests include applied mathematics and statistics, queueing theory and its applications; wireless networking and energy efficient systems, machine-to-machine and device-to-device communication.

Sergey Andreev is a Senior Research Scientist in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. He received the Specialist degree (2006) in Information Security and the Cand.Sc. degree (2009) in Wireless Communications both from St. Petersburg State University of Aerospace Instrumentation, St. Petersburg, Russia, as well as the Ph.D. degree (2012) in Technology from Tampere University of Technology, Tampere, Finland. Sergey (co-)authored more than 80 published research works. His research interests include wireless communications, energy efficiency, heterogeneous networking, cooperative communications, and machine-to-machine applications.

Mikhail Gerasimenko is a Research Assistant at Tampere University of Technology in the Department of Electronics and Communications Engineering. Mikhail received specialist degree in Saint Petersburg University of Telecommunications in 2011. In 2013 he obtained Master of Science degree in Tampere University of Technology. Mikhail started his academic career in 2011 and during 2 years he appeared as an (co-)author of several scientific journal and conference publications, as well as several patents. Moreover, he also acted as a review and participated in education activities. His main subjects of interest are wireless communications, Machine-Type Communications, Heterogeneous networks.

Yevgeni Koucheryavy is a Full Professor in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. He received the PhD degree (2004) from the TUT, Finland. Prior joining the TUT, Yevgeni spent five years in industry with R&D LONIIS in St. Petersburg, Russia, where he held various technical and managerial positions. Yevgeni actively participates in national and international research and development projects – in particular currently he is chairing COST IC0906 ”WiNeMO: Wireless Networking for Moving Objects” scheduled for 2010-2014. Yevgeni has authored or co-authored over 100 papers in the field of advanced wired and wireless networking and communications. Yevgeni is a Senior IEEE member.

Nageen Himayat is a Senior Research Scientist with Intel Labs, where she performs research on broadband wireless systems, including heterogeneous networks, cross-layer radio resource management, MIMO-OFDM techniques, and optimizations for M2M communications. She has over 15 years of research and development experience in the telecom industry. She obtained her B.S.E.E from Rice University and her Ph.D. in electrical engineering from the University of Pennsylvania in 1989 and 1994, respectively.

Shu-ping Yeh is a Research Scientist in the Wireless Communications Laboratory at Intel. She received her M.S. and Ph.D. from Stanford University in 2005 and 2010, respectively, and her B.S. from National Taiwan University in 2003, all in electrical engineering. Her current research focus includes interference mitigation in multitier networks utilizing multi-antenna techniques, machine-to-machine communications, and interworking of multiple radio access technologies within a network.

Shilpa Talwar is a Principal Engineer in the Wireless Communications Laboratory at Intel, where she is conducting research on mobile broadband technologies. She has over 15 years of experience in wireless. Prior to Intel, she held several senior technical positions in wireless industry. She graduated from Stanford University in 1996 with a Ph.D. in applied mathematics and an M.S. in electrical engineering. She authored numerous technical publications and patents.

Publication 6

© 2014 Elsevier B.V. Reprinted, with permission, from

O. Galinina, S. Andreev, A. Turlikov, and Y. Koucheryavy, "Optimizing Energy Efficiency of a Multi-Radio Mobile Device in Heterogeneous Beyond-4G Networks", *Performance Evaluation*, vol. 78, pp. 18-41, 2014.

Performance Evaluation 78 (2014) 18–41

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Optimizing energy efficiency of a multi-radio mobile device in heterogeneous beyond-4G networks

Olga Galinina [a,*], Sergey Andreev [a], Andrey Turlikov [b], Yevgeni Koucheryavy [a]

[a] Department of Electronics and Communications Engineering, Tampere University of Technology, Korkeakoulunkatu 1, FI-33720, Tampere, Finland

[b] Department of Information and Communication Systems, State University of Aerospace Instrumentation, Bolshaya Morskaya 67, 190000, St. Petersburg, Russia

ARTICLE INFO

Article history:
Received 15 March 2013
Received in revised form 2 June 2014
Accepted 6 June 2014
Available online 13 June 2014

Keywords:
Heterogeneous network
Multi-radio device
Optimal power control
Energy efficiency
Beyond-4G deployment

ABSTRACT

In this paper, we address the operation of a multi-radio mobile device in heterogeneous wireless deployments. We assume that such a device may efficiently control its radio interfaces when using the available radio access technologies. In particular, we investigate the potential of flexible transmit power allocation and develop a provably optimal power control scheme that strictly maximizes the energy efficiency of the mobile device, while at the same time satisfies the minimum required level of the user data rate. When compared against simpler (heuristic) power control strategies, our solution always demonstrates the best energy efficiency of the multi-radio device by enabling collaborative operation between several radio technologies, which makes it a useful benchmark for the future integrated beyond-4G wireless networks.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

1.1. General motivation

Wireless networks demonstrate worldwide proliferation, which has further advanced recently with the introduction of novel fourth generation (4G) communication technologies [1,2]. Adoption of these 4G technologies is becoming increasingly widespread, allowing for improved access to services and applications previously only supported through fixed broadband systems [3]. However, existing wireless deployments are still unable to deliver their users the desired ubiquitous connectivity experience due to the shortage of available capacity and lack of service uniformity [4].

Whereas there is currently no technical definition of what comes after the state-of-the-art 4G technologies, experts agree on that future *beyond-4G* wireless communications will probably be a converged set of co-existing radio access networks, rather than one single technology [5]. As wireless spectrum continues to be scarce and expensive, the success of future beyond-4G systems requires effective solutions to overcome the divide between the demanded quality-of-service (QoS) and the limited network resources.

Over the years, wireless spectrum has become one of the most valuable natural resources, which accentuates the importance of its efficient use (i.e., spectral efficiency). However, energy efficiency is also becoming increasingly important

[*] Corresponding author. Tel.: +358 465563924.
E-mail addresses: olga.galinina@gmail.com, olga.galinina@tut.fi (O. Galinina), sergey.andreev@tut.fi (S. Andreev), turlikov@vu.spb.ru (A. Turlikov), yk@cs.tut.fi (Y. Koucheryavy).

http://dx.doi.org/10.1016/j.peva.2014.06.002
0166-5316/© 2014 Elsevier B.V. All rights reserved.

primarily for small form-factor mobile devices, where wireless power consumption dominates the total device power budget. This is due to the increasing disproportion between the available and the required battery capacity, which is demanded by the ubiquitous multimedia applications [6]. To compensate for this growing gap, aggressive improvements in all aspects of wireless system design are necessary [7].

Whereas energy efficiency is accentuated by the need of extending client device operation time without recharging, the need for improved service continuity is dictated by the ubiquitous wireless multimedia applications. Currently, wireless cellular, local, and personal area networking technologies as well as supportive network architectures are evolving towards more advanced and complex converged networks. On the other hand, consumer electronics is spawning a huge explosion in the number and variety of multi-radio devices [8], driven by the user demand for "anytime, anywhere" connectivity.

The problem of energy efficient interworking between the available wireless technologies in a user multi-radio device is therefore addressed in this work in order to develop provably efficient techniques that allow for significant energy performance improvement in heterogeneous wireless environments.

1.2. Research background

Conventional wireless devices are typically communicating their data by choosing one of the fixed set of modulation and coding schemes, which sacrifices flexible power adaptation for design simplicity [9]. This often causes excessive energy consumption or pessimistic data rates selected for peak channel conditions [10]. Hence, physical layer parameters should be flexibly adjusted to actually account for the client QoS requirements as well as for the state of the wireless channel to reach a compromise between energy and spectral efficiencies [11]. In this regard, throughput optimization has long been an attractive research direction [12,13]. However, as wireless clients become increasingly mobile, the focus of recent efforts tends to shift towards energy consumption at all layers of communication systems, from architectures [14] to algorithms [15].

Energy efficiency is becoming increasingly important for wireless networks due to the limited battery lifetime of mobile clients. For maximizing energy efficiency, so-called "bits-per-Joule" [16] or "throughput-per-Joule" [17] metrics are often considered. Several approaches are known to focus on energy efficiency. These include water-filling power allocation techniques that optimize throughput with respect to the fixed total transmit power limitation [18,19], as well as adaptation of both the total transmit power and its allocation according to the channel state information [10,20].

However, the vast majority of existing information-theoretic approaches (see, e.g., [21,22]) account only for the transmit power when investigating energy consumption. Typically, a client device will also consume extra circuit power, which is independent of its data rate [23,24] and can actually be on the same order (and even comparable) with the maximum allowed transmit power. As such, the circuit power consumption should be considered explicitly when optimizing energy efficiency [25]. With the emphasis on circuit power, recent work suggests the use of optimization theory for establishing energy-optimal communication settings [10,26] to balance transmit and circuit power consumption. These findings indicate that the conventional water-filling approach to simply extend the transmission time of the device may not be attractive anymore since circuit energy consumption grows with transmission duration.

With the growing use of smaller cells to improve the capacity of 4G systems, the coverage ranges of cellular, local, and personal area networks are increasingly overlapping. In the extreme, contemporary urban wireless deployments often include areas where different communication networks are co-located [27,28]. As long as these technologies occupy non-overlapping frequency bands [28], they may coexist simultaneously without any significant performance degradation. This creates an attractive opportunity to cooperatively utilize several radio access networks for improved wireless connectivity [29].

Over the last few years, much literature has accumulated [30] exploring the interworking solutions within the core network and above, including seamless mobility between 3GPP and WLAN technologies, trusted access to 3GPP services with WLAN devices, and support for Access Network Discovery and Selection functions [31]. In particular, the network selection problem in heterogeneous wireless environment using IEEE 802.21 and IEEE 1900.4 frameworks has recently been studied [32]. We emphasize that our focus in this paper is, however, on the joint use of multiple networks that requires cooperation on the Radio Access Network (RAN) layer, which enables more flexible control of the transmission parameters [33]. We expect this work to be useful in the ongoing 3GPP discussions on WLAN/3GPP radio interworking [34].

We expect that intelligent coupling between multiple radio access technologies (such as LTE-Advanced, HSPA, WiMAX, WiFi, Bluetooth, ZigBee, etc.) will enable efficient operation of a multi-radio device and thus realize the desired uniform user experience. To achieve this, both short- and long-range technologies (e.g., WiFi and LTE-Advanced) may need to work cooperatively to augment system capacity and improve service continuity [35] in a beyond-4G network (see Fig. 1). Consequently, we seek to explore the potential of adaptive power control to improve the energy efficiency of a multi-radio device using heterogeneous connectivity at different scales.

The rest of this text is organized as follows. In Section 2, we introduce our system model with its main assumptions. Section 3 formulates a practical constrained optimization problem, where the energy efficiency of a multi-radio mobile device needs to be maximized subject to some realistic restrictions. We then solve this problem directly and obtain the exact solution. In Section 4, we provide several important numerical examples to conclude on the feasibility of our approach.

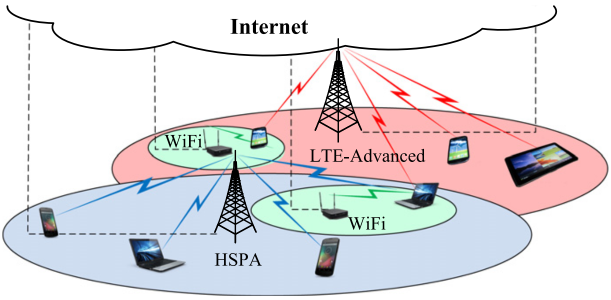


Fig. 1. Example topology of a heterogeneous network.

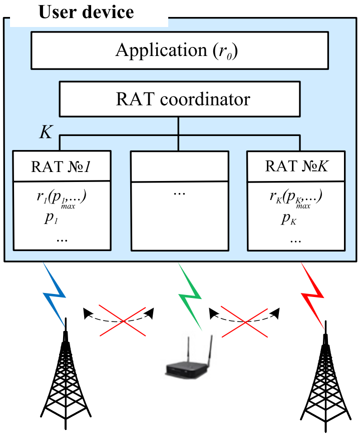


Fig. 2. Structure of a multi-radio mobile device.

Table 1
Notations of the analytical model.

Notation	Parameter description
r_0	Target bit-rate by the user
K	Number of available channels (radio interfaces)
r	Total data rate of the user device
p^{tx}	Total transmit power of the user device
p_i^{tx}	Transmit power on the channel i
$p_i^{\max}$	Maximum transmit power on the channel i
p_i^c	Circuit power on the channel i
r_i	Data rate on the channel i

2. System model

In this section, we introduce our system model and its main assumptions. We consider the uplink communication of a single user device transmitting its traffic to the Internet infrastructure (see Fig. 1). In a heterogeneous environment, such a multi-radio device may efficiently use the available radio access technologies by controlling its radio interfaces. We therefore concentrate on exploring the achievable data rate, power, and energy efficiency associated with such operation.

2.1. Main assumptions and notation

Our analytical model is based on the following assumptions and its core parameters are summarized by Table 1. We study the operation of a single user device (see Fig. 2), which needs to achieve a particular target bit-rate r_0 (specified by, e.g., an active mobile application). In other words, a data rate of at least r_0 must be guaranteed for this device to ensure that the QoS requirements are satisfied.

To abstract away a particular traffic model, the upper-layer traffic is assumed to be saturated, which often corresponds to the case of maximum available gains. We also assume that K alternative radio access technologies (RATs) are available for the device to use in a particular heterogeneous deployment. The device can send data on any or all of these RATs while attempting to meet its target bit-rate requirement.

The device power consumption when a particular RAT is used (that is, on a given communication channel) includes two components: p_i^{tx} and p_i^c, where $p_i^{tx} \leq p_i^{\max}$ is the instantaneous transmit power not exceeding the maximum power constraint, and p_i^c is the associated circuit power incurred regardless of how much transmit power is used in the respective channel. Following, e.g., [10], we further assume that the value of p_i^c is constant (we do not account for additional power consumption by the device itself). However, with the approach of this paper, this assumption could actually be relaxed without loss of generality.

On a given channel i, the relationship between the instantaneous transmit power p_i^{tx} and the achievable data rate r_i could be determined by a particular function $p_i^{tx}(r_i)$. For the sake of exposition, in this work we consider the important example, when this relationship is given by the Shannon–Hartley theorem for a point-to-point channel and $r_i(p_i^{tx}) = w_i \ln(1 + \gamma_i p_i^{tx})$ or, alternatively:

$$p_i^{tx}(r_i) = \frac{1}{\gamma_i}\left(e^{\frac{r_i}{w_i}} - 1\right), \tag{1}$$

where p_i^{tx} is the device transmit power (measured in W); γ_i is the signal-to-noise ratio, when the transmit power is 1 W; w_i is the allocated channel bandwidth (Hz); and r_i is the achievable data rate on the channel i (bps). Importantly, the main results of this paper would also apply in case of any rate function of the form $r_i = A_i \ln(1 + B_i p_i)$ for all positive parameters A_i and B_i, which may reflect additional realistic channel properties.

All K channels in our model are assumed to be non-interfering, which may be, for instance, due to the use of non-overlapping frequencies by the co-located RATs. Finally, we assume that the device can adaptively control its transmit power per channel $\{p_i^{tx}\}_{i=1}^K$ by utilizing the information about the available RATs, such as $\{w_i, \gamma_i, p_i^c, p_i^{\max}\}_{i=1}^K$.

The total data rate of the device is thus the sum of *individual* data rates on all channels: $r = \sum_{i=1}^K r_i$. Respectively, the total transmit power is the sum of transmit powers allocated on every channel: $p^{tx}(\mathbf{r}) = p^{tx}(r_1, \ldots, r_K) = \sum_{i=1}^K p_i^{tx}(r_i)$, where $\mathbf{r} = (r_1, \ldots, r_K) \in R^K$ is the vector of transmit powers. The overall device power consumption is, therefore, $p(\mathbf{r}) = p^{tx}(\mathbf{r}) + p_c$, where $p_c = \sum_{i=1}^K p_i^c$. To this end, and for the sake of simplicity, we assume that $p_i^c > 0$ for any interface i, whether it is active or not.

Alternative formulations are also possible (see, e.g., [36]) and the case when $p_i^c = 0$ if $p_i^{tx} = 0$ will be briefly discussed in Section 3.2.2. In what follows, we seek to establish the optimal power control discipline by accounting for both device data rate and its power consumption.

2.2. Discussion of the assumptions

Generally, the channel capacity is known to be the maximum data rate at which reliable communication is possible in the system. It is determined by the signal-to-noise ratio which reflects the characteristics of the signal propagation. In practice, it depends on numerous factors, such as the properties of the wireless medium, antenna heights, distance to the receiver, etc.

Conveniently, the Shannon–Hartley theorem (1) provides a comprehensive abstraction which makes performance evaluation of a practical wireless system analytically tractable. It models the fact that the user device may reduce its transmit power consumption by sacrificing some of its data rate, which is often preferred for small-scale battery powered mobile devices [37]. However, when doing so, the user device shall also ensure that its target bit-rate requirement $r_0 > 0$ is satisfied. Therefore, we expect that the transmit power may be allocated by the device intelligently to extend its battery life.

Whereas the use of the Shannon–Hartley theorem as the power–rate mapping function may provide intuition on the possible techniques for transmit power reduction, the approach of this paper is not restricted to it and can be replicated for a broader class of functions, which satisfy the following criteria:

1. The relationship between the transmit power and the data rate on the channel i is represented by a bijective function $p_i^{tx}(r_i)$, such that $p_i^{tx}(0) = 0$.
2. The derivative $\frac{dp_i^{tx}}{dr_i} > 0$ over the interval $[0, \infty)$, i.e., the function $p_i^{tx}(r_i)$ is continuously differentiable and monotonically increasing. Further, $\frac{d^2 p_i^{tx}}{dr_i^2} > 0$, i.e., the function $\frac{dp_i^{tx}}{dr_i}$ is also monotonically increasing with r_i.

3. Energy efficiency optimization problem

In this section, we focus on maximizing the energy efficiency of a multi-radio mobile device. We begin with the general problem formulation not restricted to a particular power–rate mapping function. We consider as a variable the achievable data rate on each available communication channel (which corresponds to the respective transmit power). We then list some realistic restrictions and introduce the constrained energy efficiency optimization problem based on the Shannon–Hartley theorem. Further, we aim to solve this optimization problem directly by employing the Karush–Kuhn–Tucker approach [38] and formulating a system of equations and inequalities. To obtain the optimal solution, we establish the stationary points of the objective function without any restrictions and solve the unconstrained problem to finally deal with the original task under given constraints.

3.1. General statements

3.1.1. Practical constraints

We start by introducing important realistic restrictions on the energy efficient operation of a multi-radio mobile device. According to the assumptions of Section 2.1, we consider the minimum total data rate on all channels (it implies, in turn, that at least one of K available channels is used):

$$r = \sum_{i=1}^{K} r_i \geq r_0 > 0.$$

We also set a natural restriction on the achievable data rate r_i (and, hence, on p_i^{tx}), such that it cannot be negative:

$$r_i \geq 0, \quad i = \overline{1, K},$$

where the expression $\overline{1, K}$ denotes the set of indexes $1, 2, \ldots, K-1, K$. Finally, we take into consideration the maximum allowed transmit power:

$$p_i^{tx}(r_i) \leq p_i^{\max}, \quad i = \overline{1, K}.$$

In what follows, we will refer to p_i^{tx} as p_i for the sake of brevity. Also, since the function $p_i(r_i)$ is bijective, we can use the equivalent formulation:

$$r_i \leq r_i^{\max}, \quad i = \overline{1, K},$$

where $r_i^{\max} = r_i(p_i^{\max})$ can be given by the function inverse to (1).

3.1.2. Objective function

We concentrate on maximizing the user device energy efficiency represented by the ratio of the total data rate r to the total power p spent by the user device:

$$\eta(\mathbf{r}) = \eta(r_1, \ldots, r_K) = \frac{r}{p} = \frac{\sum\limits_{i=1}^{K} r_i}{\sum\limits_{i=1}^{K} p_i(r_i) + p_c},$$

where r_i and p_i are the data rate and power allocation on the channel i, and $p_c = \sum_{i=1}^{K} p_i^c$. We thus formulate the original optimization problem in terms of the user device energy efficiency as follows.

Original Constrained Problem (OCP):

$$\max_{\{r_i\}_{i=1}^{K}} \eta(\mathbf{r}) = \max_{\{r_i\}_{i=1}^{K}} \frac{\sum\limits_{i=1}^{K} r_i}{\sum\limits_{i=1}^{K} p_i(r_i) + p_c},$$

which is subject to the constraints described in Section 3.1.1. Hence, the total energy efficient data rate follows from the optimal vector of individual data rates on each channel:

$$\mathbf{r}^* = \arg \max_{\{r_i\}_{i=1}^{K}} \eta(\mathbf{r}).$$

As the OCP may be complex to solve in its current form, we notice that it might actually be easier to instead consider the minimization of a function $U(\mathbf{r})$, which is reciprocal to $\eta(\mathbf{r})$. Such a transformation is possible if $\sum_{i=1}^{K} r_i \neq 0$, and we will be addressing this equivalent optimization problem further on. We note here that $\eta(\mathbf{r}) > 0$ for any vector $\mathbf{r}$, such that its components are non-negative, i.e., $r_i \geq 0,\ i = \overline{1, K}$.

Equivalent Constrained Problem (ECP):

$$\min_{\{r_i\}_{i=1}^{K}} \frac{1}{\eta(\mathbf{r})} = \min_{\{r_i\}_{i=1}^{K}} U(\mathbf{r}) = \min_{\{r_i\}_{i=1}^{K}} \frac{\sum\limits_{i=1}^{K} p_i(r_i) + p_c}{\sum\limits_{i=1}^{K} r_i}, \tag{2}$$

which is also subject to the constraints described in Section 3.1.1.

Finally, we reformulate the ECP as:

$$\min_{\{r_i\}_{i=1}^K} U(\mathbf{r}) = \min_{\{r_i\}_{i=1}^K} \frac{\sum_{i=1}^{K} p_i(r_i) + p_c}{\sum_{i=1}^{K} r_i},$$

subject to

$$\begin{aligned} &\phi(\mathbf{r}) = r_0 - \sum_{i=0}^{K} r_i \leq 0, \\ &f_i(r_i) = -r_i \leq 0, \quad i = \overline{1, K}, \\ &g_i(r_i) = r_i - r_i^{\max} \leq 0, \quad i = \overline{1, K}. \end{aligned} \tag{3}$$

3.1.3. Karush–Kuhn–Tucker conditions

In order to tackle the ECP under the given inequality constraints, we employ the Karush–Kuhn–Tucker (KKT) approach. Below, we list the regularity KKT conditions and obtain a system to solve in order to find the optimal solution of the ECP:

$$\begin{aligned} &\frac{\partial U(\mathbf{r})}{\partial r_i} + \sum_{i=1}^{K} \lambda_i \frac{dg_i(r_i)}{dr_i} + \sum_{i=1}^{K} \mu_i \frac{df_i(r_i)}{dr_i} + \beta \frac{d\phi(\mathbf{r})}{dr_i} = 0 \\ &\Leftrightarrow \frac{\frac{dp_i}{dr_i} \cdot r - \left(\sum_{i=1}^{K} p_i + p_c \right)}{r^2} + \lambda_i - \mu_i - \beta = 0. \end{aligned}$$

The primal feasibility conditions may be given as:

$$\begin{aligned} &r_i - r_i^{\max} \leq 0, \quad i = \overline{1, K}, \\ &r_i \geq 0, \quad i = \overline{1, K}, \\ &\sum_{i=1}^{K} r_i - r_0 \geq 0. \end{aligned} \tag{4}$$

Further, the dual feasibility conditions are represented by the following inequalities:

$$\begin{aligned} &\lambda_i \geq 0, \quad i = \overline{1, K}, \\ &\mu_i \geq 0, \quad i = \overline{1, K}, \\ &\beta \geq 0, \end{aligned}$$

where λ_i, μ_i, and β are the KKT multipliers.

Finally, the complementary slackness conditions are as follows:

$$\begin{aligned} &\lambda_i (r_i - r_i^{\max}) = 0, \quad i = \overline{1, K}, \\ &\mu_i r_i = 0, \quad i = \overline{1, K}, \\ &\beta \left(\sum_{i=1}^{K} r_i - r_0 \right) = 0. \end{aligned}$$

Summarizing, in order to obtain the optimal solution for the ECP under the above constraints, a system of $3K+1$ equations and $4K + 2$ inequalities has to be solved. Importantly, the domain bounded by the given inequalities has to be non-empty. Otherwise, the entire problem does not have a solution.

Noteworthy, the KKT conditions by themselves do not provide a method for finding the maximum/minimum points. Instead, they only determine the stationary points (where the gradient is zero) among which the minimum point can be located. In general, solving the system of many equations and inequalities is known to be difficult and in what follows we detail our approach to tackle the target optimization problem.

3.2. Solving the constrained problem

The Equivalent Constrained Problem (ECP) detailed in Section 3.1.2 can generally be formulated for an arbitrary power–rate mapping function which satisfies the conditions in Section 2.2. For the purposes of illustration, we further consider an example, where the relationship between the achievable data rate and the required transmit power is given by the Shannon–Hartley theorem (1). However, extended formulations are also possible, which may account for more advanced features of practical wireless systems.

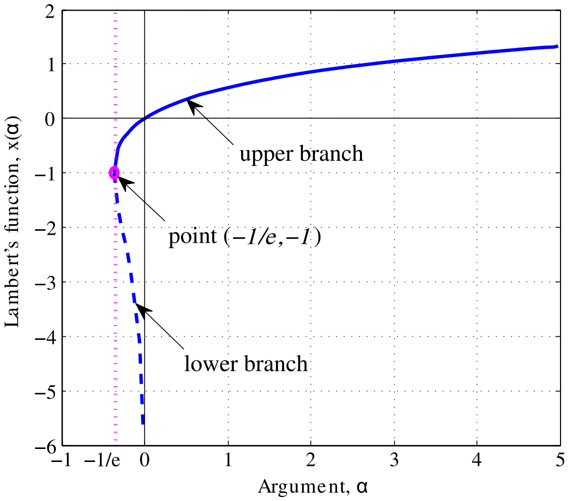


Fig. 3. A plot of the Lambert's function.

3.2.1. Considering the unconstrained problem

To locate the stationary points lying both in the considered domain and outside of it, we fall back to the simpler optimization problem with the same objective function (2) for $\mathbf{r} \neq \mathbf{0}$, but without any constraints. We will further on apply this solution to the target ECP (3).

Equivalent Unconstrained Problem (EUP)

$$\min_{\{r_i\}_{i=1}^K} U(\mathbf{r}) = \min_{\{r_i\}_{i=1}^K} \frac{\sum_{i=1}^K p_i + p_c}{\sum_{i=1}^K r_i}. \tag{5}$$

Theorem 1. *The optimal solution of EUP (5) may exist only if the transcendental equation $Xe^X = \alpha$ has real roots:*

$$\alpha = \frac{-\sum_{j=1}^K \frac{1}{\gamma_j} + \sum_{j=1}^K p_j^c}{e \sum_{k=1}^K w_k} \prod_{j=1}^K w_j^{\frac{w_j}{\sum_{k=1}^K w_k}} \prod_{j=1}^K \gamma_j^{\frac{w_j}{\sum_{k=1}^K w_k}}.$$

The optimal vector of the required transmit powers $\mathbf{p} = (p_1, \ldots, p_K)$ is then delivered by:

$$p_i = -\frac{1}{\gamma_i} + w_i e \prod_{j=1}^K \left(w_j \gamma_j\right)^{\frac{-w_j}{\sum_{k=1}^K w_k}} \exp W(\alpha), \quad i = \overline{1, K}, \tag{6}$$

where $W(\alpha)$ is a root of the transcendental equation (see Fig. 3 for illustration) and is known as the Lambert's function [39]. The procedure of selecting a particular root appropriately depends on the value of α. For the sake of exposition, it is detailed separately in Proposition 1 below.

Proof. Proof is given in Appendix A. □

The transcendental equation $Xe^X = \alpha$ has real roots if $\alpha \geq -1/e$. In this case, there might be either one or two real root(s) depending on the branches of the function $W(\alpha)$ [39]. Let us now discuss how to choose the appropriate branch of the Lambert's function $W(\alpha)$. We denote the root obtained via the upper branch as $X_0 = W_0(\alpha)$ and the one obtained via the lower branch as $X_1 = W_1(\alpha)$ for all $\alpha \in [-\frac{1}{e}, 0)$.

Proposition 1. *For the given system parameters, four different cases are possible:*

1. *If $\alpha < -\frac{1}{e}$, the equation $Xe^X = \alpha$ does not have real roots, and no stationary point of the function $U(\mathbf{r})$ exists.*
2. *If $\alpha = -\frac{1}{e}$, the equation $Xe^X = \alpha$ has one real root $X = -1$ and, therefore, no more than one stationary point of the function $U(\mathbf{r})$ may exist.*

3. *If $-\frac{1}{e} < \alpha < 0$, there exist exactly two different stationary points $\mathbf{p}^{(0)}$ and $\mathbf{p}^{(1)}$ corresponding to X_0 and X_1 respectively. Furthermore, for all $i = \overline{1,K}$: $p_i^{(0)} > p_i^{(1)}$, and at least one of the components of the vector $\mathbf{p}^{(1)}$ is negative, i.e., there is an index $i_0 \in \{i\}_{i=1}^{K}$, such that $p_{i_0}^{(1)} < 0$.*
4. *If $\alpha \in (0, \infty)$, there exists exactly one optimal solution, i.e., the optimal vector of powers $\mathbf{p}$ has only one value.*

Proof. The first, second, and fourth cases are rather trivial and follow from the properties of the Lambert's function. However, all these cases do not provide us with stationary points. For that reason, the third case is addressed in more detail in Proposition 2 below.

We proceed with the meaningful third case to solve the EUP. The equation $Xe^X = \alpha$ has two real roots *iff* $-\frac{1}{e} < \alpha < 0$. Hence, there exist exactly two different stationary points $\mathbf{p}^{(0)}$ and $\mathbf{p}^{(1)}$ related to the roots of the equation. If $\alpha \in \{-\frac{1}{e}\} \cup (0, \infty)$, the vector $\mathbf{p}$ has only one value also due to the properties of the Lambert's function. Now we consider the interval $-\frac{1}{e} < \alpha < 0$ and we will show that $p_i^{(0)} > p_i^{(1)}$ for all $i = \overline{1,K}$.

We rewrite the expression (6) for the allocated power:

$$p_i^{(0)} = \frac{w_i e^{W_0(\alpha)+1}}{\prod\limits_{k=1}^{K} (w_k\gamma_k)^{\frac{w_k}{\sum_{j=1}^{K} w_j}}} - \frac{1}{\gamma_i}, \qquad p_i^{(1)} = \frac{w_i e^{W_1(\alpha)+1}}{\prod\limits_{k=1}^{K} (w_k\gamma_k)^{\frac{w_k}{\sum_{j=1}^{K} w_j}}} - \frac{1}{\gamma_i}.$$

Since $W_0(\alpha) > W_1(\alpha)$, we conclude that $p_i^{(0)} > p_i^{(1)}$.

Further, we consider the greater value $p_i^{(1)}$ within the domain $-\frac{1}{e} \le \alpha < 0$ and account for the fact that $W_1(\alpha) < -1$:

$$p_i^{(1)} = \frac{w_i e^{W_1(\alpha)+1}}{\prod\limits_{k=1}^{K} (w_k\gamma_k)^{\frac{w_k}{\sum_{j=1}^{K} w_j}}} - \frac{1}{\gamma_i} < \frac{w_i\gamma_i - \prod\limits_{k=1}^{K} (w_k\gamma_k)^{\frac{w_k}{\sum_{j=1}^{K} w_j}}}{\prod\limits_{k=1}^{K} (w_k\gamma_k)^{\frac{w_k}{\sum_{j=1}^{K} w_j}} \gamma_i}. \tag{7}$$

We take the index i_0, such that $i_0 = \arg\min_i (w_i\gamma_i)$. Hence, $w_{i_0}\gamma_{i_0} \le w_i\gamma_i$ for any $i = \overline{1,K}$ and:

$$\prod_{k=1}^{K} (w_k\gamma_k)^{\frac{w_k}{\sum_{j=1}^{K} w_j}} \ge \prod_{k=1}^{K} \left(w_{i_0}\gamma_{i_0}\right)^{\frac{w_k}{\sum_{j=1}^{K} w_j}} = w_{i_0}\gamma_{i_0}. \tag{8}$$

This means that $p_i^{(1)} < 0$ and at least one negative component of vector $\mathbf{p}^{(1)}$ exists, which completes the proof. □

Based on this property, we omit the consideration of the lower branch of the function $W(\alpha)$. Further, we consider separately the two possible ranges of α, i.e., $\alpha \ge -1/e$, when there is at least one stationary point, and $\alpha < -1/e$, when there is none.

Case $\alpha \ge -1/e$ (a stationary point exists). Here, by using Proposition 1, we formulate the condition, when a stationary point is the sought maximum.

Proposition 2. *If the vector $\mathbf{p}$ satisfies the condition of Theorem 1 and $\alpha > -1/e$, then:*

1. *The corresponding optimal vector $\mathbf{r}$ obtained via the upper branch of the Lambert's function is a local minimum of the function $U(\mathbf{r})$ and, respectively, a local maximum of the function $\eta(\mathbf{r})$.*
2. *The corresponding optimal vector $\mathbf{r}$ obtained via the lower branch of the Lambert's function ($-1/e < \alpha < 0$) is a local maximum of the function $U(\mathbf{r})$ and, respectively, a local minimum of the function $\eta(\mathbf{r})$.*

Proof. Let us consider the cases 2, 3, and 4 of Proposition 1 ($\alpha \ge -1/e$ in all of them). We start with the general reasoning for the case $\alpha > -1/e$ and then split the combined cases 3 and 4 of Proposition 1 into the following possibilities: (1) $\alpha > 0$ or $-1/e < \alpha < 0$, when we consider the upper branch of the Lambert's function and (2) $-1/e < \alpha < 0$, when the lower branch of the Lambert's function is considered.

In order to prove the statement of the proposition, we need to calculate the Hessian matrix of the function $U(\mathbf{r})$. Therefore, we begin by calculating the term $\frac{\partial^2 U(\mathbf{r})}{\partial r_j \partial r_k}$, $k \ne j$ and then proceed with $\frac{\partial^2 U(\mathbf{r})}{\partial r_j^2}$. For both calculations we exploit the expression for the first derivative (see the proof of Theorem 1 in Appendix A):

$$\frac{\partial U(\mathbf{r})}{\partial r_j} = \frac{\frac{1}{\gamma_j w_j} e^{\frac{r_j}{w_j}} \sum\limits_{i=1}^{K} r_i - \left[\sum\limits_{i=1}^{K} \frac{1}{\gamma_i}\left(e^{\frac{r_i}{w_i}} - 1\right) + p_c\right]}{\left(\sum\limits_{i=1}^{K} r_i\right)^2}, \quad j = \overline{1,K}.$$

Hence, the non-diagonal element of the Hessian matrix, i.e., the second derivative with respect to $r_j, r_k,\ k \neq j$, follows as:

$$\frac{\partial^2 U(\mathbf{r})}{\partial r_j \partial r_k} = \frac{\frac{1}{\gamma_j w_j} e^{\frac{r_j}{w_j}} - \frac{1}{\gamma_k w_k} e^{\frac{r_k}{w_k}}}{\left(\sum\limits_{i=1}^{K} r_i\right)^2} = -\frac{\partial^2 U(\mathbf{r})}{\partial r_j \partial r_k} = 0, \quad j, k = \overline{1, K},$$

since $\frac{U(\mathbf{r})}{\partial r_j \partial r_k} = \frac{U(\mathbf{r})}{\partial r_k \partial r_j}$. Therefore, the Hessian matrix is diagonal for both the cases and we may further discuss them separately.

Let us find the diagonal element of the Hessian matrix, that is, the second derivative with respect to r_j:

$$\left.\frac{\partial^2 U(\mathbf{r})}{\partial r_j^2}\right|_{\mathbf{r}} = \frac{\frac{1}{\gamma_j w_j} \frac{1}{w_j} e^{\frac{r_j}{w_j}} \sum\limits_{i=1}^{K} r_i + 2\left(\sum\limits_{i=1}^{K} r_i\right)^{-1} \cdot 0}{\left(\sum\limits_{i=1}^{K} r_i\right)^2}, \quad j = \overline{1, K},$$

where we take into account the fact that the optimal vector satisfies the equality $\frac{\partial U(\mathbf{r})}{\partial r_j} = 0$ for all $j = \overline{1, K}$.

Using the expression (A.12) in the proof of Theorem 1, we may calculate $\sum_{i=1}^{K} r_i$:

$$\sum_{i=1}^{K} r_i = \sum_{i=1}^{K} w_i \ln(\gamma_i w_i) + \sum_{i=1}^{K} w_i \left\{ 1 - \frac{\sum\limits_{j=1}^{K} w_j \ln(w_j \gamma_j)}{\sum\limits_{k=1}^{K} w_k} + W(\alpha) \right\}$$
$$= [1 + W(\alpha)] \sum_{i=1}^{K} w_i,$$

where α is given by Theorem 1. Let us consider two different cases described in the formulation of this theorem.

1. It is known that for the upper branch of the Lambert's function $W(\alpha) \geq -1$ if $\alpha \geq -1/e$. Furthermore, $W(\alpha) = -1$ only at the point $\alpha = -1/e$. Hence, the expression $\sum_{i=1}^{K} r_i$, and, consequently, $\frac{\partial^2 U(\mathbf{r})}{\partial r_j^2}$ are always greater than zero at the stationary point $\mathbf{r}$ as long as $\alpha > -1/e$ and we take the upper branch of the Lambert's function.

   If $\alpha > -1/e$ and the upper branch of the Lambert's function is considered, then $\frac{\partial^2 U(\mathbf{r})}{\partial r_j^2} > 0$ and the Hessian matrix is positive definite at the point $\mathbf{r}$, which implies the local minimum of the function $U(\mathbf{r})$.
2. If $-1/e < \alpha < 0$ and the lower branch of the Lambert's function is considered, we have $W(\alpha) < -1$ and, hence, negative eigenvalues of the diagonal Hessian matrix $\frac{\partial^2 U(\mathbf{r})}{\partial r_j^2} < 0$. This, in turn, makes the Hessian matrix negative definite and the point $\mathbf{r}$ is then a local maximum of the function $U(\mathbf{r})$.

Finally, we note that if $\alpha = -1/e$ then $W(\alpha) = -1$ and $\frac{\partial^2 U(\mathbf{r})}{\partial r_j^2}$ equals zero, which leads to a singular Hessian matrix. However, in this case $\sum_{i=1}^{K} r_i = 0$ as well and the obtained point clearly does not lie within the domain of the function $U(\mathbf{r})$. Moreover, this point is also out of our interest in terms of the initial optimization problem, since the condition $\sum_{i=1}^{K} r_i = r_0 > 0$ has been given in advance. □

According to Proposition 1 in case when $\alpha > -1/e$, there is one and only one local minimum of the function $U(\mathbf{r})$ (that is, local maximum of $\eta(\mathbf{r})$).

Corollary 1. *If $\alpha > -1/e$ and the upper branch of the Lambert's function is taken, the stationary point $\mathbf{r} > 0$ constitutes the global maximum of the function $\eta(\mathbf{r})$.*

We note that in case $\alpha = -1/e$, we obtain $\sum_{i=1}^{K} r_i = 0$, that does not belong to the considered domain of the function $U(\mathbf{r})$ and is not in the area of our interest due to the target bit-rate requirement $r_0 > 0$.

The above approach will locate the maximum, if the stationary point exists. However, it may also be the case that $\alpha < -\frac{1}{e}$ and the Lambert's function does not have real roots. Then the stationary points do not exist and there is no local minimum for the considered function.

Case $\alpha < -1/e$ (a stationary point does not exist). The system of inequalities (3) induces the domain for the objective function, which is compact due to boundedness and closedness in R^K. Since we consider a compact set in R^K as the domain within which the objective function is bounded and continuous, the maximum/minimum can be reached only at the border of this domain when no stationary point exists ($\alpha < -\frac{1}{e}$).

Proposition 3. *If $\alpha < -\frac{1}{e}$, then there is no local maximum of the function $\eta(\mathbf{r})$ (or, local minimum of $U(\mathbf{r})$). Assuming that all the components of the rate vector are naturally not less than zero, i.e., $\mathbf{r} \geq 0$, the solution to the problem at hand (constrained only by $\mathbf{r} \geq 0$) can then be found on the plane given by the equality $r_i = 0,\ 1 \leq i \leq K$. The selection of the component i is detailed below.*

Proof. Let us fix the index i and the values $r_j, j \neq i$, and consider the function $\tilde{\eta}(r_i) = \frac{r_i + \sum_{j=1,i\neq j}^{K} r_j}{p_i(r_i) + \sum_{j=1,i\neq j}^{K} p_j(r_j)}$. We demonstrate below that $\tilde{\eta}(r_i) \to 0$ in case of $r_i \to +\infty$ and $\tilde{\eta}(0)$ tends to a positive constant, if $r_i \to +0$, which (due to monotonic $\tilde{\eta}(r_i)$) implies that the maximum should be sought on one of the planes defined at $r_i = 0$. Due to the assumption that the function $\frac{dp_i}{dr_i}$ is continuously differentiable and monotonically increasing within $(0, \infty)$, and accounting for L'Hospital's rule:

$$\lim_{r_i \to +\infty} \frac{r_i + \sum\limits_{j=1,i\neq j}^{K} r_j}{p_i(r_i) + \sum\limits_{j=1,i\neq j}^{K} p_j(r_j)} = \lim_{r_i \to +\infty} \frac{1}{\frac{dp(r_i)}{dr_i}} = 0. \tag{9}$$

Similar reasoning may be repeated for the function $\tilde{\eta}(\mathbf{r}) = \frac{\sum_{i=1}^{K} r_i}{\sum_{i=1}^{K} p_i + p_c}$, where a set $\{r_i^{(0)}\}_{i=1}^{k_1}$, $k_1 = \overline{0, K-1}$ is fixed and for all $i = \overline{k_1+1, K}$ let $r_i \to \infty$. Then, it delivers us:

$$\lim_{r_i \to \infty, i=\overline{k_1+1,K}} \frac{\sum\limits_{i=1}^{k_1} r_i^{(0)} + \sum\limits_{i=k_1+1}^{K} r_i^{(0)}}{\sum\limits_{i=1}^{k_1} \frac{1}{\gamma_i} e^{\frac{r_i^{(0)}}{w_i}} + \sum\limits_{i=k_1+1}^{K} \frac{1}{\gamma_i} e^{\frac{r_i}{w_i}} + p_c} = \lim_{r_i \to \infty, i=\overline{k_1+1,K}} \frac{\sum\limits_{i=k_1+1}^{K} r_i^{(0)}}{\sum\limits_{i=k_1+1}^{K} \frac{1}{\gamma_i} e^{\frac{r_i}{w_i}}} = +0. \tag{10}$$

The expression above implies that when all or several components of vector $\mathbf{r}$ tend to infinity, the value of the function $\eta(\mathbf{r})$ is infinitesimal.

If the system parameters do not satisfy the condition $\alpha > -1/e$, then the gradient $\nabla\eta \neq 0$. This means that the non-negative function $\eta(\mathbf{r})$ achieves its maximum at the border of any compact subset $\tilde{r} = \{\mathbf{r} | r_i \in [0, C_1]\}$, where $C_1 \in R^+$ is any number large enough. We note that for at least one of the non-negative functions $p_i(r_i)$: $\frac{dp(r_i)}{dr_i} > 0$. We derive the limit of the function $\eta(\mathbf{r})$ if $r_i \to 0$:

$$\lim_{r_i \to +0} \frac{r_i + \sum\limits_{j=1,i\neq j}^{K} r_j}{p_i(r_i) + \sum\limits_{j=1,i\neq j}^{K} p_j(r_j)} = \lim_{r_i \to +0} \frac{1}{\frac{dp(r_i)}{dr_i}} = \text{const} > 0. \tag{11}$$

From the expression for the derivative of $U(\mathbf{r})$, we can conclude that $\frac{\partial U(\mathbf{r})}{\partial r_j} < 0$ for the domain $r_j > r_j^*, j = \overline{1, K}$, where $\mathbf{r}^*$ is a stationary point of the function $U(\mathbf{r})$. Hence, the function $U(\mathbf{r})$ is monotonically decreasing within the domain $r_j > r_j^*,\ j = \overline{1, K}$.

From the statements (9) and (11), we also conclude that the maximum value for the function η lies on one of the borders $r_i = 0$, which completes the proof. □

Therefore, we establish that the point of the global maximum for the target energy efficiency function has at least one zero component. Hence, we need to decrease the dimension of the problem on hand and for that matter we exclude the component with the least contribution to the growth of the function $\eta(\mathbf{r})$, i.e., find:

$$i = \arg\max_{j=1,K} \left\{ \left. \frac{dp_j(r_j)}{dr_j} \right|_{r_j=0} \right\}.$$

Then we set $r_i = 0$ and thus decrease the dimension of the optimization task to $K - 1$, while considering the following:

$$\min_{\{r_j\}_{j=1,j\neq i}^{K}} U(\mathbf{r}) = \min_{\{r_j\}_{j=1,j\neq i}^{K}} \frac{\sum\limits_{j=1,j\neq i}^{K} p_j(r_j) + p_i + \sum\limits_{j=1}^{K} p_j^r}{\sum\limits_{j=1,j\neq i}^{K} r_j + r_i},$$

subject to the same conditions. Here, r_i and $p_i = p_i(r_i)$ are the fixed values for the dimension i, whereas r_i can either be equal to zero or follow from another border condition (more details are given below). We proceed by solving the optimization problem of the dimension $K-1$ or by decreasing the dimension further if necessary, as long as $K \geq 1$. If $K = 1$, then we arrive at the one-dimensional problem with a particular solution r_j^*, which has to be again checked for the border restrictions.

3.2.2. Considering the constrained problem

Using the obtained solution r^* of the unconstrained problem EUP, we now describe our approach to solving the system of inequalities (4) step by step. First, we assume that the argument of the Lambert's function exceeds $-\frac{1}{e}$, so that the objective function $U(\mathbf{r})$ of the constrained problem ECP (3) has a local minimum in the corresponding point r^*. Naturally, r^* may not meet the conditions of the system (4).

In order to find the solution of the ECP, we need to check that the local minimum r^* satisfies all the inequalities:

$$\begin{aligned}
&r_i \geq 0, \quad i = \overline{1, K},\\
&r_i - r_i^{\max} \leq 0, \quad i = \overline{1, K},\\
&\sum_{i=1}^{K} r_i - r_0 \geq 0.
\end{aligned}$$

If the point r^* satisfies the relations above, it is the solution of the system (4) and, consequently, the optimal solution of the problem (3). Otherwise, if at least one of the inequalities fails, the local minimum of the function $U(\mathbf{r})$ is located outside of the domain specified by the given constraints, and the optimal solution belongs to the border of this domain. The general scheme of our solution is detailed by Algorithm 1.

In particular, if we consider effective constraints $r_i - r_i^{\max} = 0$, decreasing the dimension of the ECP can be done similarly until $K = 1$. The difference with the EUP is in the fact that when moving to the problem of dimension $K-1$, we fix the components at non-zero values $r_K = r_K^{\max}$ if the constraint of index K fails (without loss of generality, we can rearrange the vector). Hence, we arrive at slightly different expressions as given by Theorem 2 below.

Theorem 2. *If the ECP has a solution $r^* \in R^K$ and there is at least one index i, such that $r_i > r_i^{\max}$, then the optimal solution should be located among:*

$$\begin{aligned}
&p_i = e^{1-\frac{r_K}{w_K}} \prod_{k=1}^{K-1} (w_k \gamma_k)^{\frac{-w_k}{\sum_{j=1}^{K-1} w_j}} \cdot e^{W(\alpha)} - \frac{1}{\gamma_i}, \quad i = \overline{1, K-1},\\
&p_K = \frac{1}{\gamma_K}\left(e^{\frac{r_k^{\max}}{w_K}} - 1\right).
\end{aligned}$$

Proof. Proof is given in Appendix B. □

For the case when the constraint $\sum_{i=1}^{K} r_i - r_0 = 0$ does not hold, we apply the similar procedure and set $r_K = r_0 - \sum_{i=1}^{K-1} r_i$. The final expressions are given in Theorem 3.

Theorem 3. *If the ECP has a solution $r^* \in R^K$ and $\sum_{i=1}^{K} r_i - r_0 < 0$, then the optimal solution should be located among:*

$$p_i = \frac{1}{\gamma_i}\left(e^{\frac{r_i}{w_i}} - 1\right), \quad i = \overline{1, K},$$

where the vector $\{r_i\}_{i=1}^{K}$ is given as follows:

$$\begin{aligned}
&r_i = w_i \frac{r_0 - \sum_{j=1}^{K} w_j \ln(\gamma_j w_j)}{\sum_{j=1}^{K} w_j} + w_i \ln(\gamma_i w_i), \quad i = \overline{1, K-1},\\
&r_K = r_0 - \sum_{j=1}^{K-1} r_j.
\end{aligned}$$

Proof. Proof is given in Appendix C. □

Finally, let us also discuss the situation when $p_i^c = 0$, if a particular channel (radio technology) is not exploited (i.e., $p_i = 0$). In Section 2, we assume that $p_i^c = q_i > 0$, where q_i is a fixed constant. However, due to the recursion in calculations,

Algorithm 1 Solving the constrained problem ECP

```
$\mathbf{r}^* \leftarrow$ FindSolution($K$, $\{i\}_{i=1}^K$)

function FindSolution($K$, $\{i_j\}_{j=1}^K$)
    while $K \geq 1$ do
        Solve the EUP (unconstrained problem) as per Section 3.2.1
        $\mathbf{r}^* \leftarrow$ stationary point
        if $\mathbf{r}^* \in \mathbb{R}^K$ then
            if $r_i^*$<0 then
                $i_0 \leftarrow \max_{\{i\}_{i=1}^K | r_i^* < 0} \frac{dp_i}{dr_i}$
                $r_{i_0}^* \leftarrow 0$
                $\{r_i^*\}_{i=1, i \neq i_0}^K$ $\leftarrow$FindSolution($K - 1$, $\{i\}_{i=1, i \neq i_0}^K$)
            end if

            if $r_i^* > r_i^{\max}$ then
                $i_0 \leftarrow \max_{\{i\}_{i=1}^K | r_i^* > r_i^{\max}} \frac{dp_i}{dr_i}$
                $r_{i_0}^* \leftarrow r_i^{\max}$
                $\{r_i^*\}_{i=1, i \neq i_0}^K$$\leftarrow$FindSolutionMax($K - 1$, $\{i\}_{i=1, i \neq i_0}^K$)
            end if

            if $\sum_{i=1}^K r_i^* < r_0$ then
                $r_K^* = r_0 - \sum_{i=1}^K r_i^*$
                $\{r_i^*\}_{i=1}^{K-1}$ $\leftarrow$ FindSolutionSum($K - 1$, $\{i\}_{i=1}^{K-1}$)
            end if
            return $r^*$ // $r^*$ is the point of interest

        else
            // $\mathbf{r}^* \notin R^K$, there is no local maximum
            $r^*$ is on the border $r_i^* = 0$
            $i_0 \leftarrow \max_{\{i\}_{i=1}^K} \frac{dp_i}{dr_i}$
            $r_{i_0}^* \leftarrow 0$
            $\{r_i^*\}_{i=1, i \neq i_0}^K$ $\leftarrow$ FindSolution($K - 1$, $\{i\}_{i=1, i \neq i_0}^K$)
        end if
    end while

end function

function FindSolutionMax($K$, $\{i_j\}_{j=1}^K$)
    Solve according to Theorem 2
    return $r^*$
end function

function FindSolutionSum($K$, $\{i_j\}_{j=1}^K$)
    Solve according to Theorem 3
    return $r^*$
end function
```

we may relax this condition and consider the function $p_i^c = q_i \cdot \mathbf{I}\{p_i = 0\}$, where $\mathbf{I}\{p_i = 0\}$ is the indicator function of an event. Therefore, after obtaining $p_i = 0$, we would continue with the optimization at the next iteration using $\tilde{p}_c = p_c - p_i^c$, so that the considered radio interface is inactive.

4. Numerical results

In this section, we provide three illustrative scenarios to investigate the achievable data rate, as well as the associated transmit power, and energy efficiency of the mobile user device. These are intended to exemplify the energy efficient operation achieved with our approach and compare it against simpler heuristic power control schemes. Below we introduce several (abstract and realistic) network geometries to study the user device behavior.

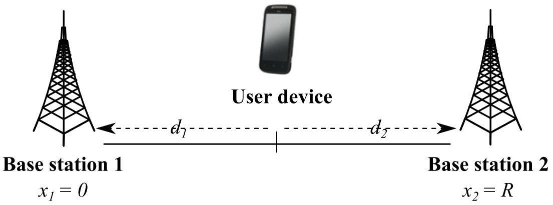


Fig. 4. Network geometry for the case of two RATs.

Table 2
System parameters.

	Parameter description	LTE	WiFi
r_0	Target user bit-rate	2.5 Mbps	2.5 Mbps
$p_i^{\max}$	Maximum device Tx power	23 dBm	20 dBm
$r_i^{\max}$	Maximum user data rate	5.7 Mbps	5.0 Mbps
p_i^c	Circuit power	0.20 W	0.10 W
w_i	Channel bandwidth (excluding overheads)	10 MHz	10.8 MHz
N_0	Noise power	−90 dB	−90 dB
f_c	Carrier frequency	2.0 GHz	2.4 GHz
h_s	Base station antenna height	25 m	10 m
h_u	Device antenna height	1.5 m	1.5 m
	Number of users (sector/AP)	10	10

4.1. Symmetric case. Two radio access technologies

Here we assume that there are two RATs ($K = 2$) available for a mobile device to use at any given moment of time (see Fig. 4). In the figure, the base station of the first RAT is located at the point $x_1 = 0$, while the second one is placed at the point $x_2 = R$. The user device is moving along the x-axis between the base stations and its current location is $x \in (0, R)$.

We employ the standard propagation model for micro cells in urban areas [40], where the path loss (dB) is determined as:

$$\begin{aligned} &\theta = 40\log_{10} d + 7.8 - 18\log_{10} h_s h_u + 2\log_{10} f_c, \quad \text{if } 10 < d < d_{BP}, \\ &\theta = 22\log_{10} d + 28 + 20\log_{10} f_c, \quad \text{if } d_{BP} < d < 5000, \end{aligned} \tag{12}$$

where h_s and h_u are the effective heights of the base station (or access point, AP) and the device (m); d and d_{BP} are the distances to the base station (access point) and the break point distance (m); and f_c is the center frequency (GHz). Then, taking into account noise power, we may calculate the SNR per unit of power as $\gamma_i = \frac{1/\theta}{N_0}$.

We study our optimization problem under realistic parameters partly borrowed from [41]. All corresponding parameters are given in Table 2.

For the sake of intuitive illustration, we first provide results for a simplistic scenario without any maximum power and minimum bit-rate restrictions. In Fig. 5, the dependence of the achievable data rate, transmit power, and energy efficiency on the current user device coordinate x is demonstrated. The user device begins by exploiting one RAT as it moves from one base station to another (left to right). We clearly see that our proposed power control increases the transmit power to compensate for the growing path loss (according to the model in (12)). Starting from a certain point ($x = 31$ m in the figure), the user device enables the second RAT by allocating some transmit power on the respective channel. Consequently, it allows decreasing the power on the first channel, while still maintaining the highest energy efficiency.

In the middle point, the transmit power levels on the two channels match due to the symmetry of this example. Generally, the user data rate decreases as the channel quality deteriorates. However, we also observe an interesting effect of some increase in the data rate due to the combined operation by the two RATs. This data rate growth is achieved for the cost of additional transmit power. Hence, at the end, this does not affect the optimal energy efficiency level, which monotonically decreases, as the user device moves form zero to the middle point, and then increases symmetrically.

4.2. Symmetric case. Three radio access technologies

Here, we illustrate the behavior of the proposed optimal device operation on the plane by again using the considered propagation model (12). Now we assume the coexistence of three RATs ($K = 3$) available for use at any given moment of time. As a result, the considered system comprises three similar (e.g., 3GPP LTE) base stations on different channels: at the points $x_1 = (0, R)$ and $x_2 = (R, 0)$, and at the center of coordinates $x_3 = (0, 0)$. The user device is allowed to move across the plane in between all these base stations (see Fig. 6), while its power control is defined by Algorithm 1.

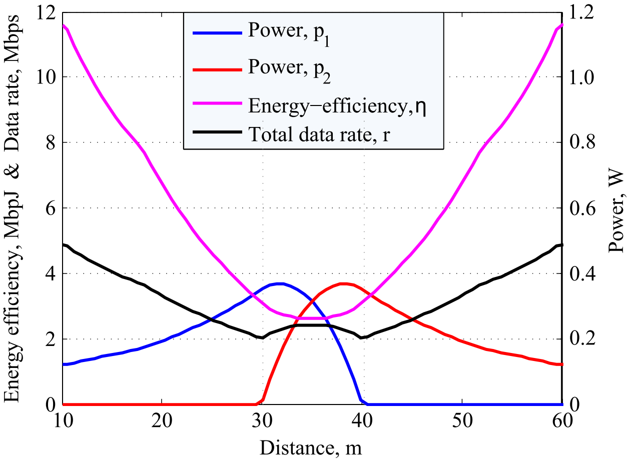


Fig. 5. Performance for $K = 2$, no restrictions.

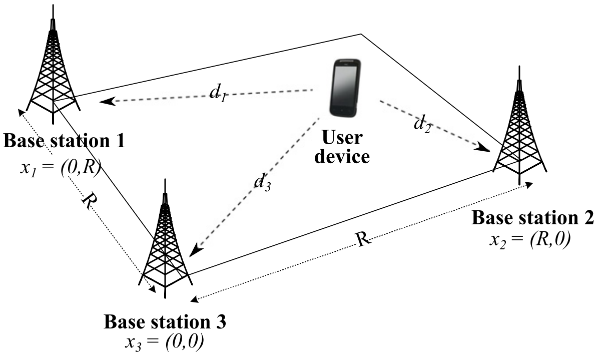


Fig. 6. Network geometry for the case of three RATs.

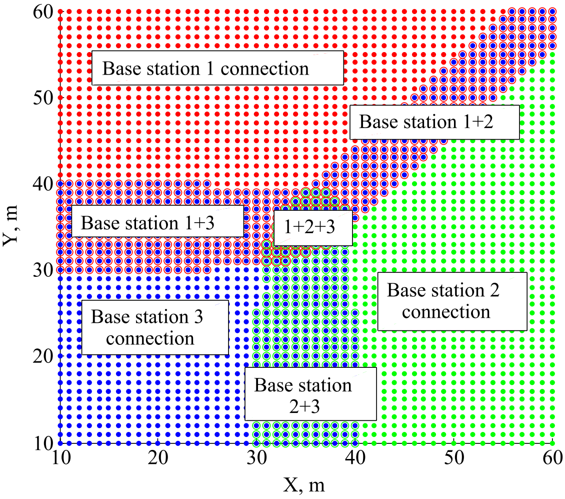


Fig. 7. Areas of user device transmission on different channels.

In Fig. 7, we highlight the areas of combined RAT operation, which are defined by the optimal power levels established with the proposed analysis. We observe several types of regions: three regions where the user device transmits on a single

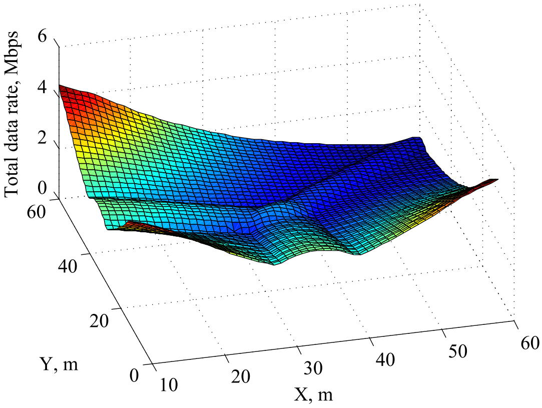


Fig. 8. Data rate for $K = 3$.

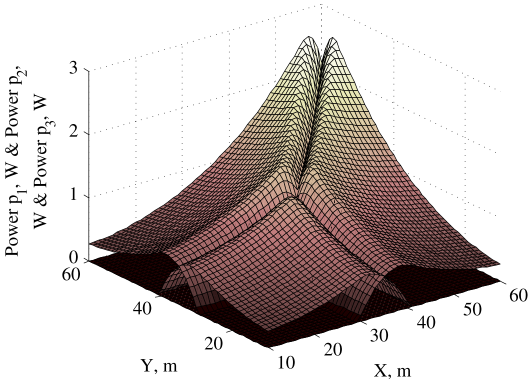


Fig. 9. Transmit power for $K = 3$.

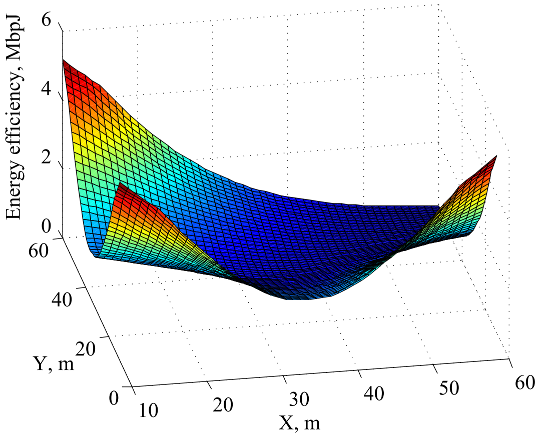


Fig. 10. Energy efficiency for $K = 3$.

channel, three *support regions* between the pairs of base stations (with more than one radio interface activated to support the target bit-rate), and one support region where all the three radio interfaces are active.

Figs. 8–10 demonstrate the dependences of the achievable data rate, transmit power, and energy efficiency on the user device coordinates respectively. The situation is generally similar to the above two-dimensional case ($K = 2$): the data rate decreases as the user device moves away from a base station. We can also clearly see the effect of boosted data rate in

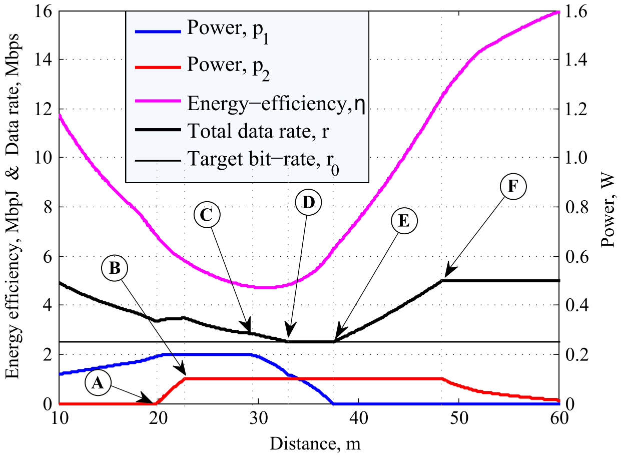


Fig. 11. Performance for $K = 2$ with max power and min bit-rate restrictions.

the areas of collaborative operation by several RATs, which is associated with the decreased transmit power given by our optimal power control.

4.3. Asymmetric case. Two radio access technologies

In what follows, we consider two different radio technologies: 3GPP LTE and WiFi, assuming that a particular user is moving between the LTE base station and the (outdoor) WiFi access point. For simplicity, we assume that WiFi operation (as per IEEE 802.11-2012 technology) is controlled by the Point Coordination Function (PCF) and, therefore, a round-robin scheduler (e.g., in commercial WiFi deployments on traffic lights or lamp posts). All corresponding technology-specific parameters are summarized in Table 2. For both LTE and outdoor WiFi, we employ the propagation model for micro cells in urban areas [40] as before.

Fig. 11 concentrates on the case when the maximum power and the minimum bit-rate restrictions take effect. In the region up to the point Ⓐ, only the LTE interface is sending data, and the transmit power increases due to the growing path loss; from the point Ⓐ to the point Ⓑ the transmit power of WiFi rapidly grows towards its maximum level (due to how WiFi PHY operates). Both the radio interfaces keep the maximum power level up to the point Ⓒ, after which the usage of LTE reduces due to lower channel quality. At the point Ⓓ, the data rate approaches the minimum level, and, therefore, the decrease in LTE transmit power slows down slightly to sustain the target bit-rate. When WiFi becomes more effective at the point Ⓔ, LTE is not sending anymore. Then, energy efficiency and total data rate increase due to WiFi radio link improvement up to the point Ⓕ, and our scheme arrives at the maximum level of the data rate, when further SNR growth does not lead to the data rate improvement.

Further, in order to compare our energy efficiency optimal power control with possible (simpler) alternatives, we continue by defining two primitive power control policies, where the transmit power is fixed. Specifically, we set the allocated power level equal to the maximum allowed power. Additionally, as a more intelligent power control strategy, below we introduce an intuitive heuristic transmission policy.

- *Simple policy* 1. The user device transmits on one channel with the maximum quality by allocating a fixed amount of power to it. We determine the best channel by choosing the highest value $\gamma_i w_i$, which takes into account both SNR and channel bandwidth.
- *Simple policy* 2. The user device transmits on all channels simultaneously by allocating a fixed amount of power to each channel.
- *Intuitive heuristic policy*. The user device follows a heuristic power allocation strategy by using firstly the best-quality channel and then, if necessary, leveraging the rest of the required bit-rate (up to the target value r_0) on other channels in the order of channel quality reduction (see Algorithm 2 for details). The available channels may be compared using the same criterion as above, i.e., the channel with higher $\gamma_i w_i$ value would be preferred.

In Fig. 12, the user device energy efficiency and the data rate achieved with the proposed optimal power control are compared against the performance of the three simpler power control strategies detailed above. The data rate with the first strategy is the lowest in the middle region, where the quality of both the channels is poor, leading to QoS violation when the target bit-rate cannot be satisfied even with the maximum transmit power. The second strategy generally provides higher data rates at the cost of excessive power consumption. Our energy efficiency optimal approach allows improving user data rate almost up to the level of the second strategy performance, while guaranteeing the target bit-rate r_0. By cutting down

Algorithm 2 Heuristic power allocation strategy

```
$r_{\text{rest}} \leftarrow r_0$
$I = \{i\}_{i=1}^{K}$
**while** $r_{\text{rest}} > 0$ or $K > 0$ **do**
    Find a channel with the best quality $i_0 \leftarrow \max_{i \in I} \gamma_i$
    **if** $r_{\text{rest}} > w_{i_0} \log\left(1 + p_{i_0}^{\max} \gamma_{i_0}\right)$ **then**
        $p_{i_0} \leftarrow p_{i_0}^{\max}$
        $r_{i_0} \leftarrow w_{i_0} \log\left(1 + p_{i_0}^{\max} \gamma_{i_0}\right)$
    **else**
        $r_{i_0} \leftarrow r_{\text{rest}}$
        $p_{i_0} \leftarrow 1/\gamma_{i_0} \left(e^{r_{i_0}/w_{i_0}} - 1\right)$
    **end if**
    $r_{\text{rest}} \leftarrow r_{\text{rest}} - r_{i_0}$
    $K \leftarrow K - 1$
    $I \leftarrow I \backslash \{i_0\}$
**end while**
```

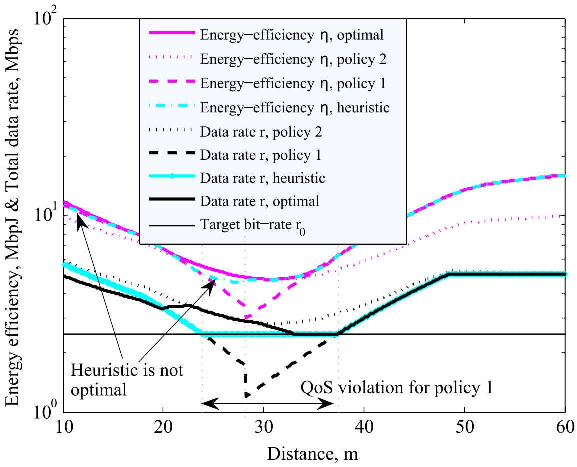


Fig. 12. Performance comparison of power control policies: optimal vs. simpler approaches.

on the unnecessary power consumption, our scheme clearly results in the maximum energy efficiency. Most interestingly, the third (intuitive heuristic) policy delivers close to the optimal energy efficiency. However, at certain distances, it fails to approach the energy efficient optimum.

Finally, in Fig. 13 we quantify the relative decrease in the energy efficiency when using one of the three alternative policies. We conclude that whenever the target bit-rate is supported, the proposed optimal scheme achieves much higher energy efficiency than the two primitive power allocation schemes. The third (heuristic) strategy approaches our optimal solution in many regions, but also leaves some room for improvement.

5. Conclusion

In this work, we have addressed energy efficient power control for a wireless deployment with multiple available radio access technologies. The problem of strict energy efficiency maximization at a mobile user device has been solved analytically for an arbitrary number of RATs and under several practical restrictions, such as minimum target bit-rate and maximum allowed transmit power. Our illustrative numerical examples for two and three RATs confirm that the proposed power control scheme reduces mobile device's power expenditure, while at the same time maintaining the required level of user data rate.

By contrast to the previous work, the use of our approach establishes support regions where two or more RATs work collaboratively to result in more energy efficient device operation when compared against simpler power control techniques. Our results suggest that the proposed power control strategy might become an attractive choice for the future integrated

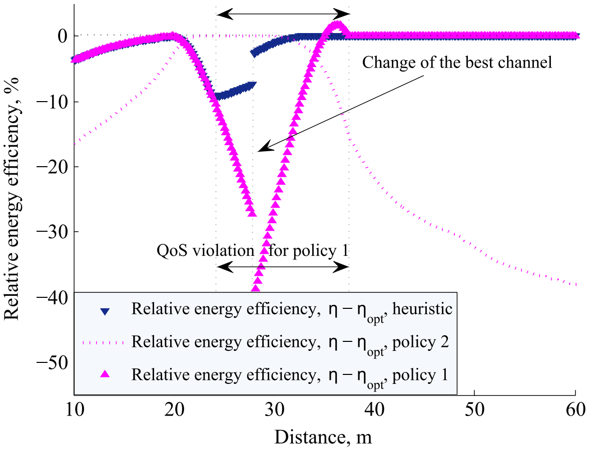


Fig. 13. Relative energy efficiency loss: optimal power control scheme vs. simpler approaches.

beyond-4G wireless systems and thus contribute to the related research. The choice of more adequate heuristic power allocation schemes that would achieve near-optimal performance at all times together with the characterization of dynamic traffic models better suited for multimedia mobile traffic are the directions of our current work.

Acknowledgments

This work is supported by GETA and the Internet of Things program of Digile, funded by Tekes. The work of the second and third authors is supported by the Academy of Finland. The authors are grateful to Alexander Pyattaev (Tampere University of Technology, Tampere, Finland) and Dr. Nageen Himayat (Intel Corp., Santa Clara, CA, USA) for insightful discussions as well as to the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of this paper.

Appendix A

In this appendix, we prove Theorem 1.

Proof. To establish the optimal solution for the optimization problem without constraints (EUP), we need to locate it among the stationary points satisfying the following condition:

$$\frac{\partial U(\mathbf{r})}{\partial r_j} = 0 \Leftrightarrow \frac{\frac{dp_j}{dr_j}\sum_{i=1}^{K} r_i - \left(\sum_{i=1}^{K} p_i + p_c\right)}{\left(\sum_{i=1}^{K} r_i\right)^2} = 0$$

$$\Leftrightarrow \frac{\frac{dp_j}{dr_j}\cdot c - \left(\sum_{i=1}^{K} p_i + p_c\right)}{r^2} = 0. \tag{A.1}$$

Further on, we demonstrate how to define the sought points. First, we calculate the derivatives for the individual power–rate mapping functions:

$$\frac{dp_j}{dr_j} = \frac{1}{\gamma_j w_j} e^{\frac{r_j}{w_j}}, \quad j = \overline{1, K}. \tag{A.2}$$

We substitute the expression (A.2) for the derivative of the function p_j into the condition for the stationary points (A.1):

$$\frac{\partial U(\mathbf{r})}{\partial r_j} = \frac{\frac{1}{\gamma_j w_j} e^{\frac{r_j}{w_j}} \sum_{i=1}^{K} r_i - \left[\sum_{i=1}^{K} \frac{1}{\gamma_i}\left(e^{\frac{r_i}{w_i}} - 1\right) + p_c\right]}{\left(\sum_{i=1}^{K} r_i\right)^2} = 0, \quad j = \overline{1, K}. \tag{A.3}$$

Hence, from (A.3) we obtain the sufficient condition for the stationary points:

$$\frac{1}{\gamma_j w_j} e^{\frac{r_j}{w_j}} \sum_{i=1}^{K} r_i - \sum_{i=1}^{K} \frac{1}{\gamma_i} e^{\frac{r_i}{w_i}} + \left[\sum_{i=1}^{K} \frac{1}{\gamma_i} - p_c \right] = 0, \quad j = \overline{1, K}.$$

Collecting the elements that depend and do not depend on j separately, we obtain for every $j = \overline{1, K}$:

$$\frac{1}{\gamma_j w_j} e^{\frac{r_j}{w_j}} = \frac{\sum_{i=1}^{K} \frac{1}{\gamma_i} e^{\frac{r_i}{w_i}} - \left[\sum_{i=1}^{K} \frac{1}{\gamma_i} - p_c \right]}{\sum_{i=1}^{K} r_i}, \quad j = \overline{1, K}. \tag{A.4}$$

We denote the right part of Eq. (A.4), which is constant with respect to the index j, as D:

$$\frac{\sum_{i=1}^{K} \frac{1}{\gamma_i} e^{\frac{r_i}{w_i}} - \left[\sum_{i=1}^{K} \frac{1}{\gamma_i} - p_c \right]}{\sum_{i=1}^{K} r_i} = D. \tag{A.5}$$

Therefore, for the left part, it holds the following:

$$\frac{1}{\gamma_j w_j} e^{\frac{r_j}{w_j}} = D, \quad j = \overline{1, K}. \tag{A.6}$$

Here, r_j is expressed via D as follows:

$$r_j = w_j \ln(w_j D) + w_j \ln \gamma_j, \quad j = \overline{1, K}. \tag{A.7}$$

We note that $\frac{1}{\gamma_j} e^{\frac{r_j}{w}} = w_j D,\ j = \overline{1, K}$ and then the substitution of (A.7) into (A.5) gives us:

$$D \sum_{j=1}^{K} w_j - \left[\sum_{j=1}^{K} \frac{1}{\gamma_j} - \sum_{j=1}^{K} p_j^c \right] = D \left(\sum_{j=1}^{K} w_j \ln(w_j D) + \sum_{j=1}^{K} w_j \ln \gamma_j \right).$$

Simplifying the above equation, we obtain the following:

$$D \sum_{j=1}^{K} w_j \ln(w_j D) + D \sum_{j=1}^{K} w_j \ln \gamma_j - \sum_{j=1}^{K} w_j D = - \left[\sum_{j=1}^{K} \frac{1}{\gamma_j} - \sum_{j=1}^{K} p_j^c \right].$$

Rearranging the above expression, we establish that:

$$D \ln \left\{ \prod_{j=1}^{K} w_j^{\frac{w_j}{\sum_{k=1}^{K} w_k}} \prod_{j=1}^{K} \gamma_j^{\frac{w_j}{\sum_{k=1}^{K} w_k}} \frac{D}{e} \right\} = \frac{- \sum_{j=1}^{K} \frac{1}{\gamma_j} + \sum_{j=1}^{K} p_j^c}{\sum_{j=1}^{K} w_j}. \tag{A.8}$$

We also introduce a new variable:

$$X = \ln \left[\prod_{j=1}^{K} w_j^{\frac{w_j}{\sum_{k=1}^{K} w_k}} \prod_{j=1}^{K} \gamma_j^{\frac{w_j}{\sum_{k=1}^{K} w_k}} \frac{D}{e} \right].$$

Then, for the value of D, it holds:

$$D = e^{(X+1)} \prod_{j=1}^{K} w_j^{\frac{-w_i}{\sum_{k=1}^{K} w_k}} \prod_{j=1}^{K} \gamma_j^{\frac{-w_j}{\sum_{k=1}^{K} w_k}}.$$

By changing the variables, we rewrite Eq. (A.8) as:

$$Xe^X = \frac{-\sum\limits_{j=1}^{K}\frac{1}{\gamma_j}+\sum\limits_{j=1}^{K}p_j^c}{e\sum\limits_{j=1}^{K}w_j}\prod_{i=1}^{K} w_j^{\frac{w_j}{\sum_{k=1}^{K} w_k}} \prod_{j=1}^{K}\gamma_j^{\frac{w_j}{\sum_{k=1}^{K} w_k}}. \tag{A.9}$$

From Eq. (A.9), we can obtain the value of X and, accordingly, the expression for D:

$$X = W\left(\frac{-\sum\limits_{j=1}^{K}\frac{1}{\gamma_j}+\sum\limits_{j=1}^{K}p_j^c}{e\sum\limits_{j=1}^{K}w_j}\prod_{j=1}^{K} w_j^{\frac{w_j}{\sum_{k=1}^{K} w_k}} \prod_{j=1}^{K}\gamma_i^{\frac{w_j}{\sum_{k=1}^{K} w_k}}\right),$$

$$D = e\left(\prod_{j=1}^{K} w_j^{\frac{w_j}{\sum_{k=1}^{K} w_k}} \prod_{j=1}^{K}\gamma_i^{\frac{w_i}{\sum_{k=1}^{K} w_k}}\right)^{-1} \times \exp\left\{W\left(\frac{-\sum\limits_{j=1}^{K}\frac{1}{\gamma_j}+\sum\limits_{j=1}^{K}p_j^c}{e\sum\limits_{j=1}^{K}w_j}\prod_{j=1}^{K} w_j^{\frac{w_j}{\sum_{k=1}^{K} w_k}} \prod_{j=1}^{K}\gamma_j^{\frac{w_j}{\sum_{k=1}^{K} w_k}}\right)\right\}, \tag{A.10}$$

where $W(x)$ is the Lambert's W-function [39] representing the solution to the following equation:

$$Xe^X = \alpha. \tag{A.11}$$

We note that the transcendental equation (A.11) has exactly one real root $x = W(\alpha)$ in the domain $\{-\frac{1}{e}\} \cup [0, \infty)$, and also has two real roots when $\alpha \in (-\frac{1}{e}, 0)$ (it follows from the shape of the Lambert's function, which has several branches, see Fig. 3). Here, we consider the value from the main (upper) branch, which is explained by Proposition 1. In case when $\alpha < -\frac{1}{e}$, we conclude that there is no optimal solution in R^K for our optimization problem. If a real value $W(\alpha)$ exists, we substitute the expression for D (A.10) into (A.7):

$$\begin{aligned} r_i &= w_i\ln(w_iD)+w_i\ln\gamma_i = w_i\ln\gamma_i + w_i\ln w_i + w_i\ln D \\ &= w_i\ln\gamma_i + w_i\ln w_i + w_i\left\{1-\sum_{j=1}^{K}\frac{w_j}{\sum\limits_{k=1}^{K}w_k}\ln\left(w_j\gamma_j\right)\right\}\times W\left(\frac{-\sum\limits_{j=1}^{K}\frac{1}{\gamma_j}+\sum\limits_{j=1}^{K}p_j^c}{e\sum\limits_{k=1}^{K}w_k}\prod_{j=1}^{K} w_j^{\frac{w_j}{\sum_{k=1}^{K} w_k}} \prod_{j=1}^{K}\gamma_i^{\frac{w_j}{\sum_{k=1}^{K} w_k}}\right). \end{aligned}$$

The final expression for the achievable data rate r_i is given as:

$$r_i = w_i\ln\left(\gamma_i w_i\right) + w_i\left\{1-\frac{\sum\limits_{j=1}^{K}w_j\ln\left(w_j\gamma_j\right)}{\sum\limits_{k=1}^{K}w_k}\right\} + w_iW\left(\frac{-\sum\limits_{j=1}^{K}\frac{1}{\gamma_j}+\sum\limits_{j=1}^{K}p_j^c}{e\sum\limits_{k=1}^{K}w_k}\prod_{j=1}^{K}\left(w_j\gamma_j\right)^{\frac{w_j}{\sum_{k=1}^{K} w_k}}\right). \tag{A.12}$$

Here, we have obtained the stationary point $\mathbf{r} = \{r_i\}_{i=1}^K$ for our optimization problem. The optimal power level to operate on a single channel can be indicated by formula (1), and the final formula for $p_i = w_iD - \frac{1}{\gamma_i}$ is:

$$p_i = -\frac{1}{\gamma_i} + w_ie\prod_{j=1}^{K}\left(w_j\gamma_j\right)^{\frac{-w_j}{\sum_{k=1}^{K} w_k}} \times \exp\left\{W\left(\frac{-\sum\limits_{j=1}^{K}\frac{1}{\gamma_j}+\sum\limits_{j=1}^{K}p_j^c}{e\sum\limits_{k=1}^{K}w_k}\prod_{j=1}^{K}\left(w_j\gamma_j\right)^{\frac{w_j}{\sum_{k=1}^{K} w_k}}\right)\right\}, \tag{A.13}$$

which completes the proof. □

Appendix B

Solving the constrained problem (ECP), if we establish that the optimal solution does not belong to the required domain, we proceed by considering all the possible options for border location of the maximum of energy efficiency. First, we consider the case when $r_i > r_i^{\max}$. This means that the ith component should be set on the border, i.e., $r_i = r_i^{\max}$. If several components do not satisfy the inequalities (4), we select the component with the least contribution to the energy efficiency growth, i.e., find:

$$i = \arg\min_{j=\overline{1,K}} \left\{ \frac{dp_j(r_j)}{dr_j} \middle| r_j > r_i^{\max} \right\}.$$

Then, we set $r_i = r_i^{\max}$ and thus decrease the dimension of the optimization task to $K-1$, considering the following:

$$\min_{\{r_j\}_{j=1,j\neq i}^{K}} U(\mathbf{r}) = \min_{\{r_j\}_{j=1,j\neq i}^{K}} \frac{\sum\limits_{j=1,j\neq i}^{K} p_j(r_j) + p_i(r_i^{\max}) + \sum\limits_{j=1}^{K} p_j^r}{\sum\limits_{j=1,j\neq i}^{K} r_j + r_i^{\max}},$$

subject to the same conditions. We rearrange the order of variables r_i without loss of generality, so that $r_i^{\max}$ becomes the last in a series under the index K, when $r_K = r_i^{\max}$.

Below we suggest the proof of Theorem 2, which delivers an optimal vector $\mathbf{p}$, if for the solution $r^* \in R^K$ of ECP there is at least one index i, such that $r_i > r_i^{\max}$.

Proof. Letting $p_i(r_K^{\max}) = p_K$ and $r_K^{\max} = r_K$, we find that for any $j = \overline{1,K}$, $i \neq j$:

$$\frac{\partial U(\mathbf{r})}{\partial r_j} = \frac{\frac{1}{\gamma_j w_j} e^{\frac{r_j}{w_j}} \sum\limits_{j=1}^{K} r_j - \left[\sum\limits_{j=1}^{K} \frac{1}{\gamma_j}\left(e^{\frac{r_j}{w_j}} - 1\right) + \sum\limits_{j=1}^{K} p_j^c\right]}{\sum\limits_{j=1}^{K-1} r_j + r_K}.$$

Thus, we obtain the necessary condition of the stationary points:

$$\frac{1}{\gamma_j w_j} e^{\frac{r_j}{w_j}} \sum_{j=1}^{K} r_j - \sum_{j=1}^{K-1} \frac{1}{\gamma_j} e^{\frac{r_j}{w_j}} + \left[\sum_{j=1}^{K} \frac{1}{\gamma_j} - \sum_{j=1}^{K} p_j^c - p_K\right] = 0,$$

which holds for all $i = \overline{1, K-1}$. Following the same logic as before, we can obtain the expression for the index $j = \overline{1, K-1}$:

$$\frac{1}{\gamma_j w_j} e^{\frac{r_j}{w_j}} = \frac{\sum\limits_{j=1}^{K-1} \frac{1}{\gamma_j} e^{\frac{r_j}{w_j}} + \frac{1}{\gamma_K} e^{\frac{r_K}{w_K}} - \left[\sum\limits_{j=1}^{K} \frac{1}{\gamma_j} - \sum\limits_{j=1}^{K} p_j^c\right]}{\sum\limits_{j=1}^{K-1} r_j + r_K} = D.$$

Letting, for the sake of tractability, that:

$$\alpha = \frac{\frac{1}{\gamma_K} e^{\frac{r_K}{w_K}} - \sum\limits_{i=1}^{K} \frac{1}{\gamma_i} + p_c}{e \sum\limits_{i=1}^{K-1} w_i} \prod_{i=1}^{K-1} w_i^{\frac{w_i}{\sum_{j=1}^{K-1} w_j}} \prod_{i=1}^{K-1} \gamma_i^{\frac{w_i}{\sum_{j=1}^{K-1} w_j}} \cdot e^{\frac{r_K}{w_K}}, \tag{B.1}$$

we derive the expression for D:

$$D = e \left(\prod_{i=1}^{K-1} w_i^{\frac{w_i}{\sum_{j=1}^{K-1} w_j}} \prod_{i=1}^{K-1} \gamma_i^{\frac{w_i}{\sum_{j=1}^{K-1} w_j}} \cdot e^{\frac{r_K}{w_K}} \right)^{-1} \cdot e^{W(\alpha)}.$$

Substituting the above into the expression for the achievable data rate (A.7), delivers us:

$$\begin{aligned} r_i &= w_i \ln(w_i D) + w_i \ln \gamma_i = w_i \ln \gamma_i + w_i \ln w_i + w_i \ln D \\ &= w_i \ln \gamma_i + w_i \ln w_i + w_i \left\{ 1 - \frac{r_K}{w_K} - \sum_{k=1}^{K-1} \frac{w_k}{\sum\limits_{j=1}^{K-1} w_j} \ln(w_k \gamma_k) \right\} \cdot W(\alpha). \end{aligned}$$

The final expressions for the achievable data rate r_i and transmit power p_i for any $i = \overline{1, K-1}$ are given as follows:

$$r_i = w_i \ln(\gamma_i w_i) + w_i \left\{ 1 - \frac{r_K}{w_K} - \frac{\sum\limits_{j=1}^{K} w_j \ln(w_j \gamma_j)}{\sum\limits_{j=1}^{K} w_j} \right\} W(\alpha), \tag{B.2}$$

$$p_i = e^{1-\frac{r_K}{w_K}} \prod_{k=1}^{K-1} (w_k \gamma_k)^{\frac{-w_k}{\sum_{j=1}^{K-1} w_j}} \cdot e^{W(\alpha)} - \frac{1}{\gamma_i}, \quad i = \overline{1, K-1}, \tag{B.3}$$

which completes the proof. □

Appendix C

Now we consider the case $\sum_{i=1}^{K} r_i - r_0 < 0$. We established that the optimal solution is located on the plane $\sum_{j=1}^{K} r_j = r_0$. Therefore, the following constrained problem should be considered:

$$\min_{\{r_j\}_{j=1}^{K}} U(\mathbf{r}) = \min_{\{r_j\}_{j=1}^{K}} \frac{\sum\limits_{j=1}^{K} p_j + \sum\limits_{j=1}^{K} p_j^r}{\sum\limits_{j=1}^{K} r_j},$$

subject to

$$\sum_{j=1}^{K} r_j = r_0. \tag{C.1}$$

Theorem 3 asserts the expression for optimal power vector $\mathbf{p}$ and vector $\mathbf{r}$ if ECP has a solution $\mathbf{r}^* \in R^K$ and $\sum_{i=1}^{K} r_i^* - r_0 < 0$. Below we provide the proof of Theorem 3.

Proof. We express r_K from Eq. (C.1) as follows:

$$r_K = r_0 - \sum_{j=1}^{K-1} r_j. \tag{C.2}$$

This allows decreasing the dimension of the optimization task. Our system thus transforms to:

$$\min_{\{r_j\}_{j=1}^{K-1}} U(\mathbf{r}) = \min_{\{r_j\}_{j=1}^{K-1}} \frac{\sum\limits_{j=1}^{K-1} \frac{1}{\gamma_i} e^{\frac{r_i}{w_i}} + \frac{1}{\gamma_K} e^{\frac{r_0 - \sum_{j=1}^{K-1} r_j}{w_K}} + p_c}{r_0},$$

which is equivalent to the following problem:

$$\min_{\{r_j\}_{j=1}^{K-1}} U_1(\mathbf{r}) = \min_{\{r_j\}_{j=1}^{K-1}} \left\{ \sum_{j=1}^{K-1} \frac{1}{\gamma_i} e^{\frac{r_i}{w_i}} + \frac{1}{\gamma_K} e^{\frac{r_0 - \sum_{j=1}^{K-1} r_j}{w_K}} \right\}. \tag{C.3}$$

The stationary point's condition for the problem (C.3) is:

$$\frac{\partial U_1(\mathbf{r})}{\partial r_i} = \frac{1}{\gamma_i w_i} e^{\frac{r_i}{w_i}} - \frac{1}{\gamma_K w_K} e^{\frac{r_0 - \sum_{j=1}^{K-1} r_j}{w_K}} = 0, \quad i = \overline{1, K-1}.$$

Further, we follow the same logic as before, denoting the constant part as D:

$$\frac{1}{\gamma_i w_i} e^{\frac{r_i}{w_i}} = \frac{1}{\gamma_K w_K} e^{\frac{r_0 - \sum_{j=1}^{K-1} r_j}{w_K}} = D, \quad i = \overline{1, K-1}.$$

Then, we express the individual data rates r_i via D:

$$r_i = w_i \ln(D\gamma_i w_i) = w_i \ln(D) + w_i \ln(\gamma_i w_i), \quad i = \overline{1, K-1},$$

and substitute them all into the expression for r_K (C.2):

$$r_0 - \sum_{j=1}^{K-1} w_j \ln(D) - \sum_{j=1}^{K-1} w_j \ln(\gamma_j w_j) = w_K \ln(D\gamma_K w_K).$$

Simplifying the expression above, we obtain:

$$r_0 - \sum_{j=1}^{K} w_j \ln(\gamma_j w_j) = \sum_{j=1}^{K} w_j \ln(D).$$

Next, we express $\ln D$ from the above equation:

$$\ln(D) = \frac{r_0 - \sum\limits_{j=1}^{K} w_j \ln(\gamma_j w_j)}{\sum\limits_{j=1}^{K} w_j}.$$

Finally, we derive an expression for the individual data rate r_i:

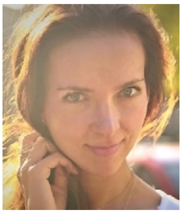
$$r_i = w_i \frac{r_0 - \sum\limits_{j=1}^{K} w_j \ln(\gamma_j w_j)}{\sum\limits_{j=1}^{K} w_j} + w_i \ln(\gamma_i w_i), \quad i = \overline{1, K-1},$$

which completes the proof. □

References

[1] 3GPP LTE Release 10 & beyond (LTE-Advanced), 2011.
[2] IEEE 802.16m-2011, Amendment to IEEE Standard for Local and Metropolitan area networks, Advanced Air Interface, 2011.
[3] B. Bjerke, LTE-Advanced and the evolution of LTE deployments, IEEE Wirel. Commun. 18 (2011) 4–5.
[4] S.-P. Yeh, S. Talwar, G. Wu, N. Himayat, K. Johnsson, Capacity and coverage enhancement in heterogeneous networks, IEEE Wirel. Commun. 18 (2011) 32–38.
[5] J. Andrews, Seven ways that HetNets are a cellular paradigm shift, IEEE Commun. Mag. 51 (2013) 136–144.
[6] K. Pentikousis, In search of energy-efficient mobile networking, IEEE Commun. Mag. 48 (2010) 95–103.
[7] I. Akyildiz, D. Gutierrez-Estevez, E. Reyes, The evolution to 4G cellular systems: LTE-Advanced, Phys. Commun. 3 (2010) 217–244.
[8] J. Zhu, A. Waltho, X. Yang, X. Guo, Multi-radio coexistence: challenges and opportunities, in: Proc. of the 16th International Conference on Computer Communications and Networks, ICCCN, pp. 358–364.
[9] White Paper: Power Consumption & Energy Efficiency, Atheros Communications, 2003.
[10] G. Miao, N. Himayat, G. Li, Energy-efficient link adaptation in frequency-selective channels, IEEE Trans. Commun. 58 (2010) 545–554.
[11] D. Raychaudhuri, N. Mandayam, Frontiers of wireless and mobile communications, Proc. IEEE 100 (2012) 824–840.
[12] G. Song, G. Li, Asymptotic throughput analysis for channel-aware scheduling, IEEE Trans. Commun. 54 (2006) 1827–1834.
[13] X. Lin, N. Shroff, R. Srikant, A tutorial on cross-layer optimization in wireless networks, IEEE J. Sel. Areas Commun. 24 (2006) 1452–1463.
[14] L. Benini, A. Bogliolo, G. de Micheli, A survey of design techniques for system-level dynamic power management, IEEE Trans. Very Large Scale Integr. (VLSI) Syst. 8 (2000) 299–316.
[15] C. Schurgers, Energy-aware wireless communications (Ph.D. thesis), University of California Los Angeles, 2002.
[16] V. Rodoplu, T. Meng, Bits-per-Joule capacity of energy-limited wireless networks, IEEE Trans. Wireless Commun. 6 (2007) 857–865.
[17] G. Miao, N. Himayat, G. Li, A. Swami, Cross-layer optimization for energy-efficient wireless communications: a survey, J. Wireless Commun. Mob. Comput. 9 (2009) 529–542.
[18] G. Song, G. Li, Cross-layer optimization for OFDM wireless networks—part I: theoretical framework, IEEE Trans. Wireless Commun. 4 (2005) 614–624.
[19] G. Song, G. Li, Cross-layer optimization for OFDM wireless networks—part II: algorithm development, IEEE Trans. Wireless Commun. 4 (2005) 625–634.
[20] G. Miao, N. Himayat, G. Li, A. Koc, S. Talwar, Interference-aware energy-efficient power optimization, in: Proc. of the IEEE International Conference on Communications, ICC.
[21] S. Verdu, Spectral efficiency in the wideband regime, IEEE Trans. Inform. Theory 48 (2002) 1319–1343.
[22] F. Meshkati, H. Poor, S. Schwartz, N. Mandayam, An energy-efficient approach to power control and receiver design in wireless networks, IEEE Trans. Commun. 53 (2005) 1885–1894.
[23] A. Wang, S. Cho, C. Sodini, A. Chandrakasan, Energy efficient modulation and MAC for asymmetric RF microsensor system, in: Proc. of the International Symposium on Low Power Electronics and Design, pp. 106–111.
[24] S. Cui, A. Goldsmith, A. Bahai, Energy-constrained modulation optimization, IEEE Trans. Wireless Commun. 4 (2005) 2349–2360.
[25] H. Kim, Exploring tradeoffs in wireless networks under flow-level traffic: energy, capacity and QoS (Ph.D. Thesis), University of Texas at Austin, 2009.
[26] P. Grover, K. Woyach, A. Sahai, Towards a communication-theoretic understanding of system-level power consumption, IEEE J. Sel. Areas Commun. 29 (2011) 1744–1755.
[27] D. Cavalcanti, D. Agrawal, C. Cordeiro, B. Xie, A. Kumar, Issues in integrating cellular networks WLANs, and MANETs: a futuristic heterogeneous wireless network, IEEE Wirel. Commun. 12 (2005) 30–41.
[28] B. Walke, S. Mangold, L. Berlemann, IEEE 802 Wireless Systems: Protocols, Multi-Hop Mesh/Relaying, Performance and Spectrum Coexistence, Wiley, 2007.

[29] T. Zetterman, A. Piipponen, K. Raiskila, S. Slotte, Multi-radio coexistence and collaboration on an SDR platform, Analog Integr. Circuits Signal Process. 69 (2011) 329–339.
[30] L. Wang, G. Kuo, Mathematical modeling for network selection in heterogeneous wireless networks—a tutorial, IEEE Commun. Surv. Tutor. 15 (2013) 271–292.
[31] K. Andersson, C. Ahlund, Optimized access network selection in a combined WLAN/LTE environment, Wirel. Pers. Commun. 61 (2011) 739–751.
[32] R. Amin, J. Martin, J. Deaton, L. DaSilva, A. Hussien, A. Eltawil, Balancing spectral efficiency, energy consumption, and fairness in future heterogeneous wireless systems with reconfigurable devices, IEEE J. Sel. Areas Commun. 31 (2013) 969–980.
[33] E. Aryafar, A. Keshavarz-Haddad, M. Wang, M. Chiang, RAT selection games in HetNets, in: Proc. of the IEEE INFOCOM.
[34] 3GPP TR 37.834. Study on WLAN/3GPP Radio Interworking, 2013.
[35] M. Bennis, M. Simsek, W. Saad, S. Valentin, M. Debbah, When cellular meets WiFi in wireless small cell networks, IEEE Commun. Mag. 51 (2013) 44–50.
[36] V. Zafeiris, E. Giakoumakis, Optimized traffic flow assignment in multi-homed, multi-radio mobile hosts, Comput. Netw. 55 (2011) 1114–1131.
[37] S. Andreev, P. Gonchukov, N. Himayat, Y. Koucheryavy, A. Turlikov, Energy efficient communications for future broadband cellular networks, Comput. Commun. 35 (2012) 1662–1671.
[38] W. Zangwill, Nonlinear Programming: A Unified Approach, Prentice-Hall, Englewood Cliffs, NJ, 1969.
[39] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, On Lambert's W function, Technical Report CS-93-03, University of Waterloo, 1993.
[40] 3GPP TR 36.814. Further advancements for E-UTRA physical layer aspects, 2010.
[41] S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, Y. Koucheryavy, Cellular traffic offloading onto network-assisted device-to-device connections, IEEE Commun. Mag. 52 (2014) 20–31.



Olga Galinina is a Ph.D. candidate in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. She received her B.Sc. and M.Sc. degrees in Applied Mathematics from the Department of Applied Mathematics, Faculty of Mechanics and Physics, Saint-Petersburg State Polytechnical University, Russia. She has publications on mathematical problems in the novel telecommunication protocols in internationally recognized journals and high-level peer-reviewed conferences. Her research interests include applied mathematics and statistics, queueing theory and its applications; wireless networking and energy efficient systems, machine-to-machine and device-to-device communication.

Sergey Andreev is a Senior Research Scientist in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. He received the Specialist degree (2006) in Information Security and the Cand.Sc. degree (2009) in Wireless Communications both from St. Petersburg State University of Aerospace Instrumentation, St. Petersburg, Russia, as well as the Ph.D. degree (2012) in Technology from Tampere University of Technology, Tampere, Finland. He has (co-)authored more than 80 published research works. His research interests include wireless communications, energy efficiency, heterogeneous networking, cooperative communications, and machine-to-machine applications. More information is available at: http://www.cs.tut.fi/~andreev.

Andrey Turlikov is a Full Professor and the Head of the Department of Information and Communication Systems at St. Petersburg State University of Aerospace Instrumentation (SUAI). Over the 30 years of experience in the area of telecommunications, he has supervised more than 120 diploma projects and M.Sc. theses in total, as well as 8 successful Ph.D. theses. Recently, he has participated in and supervised over 10 successful long-term research projects, focusing on the next-generation telecommunications technologies (3GPP LTE-Advanced, IEEE 802.16m, prominent IEEE 802.11-based solutions). His primary targets are energy efficiency improvement, collaborative techniques, heterogeneous networking, and support for advanced services towards standardization in IEEE, as well as energy-efficient schemes for Long Term HSPA Evolution (LTHE). More information is available at: http://andrey-turlikov.narod.ru/turlikov_eng.htm.

Yevgeni Koucheryavy is a Full Professor and Lab Director at the Department of Electronics and Communications Engineering at the Tampere University of Technology (TUT), Finland. He received his Ph.D. degree (2004) from the TUT. He is the author of numerous publications in the field of advanced wired and wireless networking and communications. His current research interests include various aspects in heterogeneous wireless communication networks and systems, the Internet of Things and its standardization, and nanocommunications. He is an Associate Technical Editor of IEEE Communications Magazine and Editor of IEEE Communications Surveys and Tutorials. More information is available at: http://www.cs.tut.fi/~yk.

Publication 7

© 2014 IEEE. Reprinted, with permission, from

S. Andreev, M. Gerasimenko, O. Galinina, Y. Koucheryavy, N. Himayat, S. Yeh, and S. Talwar, "Intelligent Access Network Selection in Converged Multi-Radio Heterogeneous Networks", *IEEE Wireless Communications*, vol. 21, pp. 86-96, 2014.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Tampere University of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to `http://www.ieee.org/publications_standards/publications/rights/rights_link.html` to learn how to obtain a License from RightsLink.

Intelligent Access Network Selection in Converged Multi-Radio Heterogeneous Networks

Sergey Andreev[†], Mikhail Gerasimenko, Olga Galinina, Yevgeni Koucheryavy,
Nageen Himayat, Shu-ping Yeh, and Shilpa Talwar

Abstract—**Heterogeneous multi-radio networks are emerging network architectures, which comprise hierarchical deployments of increasingly smaller cells. In these deployments, each user device may employ multiple radio access technologies to communicate with network infrastructure. With the growing numbers of such multi-radio consumer devices, mobile network operators seek to leverage spectrum across diverse radio technologies thus boosting capacity and enhancing quality of service. In this article, we review major challenges in delivering uniform connectivity and service experience to converged multi-radio heterogeneous deployments. We envision that multiple radios and associated device/infrastructure intelligence for their efficient use will become a fundamental characteristic of future 5G technologies, where the distributed unlicensed-band network (e.g., WiFi) may take advantage of the centralized control function residing in the cellular network (e.g., 3GPP LTE). Illustrating several available architectural choices for integrating WiFi and LTE networks, we specifically focus on interworking within the *radio access network* and detail feasible options for intelligent access network selection. Both network- and user-centric approaches are considered, wherein the control rests with the network or the user. In particular, our system-level simulation results indicate that load-aware user-centric schemes, which augment SNR measurements with additional information about network loading, could improve the performance of conventional WiFi-preferred solutions based on minimum SNR threshold. Comparison with more advanced network-controlled schemes has also been completed to confirm attractive practical benefits of distributed user-centric algorithms. Building on extensive system-wide simulation data, we also propose novel analytical space-time methodology for assisted network selection capturing user traffic dynamics together with spatial randomness of multi-radio heterogeneous networks.**

Index Terms—**Heterogeneous networks, multiple radio access technologies, LTE/WiFi integration, intelligent access network selection, load-awareness.**

I. Recent Advances in Multi-Radio Networking

The rapid expansion of wireless communications over the last decades has introduced fundamental changes to "anytime, anywhere" mobile Internet access, as well as posed new challenges for the research community. In 2011, the fourth generation of broadband communication standards has been completed to offer aggressive improvements in all aspects of wireless system design, including system capacity, energy efficiency, and user quality of service (QoS). As the respective technologies are being deployed today, the focus of recent research efforts is shifting to what may be referred to as fifth generation (5G) wireless networks.

Given a historical 10-year cycle for every existing generation, it is expected that 5G systems will be deployed sometime around year 2020. Whereas there is currently no complete technical definition of what comes after the state-of-the-art networking technology, the anticipated communication requirements may already be understood from the user perspective. Regardless of their current location, human users would like to be connected at all times taking advantage of the rich set of services provided by the contemporary multimedia-over-wireless networks. This creates significant challenges for 5G technology design, as user's connectivity experience should match data rate requirements and be uniform no matter where the user is, who the user connects to, and what the user service needs are [1].

Unfortunately, contemporary wireless networks are currently unable to deliver the desired ubiquitous connectivity experience. In the first place, they are lacking uniformity in the data rates, suffer from excessive time delays, or sometimes even service outage due to poor coverage and severe interference conditions. Whereas current technologies have indeed been helpful to cope with some of those challenges, it is commonly believed that they will still be insufficient to meet the anticipated growth in traffic demand (nearly 11-fold over the following 5 years [2]) aggravated by rapid proliferation in types and numbers of wireless devices. To make matters worse, billions of diverse machine-type devices connect to the network thus reshaping the Internet as we know it today. All these technological challenges accentuate the need to explore novel solutions within the context of 5G networks.

A. Major trends behind 5G technology

A transformation of mobile user experience requires revolutionary changes in both network infrastructure and device architecture, where the user equipment (UE) is jointly optimized with the surrounding network context [3]. Many believe that the only feasible solution to mitigate the increasing disproportion between the desired QoS and the limited wireless resources is by deploying the higher density of femto- and pico-cells in current cellular architecture. Owing to shorter radio links, smaller cells provide higher data rates and require less energy for uplink transmission, especially in urban environments.

S. Andreev, M. Gerasimenko, O. Galinina, and Y. Koucheryavy are with the Department of Electronics and Communications Engineering, Tampere University of Technology, FI-33720 Tampere, Finland.

N. Himayat, S.-p. Yeh, and S. Talwar are with Intel Corporation, Santa Clara, CA, USA.

[†]S. Andreev is the contact author: P.O. Box 553, FI-33101 Tampere, Finland; e-mail: sergey.andreev@tut.fi

Manuscript received March 01, 2014; revised June 14, 2014.

However, introducing an increasing number of serving stations to bridge the capacity gap incurs extra complexity due to more cumbersome interference management, higher rental fees, and increased infrastructure maintenance costs [4]. More importantly, licensed spectrum continues to be scarce and expensive, whereas the traditional methods to improve its efficient use approach their theoretical limits. Even when additional spectrum is allocated, these new frequencies are likely to remain fragmented and could require diverse transmission technologies. Consequently, there is a pressing demand to leverage additional capacity across multiple radio access technologies (RATs).

As the result, it becomes crucial to aggregate different radio technologies as part of a common *converged* radio network, in a manner transparent to the end user, and develop techniques that can efficiently utilize the radio resources available across different spectral bands potentially using various RATs [5]. In particular, we expect that the majority of immediate gains will come from advanced architectures and protocols that would leverage the unlicensed spectrum. For example, mobile users with direct device-to-device communication capability may take advantage of their unlicensed-band radios and cooperate with other proximate users to locally improve access in a cost-efficient way [6].

Further, as cell sizes shrink, the footprints of cellular, local, and personal area networks are increasingly overlapping. This creates an attractive opportunity to *simultaneously* utilize multiple RATs for improved wireless connectivity. We thus believe that intelligent multi-RAT coupling will efficiently leverage performance benefits across several dimensions of diversity, including spatial, temporal, frequency, interference, load, and others. In future 5G networks, both short- and long-range technologies may need to work cooperatively and exploit the intricate interactions between the device and the network, as well as between the devices themselves, to realize the desired uniform user experience [7].

Consequently, the incentive to efficiently coordinate between the alternative RATs is growing stronger and we envision that multiple radios together with the associated device/system intelligence for their efficient use will become a fundamental characteristic of next-generation networks [8]. More specifically, the distributed unlicensed-band network (e.g., Wireless Local Area Network, WLAN) may take advantage of the *centralized* control function residing in the cellular network to effectively perform dynamic multi-RAT network association and hence provide beyond-additive gains in network capacity and user connectivity experience.

B. Scope and core novelty of current research

According to the above, there is currently an increasing shift towards tighter interworking between different RATs. To this end, our research campaign is targeting joint RAT assignment, selection, and scheduling algorithms, which provide significant improvement in overall system performance. In what follows, our focus is set on integration between multiple RATs within heterogeneous network architecture. As our case study, we consider convergence of WLAN-based small cells with operator-managed cellular deployment to illustrate feasible architectural options for integration and their associated performance benefits. Consequently, we seek to explore the potential of a diverse range of devices requiring connectivity at different scales to augment system capacity and improve user connectivity experience.

We emphasize that interworking between WLAN and cellular networks has already been considered in the past, but largely from the perspective of inter-network (vertical) handoff [9]. Cellular standards community, represented by the Third Generation Partnership Project (3GPP), has also been involved in developing specifications that address cellular/WLAN interworking for a number of years. Several new study and work items have recently emerged to develop specifications towards tighter integration of WLAN with cellular networks. The areas of investigation range from solutions for trusted access to 3GPP services with WLAN devices, seamless mobility between 3GPP and WLAN technologies, and support for Access Network Discovery and Selection Function (ANDSF). While much of this effort has focused on loose interworking solutions only requiring changes within the core network, there has been a recent shift in 3GPP Release 12 to address interworking within the Radio Access Network (RAN) [10].

This emerging trend is driven by the need to support better QoS on unlicensed spectrum as demanded by a consortium of network operators who have introduced stringent requirements for carrier-grade WiFi. The WLAN community has also responded with new initiatives such as Hot Spot 2.0, as well as a novel "High Efficiency WLAN" standardization effort by IEEE 802.11 working group. Hence, it is timely to investigate RAN-based integration solutions, which assume increased cooperation between 3GPP Long Term Evolution (LTE) and WiFi radio technologies. Along these lines, our work details several intelligent network selection mechanisms, which deliver significant gains in overall system performance and user QoS. We address both network-centric and user-centric approaches, wherein the control of how different radio technologies are utilized rests with the network or the user respectively.

II. Enabling Architectures and Algorithms for Converged Heterogeneous Networks

As argued previously, the capacity and connectivity limitations faced by the future 5G networks will continue to drive the need for closer integration across different RATs. To this end, Fig. 1(a) illustrates our vision of an operator's multi-RAT heterogeneous network (HetNet). It features a hierarchical deployment of wide-area macro cells for ubiquitous coverage, connectivity, and seamless mobility augmented with the overlay tier of inexpensive low-power smaller cells (picos, femtos, WiFi access points, integrated WiFi-LTE modules, etc.) to enhance capacity by moving infrastructure closer to the users in areas with higher traffic demand.

Whereas the trend towards the use of WLAN in conjunction with cellular networks has emerged from the near-term need of operators to relieve congestion on cellular networks, the use of WiFi is expected to remain an integral part of operators'

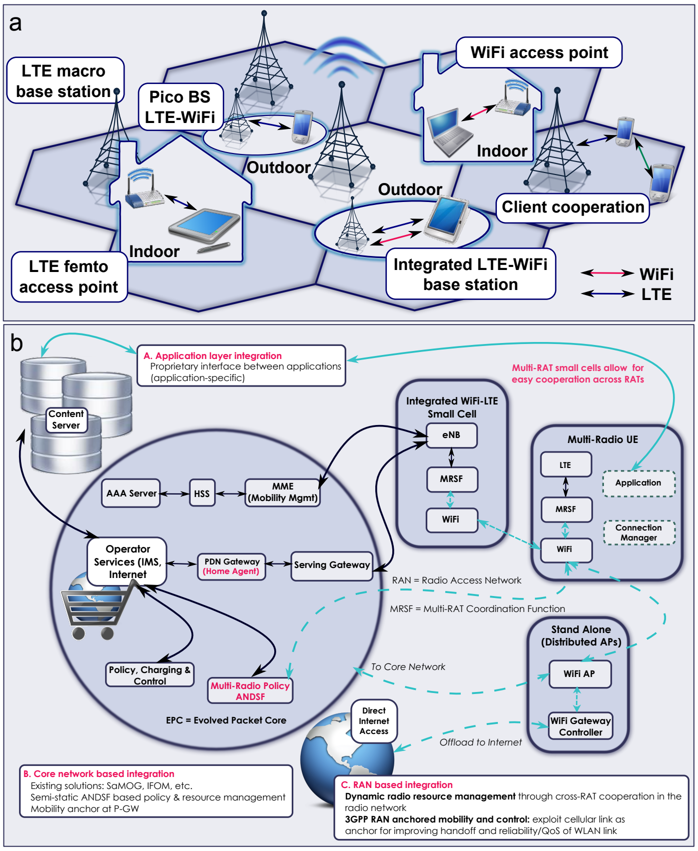

a
LTE macro base station
Pico BS LTE-WiFi
Outdoor
WiFi access point
Indoor
Outdoor
Client cooperation
Indoor
LTE femto access point
Integrated LTE-WiFi base station
WiFi
LTE
b
A. Application layer integration
Proprietary interface between applications (application-specific)
Content Server
Multi-RAT small cells allow for easy cooperation across RATs
Integrated WiFi-LTE Small Cell
eNB
MRSF
WiFi
Multi-Radio UE
LTE
Application
MRSF
Connection Manager
WiFi
AAA Server
HSS
MME (Mobility Mgmt)
Operator Services (IMS, Internet
PDN Gateway (Home Agent)
Serving Gateway
RAN = Radio Access Network
MRSF = Multi-RAT Coordination Function
Stand Alone (Distributed APs)
WiFi AP
WiFi Gateway Controller
To Core Network
Policy, Charging & Control
Multi-Radio Policy ANDSF
EPC = Evolved Packet Core
Direct Internet Access
Offload to Internet
B. Core network based integration
Existing solutions: SaMOG, IFOM, etc.
Semi-static ANDSF based policy & resource management
Mobility anchor at P-GW
C. RAN based integration
Dynamic radio resource management through cross-RAT cooperation in the radio network
3GPP RAN anchored mobility and control: exploit cellular link as anchor for improving handoff and reliability/QoS of WLAN link

Fig. 1: Topology and architecture of a converged heterogeneous network

long-term strategy to address future capacity needs. In the simplest case, no cooperation between WiFi and cellular RAN is available and the users are left to determine how the two RATs are utilized. However, when WiFi is managed as part of an operator's RAN, increased level of cooperation between WLAN and 3GPP infrastructure may become feasible.

For instance, one may envisage an architecture where integrated WiFi-LTE small cells enable full cooperation between the two RATs, allowing for WiFi to simply become a "virtual carrier" anchored on the 3GPP radio network. We note that multi-RAT small cells with collocated WiFi and 3GPP interfaces are an emerging industry trend for lowering deployment costs by leveraging common infrastructure across multiple RATs. However, given that such deployments are presently not common, current standardization efforts aim to improve UE-centric interworking architectures while assuming only limited cooperation or assistance across a multi-RAT network.

A. Options for integrating WiFi with 3GPP LTE

We continue by illustrating various architectural choices for integrating WiFi and LTE networks in Fig. 1(b). These generally deliver different mechanisms to implement important operations required for multi-RAT integration, including RAT discovery, RAT selection or assignment, control of multi-RAT radio resource management (RRM), protocols for inter-RAT mobility or session transfers, etc.

1) Application Layer Integration: In Fig. 1(b), *Case A* corresponds to the application or higher-layer integration architecture. Accordingly, there is a proprietary or higher layer interface allowing the UE and the content server to communicate directly by exchanging information over multiple RATs. As no coordination at the network layer is involved, such solutions are typically simple and have already been explored in the context of improving over-the-top applications. This choice of architecture is beneficial for boosting user quality of experience (QoE), but it remains largely application-dependent and may not fully account for underlying network conditions, especially when such conditions vary dynamically.

2) Core Network Based Integration: Further, *Case B* summarizes recent solutions proposed by 3GPP for cellular/WLAN integration basing on interworking within the core network. Accordingly, ANDSF assists in discovery of WiFi access points and may also specify policies for network selection, but the overall network selection decision remains in control of the UE. Therefore, it can combine the local radio link state information, operator policies, and user preferences to make a decision that improves user QoE.

There are a number of benefits with this integration option, as it can more adequately account for both operator policies and user preferences. However, the performance of corresponding control procedures may still be rather limited. This is due to the fact that the UE may only have local knowledge of the network conditions and is thus likely to make greedy decisions ultimately hurting overall system performance. Whereas the UE can be made to report its perceived radio link state to the core network, such information exchange cannot be updated dynamically due to prohibitive levels of associated signaling overhead. Hence, when wireless channel conditions change dynamically, local RRM directly on the RAN layer may deliver higher QoS. Therefore, advanced architectures allowing for multi-RAT integration within the RAN are of increasing interest today, as they employ network-wide knowledge of radio link conditions.

3) RAN Based Integration: Finally, *Case C* details the emerging RAN-based 3GPP/WLAN integration architecture. Here, UE assistance may facilitate information exchange between cellular and WLAN infrastructure or a dedicated interface may be introduced for that matter. The available levels of cooperation within the RAN are constrained by the capacity of the inter-cell/inter-RAT backhaul links. When high-capacity backhaul is available or in case of integrated multi-RAT small cells, full cooperation across multiple RATs may become available, thus enabling more dynamic RRM for improved system and user performance.

In addition, the cellular RAT may be employed as a mobility and control anchor: a user thus utilizes 3GPP protocols for transferring sessions to multi-radio small cells and then uses local switching to steer sessions to/from WLAN with low latency. The benefits of this solution are obvious, as adaptations to dynamic variations in interference conditions can easily be performed without undesired session interruptions and packet drops. Further, user and operator preferences may be accounted for through appropriate feedback by the UE or via a suitable configuration of the RAN by the operators.

In summary, the degrees of cooperation within the RAN can range from exploiting simple assistance information (such as network loading) by the radio network to tight coupling and joint/centralized RAN-based RRM. In what follows, we describe the various levels of cross-RAT cooperation options across a multi-RAT HetNet and then characterize the associated performance benefits. We pay particular attention to the more practical case when only limited assistance across multi-RAT network is available to users, by contrast to significantly more complex network-controlled approaches requiring higher signaling and computation overheads.

B. Algorithms for radio resource management

In what follows, we detail various options for utilizing and managing multi-RAT radio resources available in the network. Both user- and network-controlled (or assisted) RRM may be considered for the range of architectural options described above. For application or core network based integration (*options A and B*), only UE-based RRM schemes may be feasible. A richer set of choices is available for RAN-based multi-RAT integration (*option C*), which depend on the degree of inter-RAT cooperation achieved with different RAN topologies.

Generally, RAN can play a major role in multi-RAT resource management across the HetNet. Even if RAN does not directly control the RRM decisions, it may provide optimized network assistance to enable better decisions by the UE. In *virtual* RAN architectures, where the mobility and control anchor is moved from the core network to the RAN, more dynamic RRM with fast session transfers between RATs (dynamic switching) may become feasible. For integrated multi-RAT small cells or where the delay between the interfaces is negligible, tighter cooperation involving joint RAT scheduling may also be enabled.

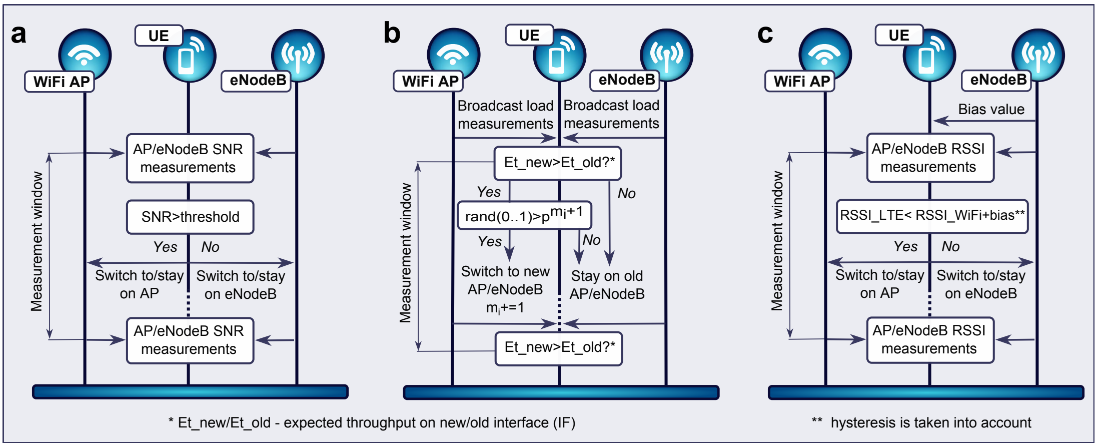
We continue by introducing specific RRM schemes that are investigated in our research. They range from typical implementations used by UEs today, where the UE always prefers to connect to the less expensive WiFi network if it is available (WiFi-preferred), to more intelligent cross-RAT access network selection for converged HetNets.

1) User-centric approaches: The simplest threshold-based algorithm serves as our baseline user-centric network selection scheme. With this solution, a UE is continuously monitoring the signaling messages from the neighboring WiFi access points (APs) to obtain timely signal-to-noise ratio (SNR) information. When a particular SNR value exceeds a predefined threshold (which we set equal to 40 dB as discussed in 3GPP), the user starts steering its traffic to the respective WiFi AP. Otherwise, it keeps transmitting on LTE network (see Fig. 2(a) for details).

Naturally, such behavior is an automatized version of what a human user would do: whenever a hot-spot with reliable signal is available, UEs switch to WiFi to enjoy higher data rates and reduce expenses associated with paid cellular traffic. Alternative user-centric algorithms include schemes based on preferring WiFi if certain minimum performance (coverage, QoS, etc.) is available, as well as solutions where the UE is able to transmit on both RATs, without any intelligent coordination across them.

2) RAN-assisted approaches: Due to its simplicity, the baseline WiFi-preferred (SNR-threshold) scheme may experience limitations in dense interference-limited scenarios which are typical for modern urban deployments. For instance, a hot-spot AP may experience overload conditions when a significant number of users try to steer their traffic through it. Moreover, nomadic WiFi users, such as those with laptops, could consume most of WLAN capacity. To make matters worse, the WiFi medium access is contention-based which results in non-linear degradation of the throughput performance with increasing number of users.

Therefore, the load-agnostic SNR-threshold scheme is not expected to remain effective in environments with varying load. In such situations, UEs may attempt to combine SNR knowledge with additional knowledge of the loading in-



* Et_new/Et_old - expected throughput on new/old interface (IF)

** hysteresis is taken into account

Fig. 2: Alternative network selection algorithms for HetNets

formation from the network infrastructure (cellular/WLAN). While accounting for WiFi load would certainly improve performance beyond the SNR-threshold scheme, it is easy to envision scenarios where accounting for WiFi load *only* will not be sufficient. Hence, we focus our further investigation on schemes that account for both LTE and WiFi loading and compare them with existing network-based schemes which have been standardized in 3GPP for small cell offload. Our proposed load-aware scheme works as follows (see simplified time diagram in Fig. 2(b)).

Throughput estimation: User attempts to listen on both interfaces in order to monitor the SNR information in its neighborhood and estimates its expected throughput. For WiFi, such estimation is conducted based on predicted network capacity divided by number of UEs connected to a particular AP (as advertised by AP through the load indicators in the beacon frames) as well as accounting for several weighting factors (SNR, contention, etc.). The motivation behind the SNR weighting is to exclude APs with low signal quality. Another coefficient may account for the contention-based nature of WiFi channel access and include signaling overheads as well as collision losses. For LTE, throughput prediction may be simply built on the scheduler advertisements by base station (BS or eNodeB) and the used power control.

Randomization: User may select the network with the highest expected throughput value probabilistically $rand(0..1) < p^{m_i+1}$, where m_i is the number of recent connections to this AP/BS and p is the number in (0, 1), which is representing the re-connection probability. The proper use of p reduces the number of concurrent re-connections to the same AP/BS, which will prevent uncontrollable hopping from one interface to another. If a network re-selection occurs, m_i is incremented for AP/BS i. Other users are taking into account this information by dividing their expected throughput value for this AP/BS by m_i+1. This allows to control dynamic re-selections on both networks.

Hysteresis: To additionally decrease the number of cell-border switchings, an appropriate hysteresis value should be added to the current expected throughput value.

Filtering throughput estimations: Further improvement in throughput estimates is obtained through averaging. After each measurement window, the actual throughput obtained over this period may be filtered with a moving average filter. The resultant value, which combines the measured and the predicted throughput, is then used as the expected throughput value for this AP/BS. This averaging is made to achieve more reliability, which could suffer due to contention-based channel access.

In summary, RAN-assisted approaches employ network assistance from the RAN to improve UE-based RAT selection decisions. Network assistance can be very simple in that RAN may transmit certain assistance parameters (e.g., network load, utilization, expected resource allocation), but with increased cross-RAT cooperation RAN assistance may also be improved.

3) RAN-controlled approaches: The above two network selection schemes are user-centric in nature. Hence, they may still result in sub-optimal system-wide performance, which may otherwise be improved through network-based centralized mechanisms. Consequently, RAN-controlled approaches place the control of the RRM in the radio network so that the BS could assign the UEs to use certain RATs. Such network control may be distributed across base stations, or may utilize a central RRM entity that manages radio resources across several cells/RATs.

Below we consider the conventional cell-range extension schemes applied in cellular networks to steer users to small cells employing a network-optimized RSSI (Received Signal Strength Indication) bias value. We use the RSSI bias to increase/decrease the effective WiFi AP coverage area depending on the network capacity expectations. One limitation of this method is that the optimal bias value needs to be adapted based on network-wide knowledge of user distribution. For example, our results show that the optimal bias depends on user deployment model as well as the interference levels in the network, which may not always be available as typically WiFi cells may not have a direct interface to cellular BS. In what follows, we evaluate RSSI-based cell-range extension with bias values optimized for the target scenario. We also use hysteresis for the RSSI-based algorithm. The time-diagram of this method is shown in Fig. 2(c).

More generally, network-controlled schemes may utilize proprietary or standardized interfaces between cells/RATs. Distributed network-controlled schemes have recently been discussed as part of the 3GPP study on WLAN/3GPP RAN interworking. Here, the network establishes certain triggers for UE to report measurement on their local radio environment. The final RAT selection decisions are then made by the 3GPP BS based on UE measurement reports. Other examples of centralized RAN-controlled architecture is the emerging dual connectivity, or “anchor/booster” architecture, where the UE always maintains a control link to the macro cell tier and the macro cell centrally manages the user offload to smaller cells. Hence, the macro cell can centrally determine the optimal offload mechanisms.

III. Analyzing Intelligent Access Network Selection

In what follows, we concentrate on the important problem of network selection between LTE and WiFi RATs [11], assuming that WLAN is a part of an operator deployed and managed multi-RAT HetNet. We target feasible practical extensions to improve performance of UE-centric network selection schemes. To be consistent with current network deployments, we consider distributed small cell overlay with standalone WiFi APs, assuming that there is no interface between the WiFi and the 3GPP radio networks [10]. Additionally, we discuss benefits of deploying integrated WiFi-LTE small cells and quantify the respective performance gains.

In particular, we investigate distributed RAT selection schemes that account for network loading information across the LTE and the WiFi technologies and compare them with solutions that only rely on signal strength measurements. We also benchmark the performance of UE-centric RAT selection with optimized network-based load balancing mechanisms. Intuitively, network-centric solutions may seem to offer better performance compared to UE-based approaches as network-wide radio link information across users can be employed to develop optimum RAT assignment algorithms. However, with distributed architectures assuming no direct cooperation between LTE and WiFi RATs, such solutions may only be implemented through extensive UE feedback which could result in significant overheads. UE-centric RAT selection may also be preferred as the UE can better account for user preferences and application QoE.

A. System-level evaluation scenario and results

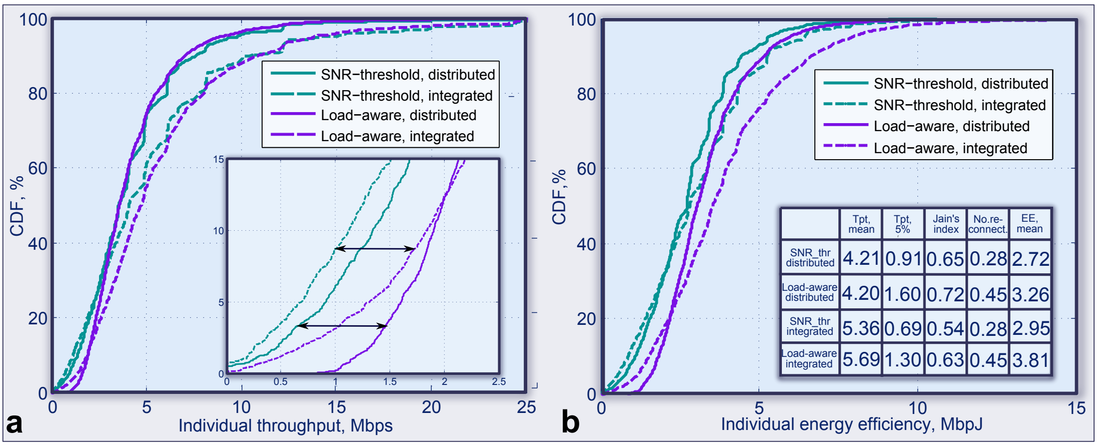
In the course of this study, we have developed an advanced system-level simulator (SLS) that mimics a complete LTE-WiFi system deployment compatible with 3GPP LTE Release-10 and IEEE 802.11-2012 specifications. Presently, neither free nor commercially-available simulation tools are readily applicable for evaluating heterogeneous multi-RAT systems, as they are missing the necessary features, as well as lacking scalability to adequately capture the dependencies between the studied variables. By contrast, our SLS is a flexible tool designed to support diverse deployment strategies, traffic models, channel characteristics, and wireless protocols. It comprises several software modules modeling the deployment of wireless infrastructure and user devices, control events related to transmission of signals between several distinct types of transmitters and receivers, abstractions for wireless channels, mechanisms for collecting measurements and statistics for quantifying system performance, etc.

TABLE I: Important simulation parameters

Parameter	Value
LTE/WiFi configuration	10 MHz FDD / 20 MHz
Macro cell layout	7 cells, 3 sectors each
LTE signaling mode	2 out of 20 special subframes, short CP, 10 ms frame
Inter-site distance (ISD)	500 m
LTE macro Antenna configuration	1x2 (diversity reception)
UE to eNodeB/pico/AP pathloss	ITU UMa/UMi
eNodeB antenna gain	14 dB
eNodeB/AP/UE maximum power	43/20/(23/20 LTE/WiFi) dBm
LTE power control	Fractional (α=1.0) [12]
WiFi power/rate control	Max-power/ARQ
UE/eNodeB/AP antenna height	1.5/25/10 m
UE noise figure/feeder loss	9 dB/0 dB
Feedback/control channel errors	None
Traffic model	Full-buffer
Number of UEs/APs	30/4 per macro cell (3 sectors)
AP/UE deployment type	Uniform/clustered (4b in [12])
AP/UE-eNodeB, AP/UE-UE distance	$>$ 75/35 m, 40/10 m
WiFi MPDU	1500 bytes
Modeling time	3 s
Number of trials per experiment	30

Below we construct a multi-RAT simulation model representative of an urban deployment, where WiFi small cells are overlaid on top of the 3GPP cellular network. Outdoor deployments are considered and are based on recommendations in [12]. A brief summary of the parameters is provided in Table I. Specifically, we consider a loaded (full-buffer) WiFi network with WLAN APs uniformly distributed across the cellular coverage area. Most UEs cluster around the APs, which recreates a hot-spot area (airport, restaurant, shopping mall, or university campus) with many bandwidth-hungry users loading the WiFi network. Moreover, around one third of UEs are still deployed uniformly across the cellular network mimicking regular mobile users. Whereas this scenario may not be characteristic of all practical urban conditions, it represents a harmonized 3GPP vision of a characteristic HetNet deployment.

The major expected outcome of leveraging WiFi small cells is efficient offloading of cellular user traffic resulting in significant user benefits. For that reason, our primary metric of interest is the *uplink* UE throughput (by contrast to



	Tpt, mean	Tpt, 5%	Jain's index	No.re-connect.	EE, mean
SNR_thr distributed	4.21	0.91	0.65	0.28	2.72
Load-aware distributed	4.20	1.60	0.72	0.45	3.26
SNR_thr integrated	5.36	0.69	0.54	0.28	2.95
Load-aware integrated	5.69	1.30	0.63	0.45	3.81

Fig. 3: Comparing SNR-threshold (WiFi-preferred) and our load-aware (RAN-assisted) network selection schemes

many existing studies concentrating on downlink performance) which, in turn, determines the overall system capacity. The cumulative distribution function (CDF) of individual user throughput comparing performance between SNR-threshold (WiFi-preferred) and our proposed load-aware (RAN-assisted) scheme is shown in Fig. 3(a). The results indicate that the load-aware scheme gives visible benefits at the cell edge (e.g., over 75% of improvement is observed in 5% quantile), as well as some improvement in the average throughput for integrated deployments (i.e. with colocated WiFi-LTE interfaces).

Energy efficiency (EE) is also becoming increasingly important for 5G wireless systems due to the limited battery resources of mobile clients [13] and we confirm significant gains in bits-per-Joule metric for both distributed (19%) and integrated (29%) scenarios in Fig. 3(b). Further, as QoS may be equally important, we also account for fairness between the users which indicates how large is the deviation between actual user throughput and the cell-average performance. In terms of fairness, the Jain's index (see table in Fig. 3(b)) of the load-aware scheme (0.72/0.63) is also higher than that for the SNR-threshold scheme (0.65/0.54). Stability of UE-centric schemes is another very important aspect of UE-centric RAT selection, as excessive ping-ponging between RATs is undesirable due to the overhead and latency of mobility protocols as well as due to EE considerations. In Fig. 3(b) (see table), we additionally report the number of cellular/WLAN reconnections (in no. of reconnections per second) and employ hysteresis mechanisms (optimized 3 dB value has been used in our experiments) to improve performance.

We also account for the performance of optimized cell-range extension (RAN-controlled) scheme based on RSSI bias, where the network-wide optimization is expected to result in improved performance. The main feature of the considered cell-range extension scheme is that it increases the effective WiFi/LTE small cell radius with respect to the bias level. This could work well in the scenario with uniformly-deployed UEs, but in the clustered case the interference between WiFi users needs to be considered as well, which is what our load-aware scheme does explicitly. To this end, we perform optimization of small cell offloading bias based on network-wide knowledge of user distribution in Fig. 4(a).

However, from Fig. 4(b) we learn that even with a network controlled bias value (the optimal value of 14 dB is chosen), the individual user throughput is very close to that in the load-aware case (and even smaller at the cell edge). In more detail, Fig. 4(b) (see bar chart) also highlights the average percentage of time spent by users on each interface. It may be seen from our results that the load-aware scheme is effective in utilizing the available WLAN capacity, while efficiently balancing capacities across LTE macro and pico tiers.

B. Analytical space-time methodology for converged HetNets

The above performance results addressed loaded multi-RAT HetNets, but such networks may also be substantially underutilized during the off-peak hours. Hence, the load on a heterogeneous deployment can vary significantly and it is crucial to capture network dynamics explicitly when modeling HetNet performance. However, given the associated complexity, dynamic systems have not been studied as broadly as their static counterparts with a fixed set of active users. Consequently, our proposed analytical methodology suggests assessing flow-level network performance enabling user, traffic, and environment dynamics.

Recently, we have made progress along these lines and have results that demonstrate that the locations of the network users relative to each other highly impact the overall system performance [14]. Indeed, given that users are not regularly spaced, there may be a high degree of spatial randomness which needs to be captured explicitly. We thus adopt a range of random spatial models where user locations are drawn from a particular realization of a random process. Coupling such topological randomness with system dynamics requires a fundamental difference in characterizing user signal power and interference. Fortunately, the field of stochastic geometry

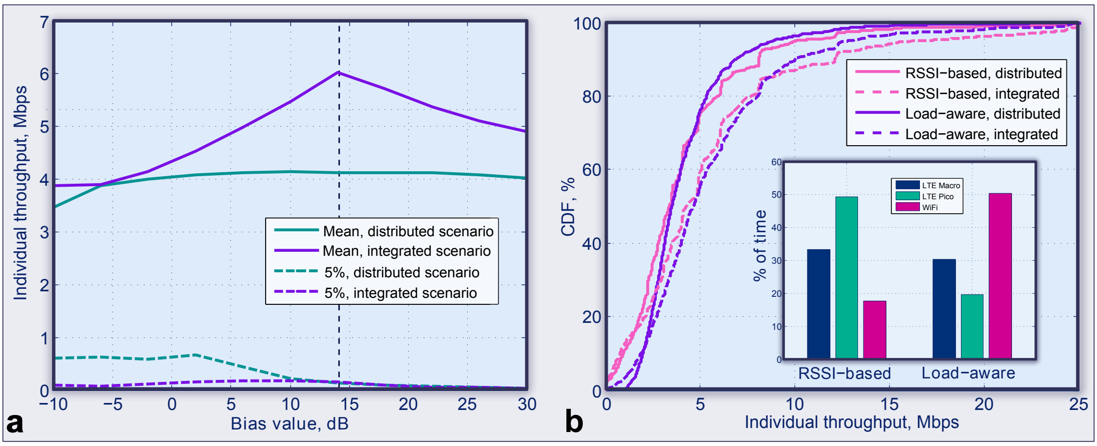


Fig. 4: Comparing our load-aware (RAN-assisted) and RSSI-based (RAN-controlled) network selection schemes

provides us with a rich set of powerful results and analytical tools that can capture the network-wide performance of a random user deployment [15].

More specifically, every data flow in a dynamic network may generally represent a stream of packets corresponding to a new file transfer, web-page browsing, or real-time voice/video session. As an example, consider an isolated cell of a macro network with radius R encompassing a macro BS together with several distributed pico base stations and WLAN APs. All the BSs/APs are capable of serving uplink data from their wireless users concurrently. The considered traffic is characteristic of real-time sessions with some target bitrate. Basing on the recent 3GPP specifications, we further assume non-overlapping frequency bands for all three tiers. However, all WLAN/pico links share the frequency bands of their respective tiers and thus interfere, whereas the macro tier may be considered interference-free (with appropriate inter-cell power control).

To explicitly model topological randomness in our network (see Fig. 5(a)), we employ several stochastic processes and, to this end, adopt a number of simplifications basing on a Poisson point process (PPP). The key novelty of this approach is that we consider a *space-time* PPP with the rate function $\Lambda(x, t)$, where $x \in R^2$ is the spatial component and $t \in R^+$ is the time component. While random network topology is the primary focus of our model, we also couple it with flow-level system dynamics. This involves an appropriate queuing model, where the session arrives and leaves the system after being served (the service time is determined by the random session length). When a new session arrives or a served session leaves the system, the centralized assisting entity in the RAN performs admission and power control on all tiers by deciding whether the session would be admitted to a particular tier or not and/or advising on the users' transmit powers.

Our general system model is illustrated in Fig. 5(b) representing areas of the macro, pico, and WLAN tiers together with the corresponding users and infrastructure nodes. We consider the following *cascade* network selection when a new session arrives into the system. First, the RAN-based network selection assistance entity attempts to offload the newly arrived session onto the nearest WLAN AP by performing the WLAN admission control managed centrally. If the session is accepted on the WLAN tier, it is served there without interruption until when it successfully leaves the system. Otherwise, if this session cannot be admitted onto the WLAN, the pico network admission control is executed and either the session is accepted on the pico tier and served by the nearest pico BS or the macro network itself attempts to serve this session. Eventually, if the session cannot be admitted onto the macro tier either, it is considered permanently blocked and leaves the system unserved.

In Fig. 5(c), we detail the overall blocking probabilities for the converged HetNet as well as for the three tiers individually: macro, pico, and WLAN. Our main observation is that with two additional overlay tiers, the HetNet performance improves significantly over what can be achieved in the macro-only networks (cellular baseline). Remarkably, we actually witness visible performance improvement even with only a few additional infrastructure nodes, such as 2 WLAN APs and 2 pico BSs in this example. Therefore, we believe that multiple RATs and the associated network selection intelligence for their efficient use will become a characterizing feature of future 5G HetNets.

IV. Main Takeaways and Way Forward

In summary, this article reviews major challenges in delivering uniform connectivity and service experience to future heterogeneous 5G networks. It discusses several architecture choices and associated algorithms for intelligent access network selection in multi-RAT HetNet deployments, both when the control of how radios are utilized rests with the network and the user. In particular, it compares simulated performance of RAN-assisted load-aware network selection schemes with conventional/existing UE- and network-based

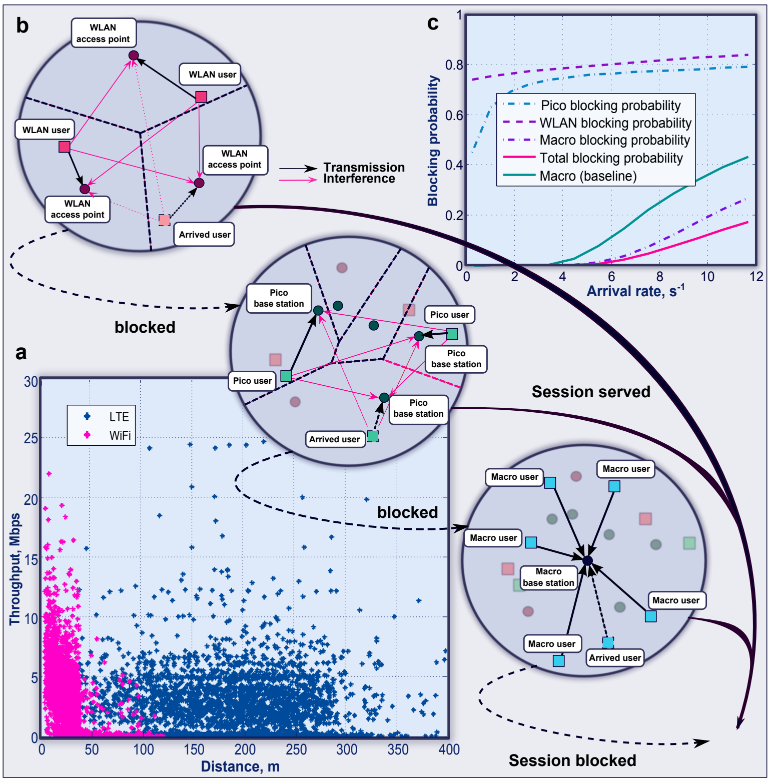


Fig. 5: Illustration of proposed space-time methodology for multi-RAT networks

solutions employed in current systems. We primarily focus on uplink performance as it has not been fully addressed in the past literature.

The main advantages of load-aware schemes stem from the fact that the SNR-threshold (WiFi-preferred) scheme, as well as the network-centric cell-range extension scheme, do not explicitly account for the loading and interference on the WiFi APs typically encountered in clustered UE deployments. Our results show that the load-aware user-centric scheme, which augments SNR measurements with additional information about network load, could improve the performance of WiFi-preferred scheme based on minimum SNR threshold. We observed over 75% improvement in 5% cell-edge throughput as well as significant gains in energy efficiency for both distributed and integrated deployment scenarios.

Comparison with more advanced network-controlled schemes has also been completed across various heterogeneous deployments to confirm attractive practical benefits of distributed user-centric solutions. Next steps include further investigation of UE-based algorithms while explicitly considering load variation in the network and accounting for application-layer statistics. System behavior in the presence of uncoordinated (rogue) WiFi interference must also be accounted for and hysteresis mechanism may further be improved to combat the uncertainty in estimating user throughput.

Building on the system-wide simulation data, we also propose a novel dynamic methodology for RAN-assisted network selection capturing the spatial randomness of HetNets together with unsaturated uplink traffic from its users. Our stochastic geometry based analysis enables in-depth characterization of dynamic interactions between macro and pico cellular networks, as well as WLAN, mindful of user QoS and based on intelligent RAT selection/assignment. Going further, we expect our space-time methodology to be capable of encompassing other technologies beyond LTE and WiFi, as well as additional use cases beyond simple aggregation of capacity across unlicensed bands.

More generally, studying the ultimate capacity of multi-radio wireless networks remains an open problem in the field of information theory and stochastic geometry has the potential to shed light on it given that it can explicitly capture new interference situations and hence the achievable data rates. This challenging objective may require novel advanced analytical tools to interconnect and apply techniques and methods coming from the area of point processes, probability theory, queuing theory, and percolation theory, as well as modern engineering insights.

V. Acknowledgment

This work is supported by Intel Corporation, GETA, and the IoT SRA program of Digile, funded by Tekes. The work of the first author is supported with a Postdoctoral Researcher grant by the Academy of Finland.

References

[1] B. Bangerter et al., "Networks and devices for the 5G era," *IEEE Comm. Mag.*, vol. 52, pp. 90–96, 2014.

[2] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018*, February 2014.

[3] S.-P. Yeh et al., "Capacity and coverage enhancement in heterogeneous networks," *IEEE Wireless Comm.*, vol. 18, pp. 32–38, 2011.

[4] P. Marsch et al., "Future mobile communication networks: Challenges in the design and operation," *IEEE Vehicular Tech. Mag.*, vol. 7, pp. 16–23, 2012.

[5] J. Andrews et al., "An overview of load balancing in HetNets: Old myths and open problems," *IEEE Wireless Comm.*, vol. 21, no. 2, pp. 18–25, 2014.

[6] S. Andreev et al., "Cellular traffic offloading onto network-assisted device-to-device connections," *IEEE Comm. Mag.*, vol. 52, no. 4, pp. 20–31, 2014.

[7] R. Baldemair et al., "Evolving wireless communications: Addressing the challenges and expectations of the future," *IEEE Vehicular Tech. Mag.*, vol. 8, pp. 24–30, 2013.

[8] J. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Comm. Mag.*, vol. 51, pp. 136–144, 2013.

[9] M. Bennis et al., "When cellular meets WiFi in wireless small cell networks," *IEEE Comm. Mag.*, vol. 51, pp. 44–50, 2013.

[10] *3GPP TR 37.842, 3GPP/WLAN RAN Interworking Study Item Report*, 2013.

[11] L. Wang and G.-S. Kuo, "Mathematical modeling for network selection in heterogeneous wireless networks – A Tutorial," *IEEE Comm. Surveys & Tutorials*, vol. 15, pp. 271–292, 2013.

[12] *3GPP TR 36.814, Further advancements for E-UTRA physical layer aspects*, 2010.

[13] S. Navaratnarajah et al., "Energy efficiency in heterogeneous wireless access networks," *IEEE Wireless Comm.*, vol. 20, pp. 37–43, 2013.

[14] O. Galinina et al., "Capturing spatial randomness of heterogeneous cellular/WLAN deployments with dynamic traffic," *IEEE J. on Sel. Areas in Comm.*, vol. TBD, p. TBD, 2014.

[15] H. ElSawy et al., "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Comm. Surveys & Tutorials*, vol. 5, pp. 996–1019, 2013.

Authors' Biographies

Sergey Andreev is a Senior Research Scientist in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. He received the Specialist degree (2006) in Information Security and the Cand.Sc. degree (2009) in Wireless Communications both from St. Petersburg State University of Aerospace Instrumentation, St. Petersburg, Russia, as well as the Ph.D. degree (2012) in Technology from Tampere University of Technology, Tampere, Finland. Sergey (co-)authored more than 80 published research works. His research interests include wireless communications, energy efficiency, heterogeneous networking, cooperative communications, and machine-to-machine applications.

Mikhail Gerasimenko is a Research Assistant at Tampere University of Technology in the Department of Electronics and Communications Engineering. Mikhail received specialist degree in Saint Petersburg University of Telecommunications in 2011. In 2013 he obtained Master of Science degree in Tampere University of Technology. Mikhail started his academic career in 2011 and during 2 years he appeared as an (co-)author of several scientific journal and conference publications, as well as several patents. Moreover, he also acted as a review and participated in education activities. His main subjects of interest are wireless communications, Machine-Type Communications, Heterogeneous networks.

Olga Galinina is a PhD Candidate in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. She received her B.Sc. and M.Sc. degree in Applied Mathematics from Department of Applied Mathematics, Faculty of Mechanics and Physics, Saint-Petersburg State Polytechnical University, Russia. She has publications on mathematical problems in the novel telecommunication protocols in internationally recognized journals and high-level peer-reviewed conferences. Her research interests include applied mathematics and statistics, queueing theory and its applications; wireless networking and energy efficient systems, machine-to-machine and device-to-device communication.

Yevgeni Koucheryavy is a Full Professor in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. He received the PhD degree (2004) from the TUT, Finland. Prior joining the TUT, Yevgeni spent five years in industry with R&D LONIIS in St. Petersburg, Russia, where he held various technical and managerial positions. Yevgeni actively participates in national and international research and development projects – in particular currently he is chairing COST IC0906 "WiNeMO: Wireless Networking for Moving Objects" scheduled for 2010-2014. Yevgeni has authored or co-authored over 100 papers in the field of advanced wired and wireless networking and communications. Yevgeni is a Senior IEEE member.

Nageen Himayat is a Senior Research Scientist with Intel Labs, where she performs research on broadband wireless systems, including heterogeneous networks, cross-layer radio resource management, MIMO-OFDM techniques, and optimizations for M2M communications. She has over 15 years of research and development experience in the telecom industry. She obtained her B.S.E.E from Rice University and her Ph.D. in electrical engineering from the University of Pennsylvania in 1989 and 1994, respectively.

Shu-ping Yeh is a Research Scientist in the Wireless Communications Laboratory at Intel. She received her M.S. and Ph.D. from Stanford University in 2005 and 2010, respectively, and her B.S. from National Taiwan University in 2003, all in electrical engineering. Her current research focus includes interference mitigation in multitier networks utilizing multi-antenna techniques, machine-to-machine communications, and interworking of multiple radio access technologies within a network.

Shilpa Talwar is a Principal Engineer in the Wireless Communications Laboratory at Intel, where she is conducting research on mobile broadband technologies. She has over 15 years of experience in wireless. Prior to Intel, she held several senior technical positions in wireless industry. She graduated from Stanford University in 1996 with a Ph.D. in applied mathematics and an M.S. in electrical engineering. She is the author of numerous technical publications and patents.

Publication 8

© 2015 IEEE. Reprinted, with permission, from

O. Galinina, A. Pyattaev, S. Andreev, M. Dohler, and Y. Koucheryavy, "5G Multi-RAT LTE-WiFi Ultra-Dense Small Cells: Performance Dynamics, Architecture, and Trends", *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 1224-1240, 2015.

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Tampere University of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to `http://www.ieee.org/publications_standards/publications/rights/rights_link.html` to learn how to obtain a License from RightsLink.

5G Multi-RAT LTE-WiFi Ultra-Dense Small Cells: Performance Dynamics, Architecture, and Trends

Olga Galinina, *Student Member, IEEE,* Alexander Pyattaev, *Student Member, IEEE,*
Sergey Andreev[†], *Member, IEEE,* Mischa Dohler, *Fellow, IEEE,* and Yevgeni Koucheryavy, *Senior Member, IEEE*

Abstract—**The ongoing densification of small cells yields an unprecedented paradigm shift in user experience and network design. The most notable change comes from cellular rates being comparable to next-generation WiFi systems. Cellular-to-WiFi offloading, the standard *modus operandi* of recent years, is therefore shifting towards a true integration of both technology families. Users in future 5G systems will thus likely be able to use 3GPP, IEEE, and other technologies simultaneously, so as to maximize their quality of experience. To advance this high-level vision, we perform a novel performance analysis specifically taking the system-level dynamics into account and thus giving a true account on the uplink performance gains of an integrated multi radio access technology (RAT) solution versus legacy approaches. Further, we advocate for an enabling architecture that embodies the tight interaction between the different RATs, as we lay out a standardization roadmap able to materialize the envisaged design. 3GPP-compliant simulations have also been carried out to corroborate the rigorous mathematical analysis and the superiority of the proposed approach.**

I. Introduction

Fueled by the increasing popularity of handheld mobile devices with powerful data processing capabilities, the wireless industry is witnessing an avalanche of mobile traffic. Indeed, the amount of data produced by smartphones, tablets, PDAs, and new types of mobile computing devices has recently been doubling every year with this trend very likely to continue over the following decade. This unprecedented escalation has imposed significant challenges on the design of existing wireless networks. Subsequently, we outline the state of the art in cellular design, which clearly shows that these trends cannot be met with legacy approaches. We then explain the rationale of integrating WiFi with legacy cellular systems, before discussing trends beyond state of the art, along with our specific contributions in the field.

A. Densification in cellular technologies

Over the first decade of this century, we have already seen a 1000-fold growth in capacity of wireline communication systems with another 1000-fold increase targeted by the year 2020. Consequently, mobile network operators have been taking steps to improve performance of their deployments as part of the fourth generation (4G) Long-Term Evolution (LTE) communications technology. However, whereas cellular *coverage* is now nearly ubiquitous, 4G will not suffice to provide the needed *capacity*. A large factor here is the shortage of a suitable network-side infrastructure.

Today, the macro cell network capacity cannot continue to scale any further. The expensive tower-mounted macro cells generally demand high installation, maintenance, and backhauling costs as well as elaborate site planning, and thus suffer from the lack of available sites. Therefore, it is cumbersome for operators to further optimize their macro cell deployments. Furthermore, the conventional wireless link throughputs are very close to their theoretical limits, whereas adding more bandwidth often has prohibitively high costs [1].

It has therefore been recognized that an increase in capacity per unit area may be achieved by shrinking the cell [2]. Correspondingly, LTE networks have evolved to include numerous nested small nodes, such as pico cells (with the coverage of under 100 meters) [3] and femto cells with a WiFi-like range [4], as well as a multitude of relay nodes and distributed antenna systems. With this trend, known as *network densification*, the massive numbers of network radio units are brought closer to the user equipment (UE) [5].

While many of today's pico cells are still installed by the mobile network operators in a (semi-) planned manner, the deployment of femto cells remains mostly unplanned [6]. This is due to the presence of other parties (building owners, as well as the end-users), who install additional femto cells in homes, small offices, and enterprises. It is obvious that improved traffic capacity may be required in home and office environments, as well as in public places (shopping malls, subway stations, etc.) or other dense urban environments, where large numbers of users are located within a limited geographic area [7]. Whereas increasing densities of such residential, enterprise, and hotspot small cells are providing a more cost-efficient approach to spatial densification [8], there may soon be more base stations than there are users [9]. This extreme network densification – being a combination of spatial densification (with distances between access nodes in tens of meters) and spectral aggregation (across licensed and unlicensed spectrum) – is a major paradigm shift which has recently been named one of the "big three" in the fifth generation (5G) technology ecosystem (see [10] and references therein).

As 5G-centric research is taking shape, it becomes apparent that a single (new) radio technology will not be able to satisfy all the associated performance requirements and character-

O. Galinina, A. Pyattaev, S. Andreev, and Y. Koucheryavy are with the Department of Electronics and Communications Engineering, Tampere University of Technology, Tampere, Finland.

M. Dohler is with the Department of Informatics, King's College London, London, UK.

[†]S. Andreev is the contact author: P.O. Box 553, FI-33101 Tampere, Finland; e-mail: sergey.andreev@tut.fi

Manuscript received 21.07.2014; revised 17.12.2014, 15.03.2015.

istics [11]. However, *ultra-dense* HetNets, enabling efficient reuse of spectrum across a certain *area of interest* [12], [13], are very likely to become the only viable solution on the road to 1000x capacity in a less than ten year time frame [14].

B. Integration with WiFi networks

Owing to their more compact form factor, LTE small cells are preferred due to deployment flexibility, lower capital and operational costs, as well as reduced energy usage [15]. As a result, they become similar to WiFi access points that have historically been a de-facto solution for local-area connectivity [16]. This creates an attractive opportunity for an increasing cooperation between LTE and WiFi radio access technologies (RATs) to realize effective offloading and/or balancing of user data traffic, and thus leverage resources available in both systems [17].

The latest standardization work by 3GPP pays lots of attention to the integration of WiFi, with an array of options discussed. This includes trusted access to cellular services for WiFi-only devices, seamless WiFi/LTE mobility, as well as the Access Network Discovery and Selection Function (ANDSF), among others. More recently, in 3GPP Release-12 of LTE, solutions for tighter coupling between cellular and local-area communication at radio access network (RAN) level have been investigated. The degrees in which RAN-level integration may improve LTE/WiFi cooperation are numerous, ranging from providing simple assistance information (such as network loading [18]) to full-scale joint (centralized) radio resource management [19]. In general, the consensus today is that if there is an opportunity to install WiFi in co-location with an LTE deployment, it should be preferred, primarily due to a marginal increase in associated costs [20] for a significant added value of harnessing license-exempt spectrum that WiFi systems can utilize. With advanced multi-RAT integration options, a mobile device could potentially transmit data on both radio interfaces *at the same time*, which is expected to improve its performance [21]. However, novel multi-RAT and multi-tier solutions require additional infrastructure enablers, such as new network management interfaces [22], able to deliver flexible core network connectivity for the envisioned system architecture of next-generation 5G systems [23].

In summary, diverse types of low-power and low-cost small cells that operate in both licensed (e.g., LTE) and unlicensed (e.g., WiFi) frequency bands and connected to the core network(s) with various types of backhaul links, comprise the emerging vision of a *heterogeneous network* (HetNet) [24].

C. Scientific contribution

The consideration of ultra-dense small cell networks with co-located LTE and WiFi radio interfaces requires a fundamental change in system characterization, including respective modeling (analysis, simulation) and visualization. In this paper, **we offer the following contributions** to conduct a thorough performance investigation of future ultra-dense HetNets.

1) **Architecture.** We provide a comprehensive review of the available options to provide flexible and intelligent LTE/WiFi integration over the current 3GPP architecture. In particular, our proposed architecture enables dynamic data flow splitting across integrated dual-RAT infrastructure, as well as stand-alone deployments with arbitrary composition of LTE/WiFi small cells.
2) **Analytical Framework.** We introduce a novel analytical methodology for ultra-dense LTE/WiFi HetNets, coupling spatial randomness of user distribution [25] with their uplink data dynamics [26]. Given that the load of each small cell varies significantly over time and space [27], we consider an area of interest where a small cell with co-located LTE/WiFi interfaces is deployed ”on each lamppost”, and deliver a comprehensive analytical model.
3) **Corroborating Simulations.** We detail empirical results produced with our 3GPP-compatible system-level simulator, which is calibrated with real-world deployments and accurately mimics the behavior of a practical ultra-dense HetNet. The quantitative evidence obtained with this tool is then used to substantiate the core assumptions of our analytical methodology as well as make important conclusions on the degrees of its accuracy.
4) **Comparative Case-Study.** A rigorous analytical evaluation and comparison of two distinct RAT selection mechanisms: one is characteristic of how users employ their multi-radio devices today (based on preferential use of one RAT), while another scheme is made available by our proposed integrated multi-RAT architecture (simultaneous use of LTE and WiFi radios).

The paper is structured as follows. In Section II, we outline the working architecture to enable the simultaneous use of cellular and WiFi technologies. In Section III, we outline the system assumptions which underpin the subsequent mathematical analysis. The rigorous analytical framework, explicitly taking system dynamics into account, is outlined in Section IV. The simulation framework and respective results are presented and discussed in Section V. Finally, in Section VI, conclusions are drawn and future trends are outlined.

II. Integrated Multi-Radio Architecture

Instrumental to the cause of tight integration is a viable system architecture. In this section, we briefly outline operational challenges along with current integration approaches, before suggesting some improvements to the architecture as well as several new building blocks.

A. Current operator vision and challenges

Mobile network operators today control more than just the wireless last mile. In fact, a typical portfolio of a contemporary network operator includes (i) a multi-tier cellular network featuring both legacy and LTE infrastructure nodes; (ii) an IP access network, including city-wide local-area WiFi deployments; (iii) a notable set of IMS services (e.g., mobile TV and radio) facilitating delivery of bandwidth-hungry content to the users; and (iv) an abundance of unlimited data users in the service area given the popularity of smartphones and tablets.

The drastic increase of unlimited data users poses a significant challenge to the RAN capacity; all possible forms of mobile traffic offloading are thus becoming vital for mobile network operators [28]. Therefore, a truly converged multi-RAT HetNet (with several *alien* RATs, such as WiFi, operating seamlessly with the 3GPP mobile architecture) is an important enabler for cost savings. However, integration of other RATs has to be done intelligently; notably, UEs ought to operate efficiently without violating quality-of-service (QoS) requirements.

However, current 3GPP integration options are not nearly as flexible to enable efficient multi-radio connectivity [29] as one may think. For instance, any non-3GPP access technology is to have a termination point at the packet data network gateway (PDN GW). Therefore, any UE that wishes to use a non-3GPP (alien) RAT together with cellular access must route its traffic to the operator's PDN GW, from where the packets will be forwarded to the destination following the logic of mobile IP. Since mobile IP traffic is blocked by most network address translation (NAT) devices, which are abundant in today's IPv4 infrastructure, the mobile IP session needs to be tunneled to the PDN GW through a VPN tunnel. The resulting solution is extremely fragile and bulky, yet it presently constitutes the *only* available option. Naturally, if IPv6 is available, a VPN tunnel is not needed anymore, which makes the architecture somewhat more elegant. However, it does not resolve the *key challenge* – all of the traffic has to reach the PDN GW before it is routed anywhere – to the Internet or another device.

This, unfortunately, may result in unnecessary detours for the traffic, increasing delays and causing excessive congestion. Most importantly, while the use of mobile IP does make WiFi mobility somewhat similar to the cellular mobility, it is nowhere near the same: non-3GPP access cannot be used without interruptions due to lags between RAT switching and Mobile IP reactions. As a result, the existing 3GPP architecture for alien access does not solve the problem of efficient multi-radio networking, since every time the data path is switched from LTE to WiFi and back, some packets are lost, while simultaneous usage is not possible. Below we further detail the existing 3GPP architecture and offer our proposed solution to mitigate this situation.

B. Existing integration choices for LTE and WiFi systems

By convention, all user data in LTE is represented as IP packets, and all IP packets are hauled with a fixed QoS level through their respective evolved packet system (EPS) bearers, which act like virtual circuits. As a result, the LTE network internally operates as a circuit-switched system, while externally appearing to be packet-switched. This allows the flows of data to be routed and prioritized in the operator's network in any way desired, such that QoS can be maintained as the users move around. Whereas this indeed provides flexibility the cellular systems require, it is very different from how IP works. As a result, no plain IP traffic from the users is actually allowed inside the LTE network.

Above has a profound implication on a LTE-WiFi integration, notably when considering mobility. If user mobility occurs inside the LTE network, handovers can be reliably hidden with the help of the mobility management entity (MME) which facilitates the tunnel switching. If, however, the mobility handover happens towards or from a non-3GPP RAT, such as WiFi or WiGig, then even if the access point (AP) is co-located with one of LTE base stations (eNodeBs), Mobile IP needs to be used to do the flow switching which yields packet losses. In addition, the PDN GW has to act as the home agent. Finally, if IPv6 is not supported at some point along the path between the WiFi AP and the PDN GW, a VPN tunnel ought to be used to connect to the PDN GW prior to engaging with mobile IP procedures.

Similar engineering difficulties arise when developing a multi-RAT, co-located WiFi/LTE HetNet solution. One would need, at the very least, to have LTE or WiFi available for Internet access at any time; preferably, one would also need to be able to switch between the two RATs transparently. To achieve this, the WiFi AP needs to be connected to the Internet, such that the UE is able to reach the PDN GW (or Trusted Wireless Access Gateway, TWAG) with its tunnel, and then rely on that Mobile IP works sufficiently fast during handovers. Therefore, every co-located LTE-WiFi small cell must be connected to the operator's internal network (where S1, S3, S4, and X2 interfaces operate), but **also** to the Internet/TWAG, complicating deployments. And the only reason it needs to be connected there is to allow the UE to reach the PDN GW from the *outside* of the evolved packet core (EPC).

It would, however, be simpler, more efficient, and elegant to make a WiFi AP "pretend" to the EPC that it is an eNodeB, and its physical-layer technology is LTE. This would allow to run all of the IP adaptation protocols on top of the WiFi MAC. Based on this idea, we continue with our alternative view on how technologies like WiFi ought to be integrated into a viable 3GPP architecture.

C. Proposed changes to the 3GPP architecture

The aim of the envisioned HetNet system improvements are to better manage alien, i.e. non-3GPP, flows with the EPC infrastructure. Whilst going beyond the Release-10 HeNB (femto) Local IP Access (LIPA) as well as Trusted-WiFi functionalities, one would need a new entity on the interface between WiFi and EPC which masks the differences between a WiFi AP and an actual eNodeB towards the UE and the core network, and supports signaling on all appropriate interfaces. We term this new entity *alien access gateway* (AAGW) and advocate its deployment as part of the small cells or WiFi segments towards truly integrated LTE-WiFi systems, along with phasing out of the existing TWAG and PDG entities.

The AAGW acts as a transparent link-layer proxy, where the cellular side runs all the characteristic protocols (such as SCTP to communicate with other base stations and the MME), while the non-3GPP access side uses WiFi directly to transfer all the packets that would normally go over the U-interface between the eNodeB and the UE (most importantly, RRC packets would be necessary to set up the bearers). On the UE side, both WiFi and the tunnel interfaces are linked with the upper layers of the cellular stack, which is now able to open a fully functional

EPS bearer to the PDN GW over the WiFi's PHY and MAC. Surprisingly, beyond AAGW, no other changes to the 3GPP architecture are necessary, as the rest of the network can safely assume that the new entity is yet another cell or eNodeB. As a purely software solution, AAGW comes with maximal flexibility and minimal capital costs.

As a result of AAGW deployment in small cells and on WiFi APs, all non-3GPP traffic is injected before the serving gateway, by masking the injection point as yet another cell. In effect, a co-located LTE-WiFi base station would now behave as if it serves two cells; and all it needs to operate is the ISP's internal network (such that S1 and X2 interfaces with serving gateways and other eNodeBs can be activated). The deployment of co-located stations would thus be greatly simplified, as they will no longer need separate access to the Internet/TWAG (see Figure 1 for details). Lastly, with AAGW all sorts of cellular mobility scenarios between LTE and WiFi are as easy as if they were the same technology.

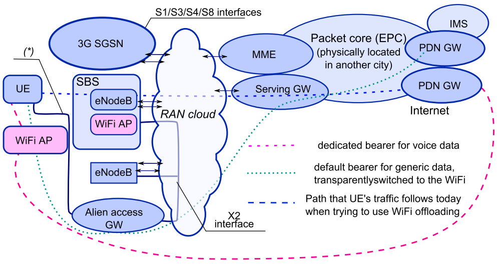

S1/S3/S4/S8 interfaces
3G SGSN
(*)
UE
SBS
eNodeB
WiFi AP
WiFi AP
eNodeB
Alien access GW
RAN cloud
X2 interface
MME
Packet core (EPC) (physically located in another city)
IMS
PDN GW
Serving GW
PDN GW
Internet
dedicated bearer for voice data
default bearer for generic data, transparentlyswitched to the WiFi
Path that UE's traffic follows today when trying to use WiFi offloading

Fig. 1. Proposed high-level architecture and resulting data flow for 5G multi-RAT LTE/WiFi networks.

We continue by considering the case when a WiFi AP is not directly connectable to the S1/X2 interfacs. An operator may want it for security reasons; or when AP is deployed by a third party (shopping mall, building owner, train station, etc.) [30]. Such situations were likely the main reason driving the 3GPP to employ access via VPN and PDN GW in the current architecture. However, unless an operator installs a PDN GW in every neighborhood, the path utilizing such topology will always remain suboptimal. In contrast, to enable cost-efficient operation without stretching the presence of EPC all over the coverage area, one could further extend the features of our proposed AAGW as we outline below. Since AAGW is a cheap software entity, and *does not have direct core access*, it could be deployed in large quantities.

With a suitable tunneling protocol (such as L2TP), one can securely and efficiently tunnel L2 messages from the alien side of the AAGW to the UE. All what is required is to replace the WiFi integration with a VPN server – and the packets from the UE could be streamed to the AAGW from any access point. To differentiate between the users, EAP/SIM-based authentication could be used. Once the UE packets reach it, the AAGW can use all of the flow management tools that LTE network provides, including transparent handovers from one AP to another. Apparently, to use the proposed AAGW (co-located or over VPN), the UE will need special drivers installed. We believe this to be the main hindrance in the deployment process, yet similar problems are encountered when using existing 3GPP architecture (as mobile IP is not deployed today on the UEs). Finally, the UE needs to know the IP address of its nearest AAGW through an appropriate mechanism, such as ANDSF, anycast or DNS. When all of the above is put together, the usage of operator's WiFi would be transparent for the network layer, which enables all of the cellular benefits, such as transparent mobility.

Finally, as all of the WiFi APs are essentially mimicked as eNodeBs/cells with our proposed HetNet architecture, it is possible to dynamically *split* the packet flows between several LTE cells and WiFi AP's by utilizing the existing CoMP signaling procedures. CoMP essentially serving an UE over different cells, thus **enabling true multi-path operation without the need for transport-layer solutions**, such as multi-path TCP. In what follows, this new capability of flexible data flow splitting is evaluated with our mathematical and simulation tools.

III. Envisioned Network Model and Assumptions

Relying on the above HetNet architecture enhancements, we proceed with the description of networking and modeling assumptions.

A. Scenario description

We focus on the *area of interest*, which is represented as a square of side D as illustrated in Figure 2. We further consider a set of *integrated small cells*, where every small base station (SBS) is equipped with LTE and WiFi radio interfaces. As recent discussions in 3GPP recommend band-separated small-cell deployment [31], which is complementing the macro cell band, all of our SBSs share the same LTE frequency. In our scenario, the macro cell uses orthogonal set of resources and handles all of the signaling and scheduling, but does not participate in the data transfers, acting as "anchor". WiFi conventionally operates in unlicensed spectrum. In practice, such scenarios are envisioned to efficiently serve, shopping centers and other locations with exceptionally high user density. A centralized entity (e.g., residing at the MME) is assumed to have full knowledge of all relevant system information, including positions of the SBSs and the associated users.

In what follows, we analyze the *uplink* traffic demand from the users located within the area of interest, whose multi-radio terminals may utilize both LTE and WiFi connections simultaneously [32] with the technology that we have outlined in the previous section. In future HetNets, downlink and uplink are two different networks [21], and our focus on the uplink transmission here is justified by a significantly higher degree of *topological randomness* in the resulting system as compared to the downlink operation [33]. To model topological and temporal randomness jointly, we introduce the following assumptions.

1) Spatial distribution of users and infrastructure: The locations of SBSs are modeled as a homogeneous Poisson point process (PPP) $\mathbf{S}_{BS}$ with intensity μ, thus the number of SBSs in a certain area is a Poisson-distributed random variable, and the numbers of SBSs in disjoint areas are independent. We

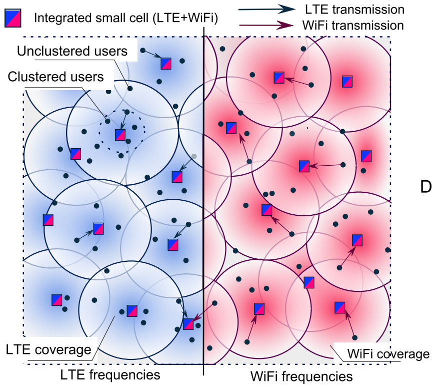


Fig. 2. Envisaged system topology with the area of interest.

TABLE I
SYSTEM EVALUATION PARAMETERS

Name	Description/definition	WiFi	LTE
D	Size of area of interest	2000m	
R	Coverage radius	100m	100m
μ	SBS density	100 per km^2	
ϵ	Share of unclustered users	0.7	
r	Cluster radius	10 m	
λ	Tx requests' arrival rate	var	
θ	Average size of data file	1.5Mbit	
w	Channel bandwidth	20	10
N_0	Noise level	-106dBm	-106dBm
$r_{\lim}$	Rate upper bound	160Mbps	80Mbps
G	Propagation constant	300	250
κ	Propagation exponent	6	6
σ	Available resource per SBS	0.7	1
η	Target SNR	20dB	15dB
$p_{\max}$	Maximum Tx power	23dBm	20dBm
ν	Resource reuse coefficient	0.4	0.4
η_0	SNR threshold for selection	20dBm	N/A

further assume *ultra-dense* SBS deployment, which is denser than the conventional grid layout, and is formally defined here as $\mu\pi(2R)^2 > 7$, where R is the the coverage radius of an individual SBS, such that the probability of not having small cell coverage at a given point of space is thus negligibly small.

From the practical side, such densities of cellular deployment would be justified in cases when it is preferred to install multiple cheaper, low-capacity stations due to capacity constraints or deployment limitations (e.g., absence of suitable sites). Examples include transit hubs, shopping malls, etc. Numerous SBS could prove to be much easier and cheaper in such settings due to stringent constraints on the radiation power when antennas are very close to the people. Even with additional WiFi radios, such devices could be much safer and cheaper to operate than conventional solutions based on leaky cables and other similar RF solutions.

The locations of arriving users are distributed according to *a mixture* of PPP and cluster processes. While a certain proportion ϵ of users are distributed according to a PPP, the rest of the users concentrate around the respective SBSs, resulting in a *cluster* point process based on the process $\mathbf{S}_{BS}$. The overall distribution of the mixed user distances is $f_d(d) = \epsilon f_u(d) + (1-\epsilon) f_c(d)$, $d \geq R$, where $f_c(d)$ is the distribution of distances for clustered users.

More specifically, the locations of the clustered nodes follow a Matern cluster process, where user positions are uniformly distributed within the circle of radius r around the points of a PPP realization. In our case, the locations are grouped around the SBSs and we assume r to be small enough, such that the users with the corresponding distances would *always* be associated with their central SBS. The above considerations are reflective of the current 3GPP documentation, discussing a weighted mixture of uniform and cluster distributions as a realistic model for user locations [20], as well as of the practical use cases when small cells are targeted to serve shops, restaurants, bars, and other populated areas.

2) Traffic dynamics: Data transmission requests from users in the area of interest arrive in time according to a stationary homogeneous Poisson process of intensity λ. We associate a newly arrived request with a new user appearing on the plane, and term it a *connection*. One may also consider connection to match some existing user changing its location while continuing data session. Due to the memoryless user location model, such interpretation implicitly captures mobility of the users.

In our model, a transmission request corresponds to a *data file* of exponential size with the average of θ bits. After the data transmission is completed, the user leaves the system. The location of the user i acquiring a new data file to transmit determines the quality of its channel to the SBS. Hence, the corresponding data rate r_i is bounded by the Shannon's limit:

$$r_i = \min\{r_{\lim}, w \log(1 + SINR_i)\}, \tag{1}$$

where $SINR_i = \frac{\gamma_i}{N_0 + I} p_i$, p_i is the output power, γ_i is the path **gain** between the transmitter and the receiver of the connection i, w is the effective channel bandwidth, N_0 is the noise level, I is the interference level at the receiver, and $r_{\lim}$ is the upper bound on the achievable data rate. The path gain γ_i between the transmitter and the receiver of a particular connection obeys the dependence of transmit power on the distance d_i between them, i.e., $\gamma_i = \frac{G}{{d_i}^{\kappa}}$, κ is the propagation exponent, and G is the propagation constant. This path gain model can be adjusted to most COST/Hata based isotropic environment models, including 3GPP micro cell models.

The described channel model is a simple yet powerful tool to characterize the limits of system densification without the complexity of small-scale power fluctuations caused by such processes as random fading, thus leading to simpler analysis. We omit the consideration of the fading effects, however they may be taken into account by introducing an appropriate fading margin or an additional random variable directly into the path gain function [34].

3) Resource allocation: Within our LTE model, the resource allocation corresponds to the *round-robin* scheduling discipline and implies equal resource sharing between all supported connections at a particular SBS, whereas WiFi users share resources equally by design of the randomized channel-access protocol. Therefore, each of n running connections of a certain SBS is allocated an equal portion of the total time-frequency resource, and the achievable data rate is thus

$\tilde{r}_i = \frac{\delta_{w/l}}{n} r_i$, where r_i is the instantaneous data rate as well as δ_w and δ_l are the resources available at this SBS for WiFi and LTE, respectively. A connection admitted with the rate $\tilde{r}_i$ is served without interruption until when it successfully leaves the system.Our choice is dictated by the absence of preliminary knowledge on the user priority, since the appropriate selection of weights constitutes a standalone research problem, which is left out of scope of this paper. Nonetheless, the following mathematical abstraction may be easily extended to the case of another scheduler.

In turn, WiFi operation assumes probabilistic time-division access between all of the users, which, over longer periods of time, is equivalent to "stochastic" round robin. Based on the above considerations, we employ round-robin scheduling in the below analysis for both LTE and WiFi systems.

B. *Power control and interference coordination*

Due to the cornerstone importance of accurate interference characterization in ultra-dense HetNets [35], we discuss our related assumptions separately in what follows.

1) We assume that the power control function for all LTE transmissions is locked to a common target SINR η [36], which is fixed across the network. If the target SINR cannot be achieved, the user transmits at its maximum allowed power $p_{\max}$, as to provide the closest possible approximation to its QoS requirement. The alternative would be to not serve the user at all, which we consider more harmful.
2) The interference for the cellular communication is maintained at a fixed level in the network by means of coordinated inter-cell scheduling [37], such that if a transmission scheduled on a certain resource at the required power would cause noticeable interference to another cell, appropriate scheduling is employed to *remove* the affected physical resources from usage there (see Figure 3).
3) The transmit power for WiFi communication is fixed at its maximum level $p_{\max}$, which, however, is different from what is used for LTE.
4) The interference for the WiFi transmissions is constrained by the DCF function with RTS/CTS handshake. Accordingly, each WiFi link reserves areas around both endpoints (for data and acknowledgments) and may be activated if neither transmitter nor receiver falls into thus reserved areas of other active WiFi transmissions. As a result, no transmissions interfering above a certain threshold may occur.

As mentioned above, to mitigate inter-cell interference in an ultra-dense deployment, the LTE SBSs may exploit advanced scheduling mechanisms [38]. However, as the size of the deployment grows, the simpler options become more practical due to better scalability. In the simplest case, if a user is located at the intersection of two SBS coverage areas and transmits to its nearest station using a set of resources X, then the neighboring SBS excludes the same set of resources X from its own pool, as illustrated in Figure 3, and the other-cell users do not cause harmful interference. Naturally, if some of the resources in X have already been excluded previously due to interaction with another small cell, no further actions will be applied to them. The share of the excluded resource is assumed to be equal $1-\nu$, where ν is the *resource reuse coefficient* depending on the deployment, i.e., on the SBS density and coverage radius. Such coordination is, in nature, suboptimal, but guarantees that no harmful interference is caused with minimal complexity, and thus may be a very practical option.

A somewhat similar procedure happens in WiFi, except that *fixed-size* areas are excluded *irrespective of the required power*, thus resulting in a random non-interfering set of links at any given moment of time. Hence, WiFi does not benefit from densification in the same way as cellular systems do.

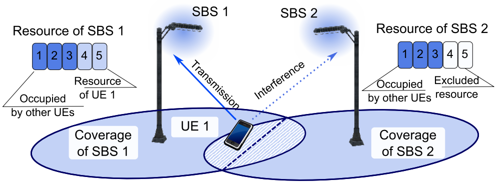


Fig. 3. Interference coordination in LTE small cells.

C. *Network selection schemes*

We benchmark two alternative schemes of integrated small cell operation, which are based on different considerations behind LTE/WiFi network selection. Our *baseline* mechanism assumes that the HetNet attempts to offload a newly arrived user onto unlicensed WiFi spectrum *first*, and only if SNR is less than some threshold η_0 it establishes an LTE connection. Such behavior is an automated version of what a human user would do today due to the connectivity costs and corresponds to the current WiFi-preferred RAT selection schemes discussed in the standards [19]. Notably, the choice of network selection scheme is not necessarily restricted to the one above, i.e. any similar algorithm may be incorporated into our model similarly.

A more advanced *simultaneous operation* scheme assumes concurrent use of both LTE and WiFi radios, whenever appropriate. It is important to note that simultaneous operation does not necessarily result in the best performance overall, but is instead provided here as an example of what future 5G architecture could do. Both schemes are rigorously analyzed below, and we offer tight approximations for the stationary distribution of users in the system and, therefore, the number of users averaged across time and space, as well as the blocking probability and the mean transmission time.

IV. Analyzing Dynamics of Integrated HetNets

A. *Section structure and general statements*

In this section, we focus on one typical (tagged) small cell in the area of interest. We analytically model the performance dynamics of the associated users by means of applying queuing theory across a number of operational scenarios. To this end, and for the sake of better readability, we begin with a preliminary discussion on the two considered underlying technologies (LTE and WiFi) and describe the corresponding system processes in terms of our mathematical abstraction. We then first outline our analysis by example of a classical PPP user distribution, whereas an important extension to the more practical case of mixed uniform and clustered users is accomplished towards the end of the section.

Our below analysis implements the following structure, where important technical derivations are detailed in the Appendix:

1) Preliminary derivations describing the evolution of the number of active users on each radio technology individually:
    - description of system abstraction and of our general mathematical approach based on aggregation of Markov processes,
    - underlying computations for individual WiFi and LTE features (including our proposed concept of ”phantom” users).

    Preliminary analysis is then used to produce solutions for different RAT selection approaches.
2) The baseline model of WiFi-preferred network selection (leading to separable data transmission processes).
3) The advanced model based on simultaneous use of LTE and WiFi RATs by a user.
4) Practical extension for mixed user deployments (both PPP and clustered arrival processes).

In what follows, we introduce our mathematical model based on the above assumptions in order to derive the steady-state distribution of the number of users in service. This stationary distribution, in turn, allows us to calculate stationary metrics of interest, averaged by the considered spatial distributions.

The user-SBS association is distance-based, when UE communicates to its closest SBS, which in homogeneous environment is equivalent to SBS preference basing on the perceived RSSI level [39]. The geometric locus of points corresponding to the location of users, which are associated with the same SBS, constitutes a convex polygon (or, a *Voronoi cell* [40]), that is further referred to as the SBS *service area*. In other words, service area defines a set of points on the plane that are located closer to the considered SBS than to any other. Below, our tagged SBS is assumed to have a circular coverage area with radius R, the effective service area of which is a polygon inside the coverage area as described above.

B. Moldeling key system dynamics

Here, for both underlying technologies, we design analytical expressions for the stationary state distribution for $N(t)$ and its derivatives. In our tagged small cell, the evolution of active users may be described by a Markov processes as follows. Let us consider this typical cell at the embedded points of user arrivals and departures. We characterize the *system state* by a set of users in service together with their individual parameters (such as rate, distance, etc.) defined by the distribution of user locations. Hence, the future process evolution is entirely determined by the current set of distances.

Therefore, we may write the system state $\tilde{S}(t)$ at the moment t as:

$$\tilde{S}(t) = (n; \xi_1, ..., \xi_n)\,,$$

where ξ_i is a distance to the user i, and n is the current number of users associated with the SBS. The system state thus includes information on the total number of users currently in service, as well as their exact locations.

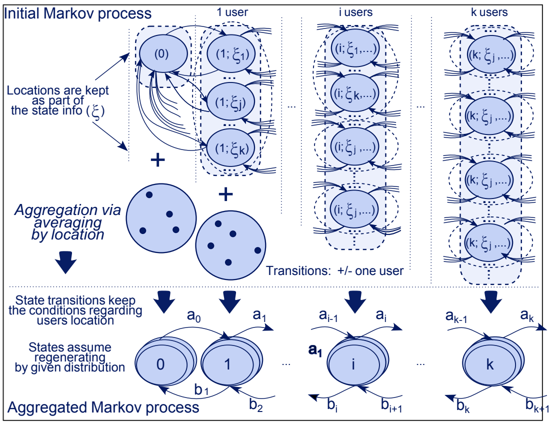

Initial Markov process
1 user
i users
k users
(0)
(1; ξ1)
(1; ξj)
(1; ξk)
(i; ξ1,...)
(i; ξk,...)
(i; ξj ,...)
(i; ξj,...)
(k; ξj ,...)
Locations are kept as part of the state info (ξ)
+
Aggregation via averaging by location
Transitions: +/- one user
State transitions keep the conditions regarding users location
States assume regenerating by given distribution
a0
a1
ai-1
ai
ak-1
ak
a1
0
1
i
k
b1
b2
bi
bi+1
bk
bk+1
Aggregated Markov process

Fig. 4. Simplification of the Markov chain.

The process $S(t)$ (see the upper part of Figure 4) is rather complex to analyze. Therefore, we apply the *state aggregation* technique [41], which allows us to provide an elegant approximation for the ”average” system behavior. In particular, we aggregate the states $\tilde{S}(t)$ into the new state $S(t) = N(t)$ (see the lower part of Figure 4), where $N(t)$ is the current total number of active users, which abstracts away user locations and associations. That means that the state $N(t)$ incorporates all the possible states of the initial process $(N(t); \xi_1, ..., \xi_n)$. The simplified system constitutes a birth-death process with the transition rates a_i and b_i from the state $N(t) = i$ to the states $N(t) = i + 1$ and $N(t) = i - 1$, respectively. The transitions from a state to a state themselves should include the implicit information on the locations’ distribution and constitute a conditional probability. Importantly, in Section V we conduct thorough verification of the feasibility of such aggregation and demonstrate very tight convergence between our analytical and simulation results.

In summary, the process $N(t)$ has been averaged spatially to preserve the memoryless property and hence may be analyzed now by employing methods from queuing theory.

Proposition 1. *The stationary distribution $\pi = \{\pi_n\}_{i=0}^{\infty}$ of the aggregated process $N(t)$ may be obtained as:*

$$\pi_n = \pi_0 \prod_{i=1}^{n} \frac{a_{i-1}}{b_i}, \tag{2}$$

where π_0 may be calculated from the normalization condition.

The above proposition enables calculation of the relevant system metrics in the stationary state, such as the average number of users in the system or their transmission times.

However, the expression (2), in turn, requires calculation of the transitions a_i and b_i between the neighboring states. The transition rates a_i for the aggregated system are determined by the average number of users per a time unit and per service area, which may be described as:

$$a_i = \lambda \frac{S_s}{D^2}, \tag{3}$$

where λ is the arrival rate per the area of interest D^2 and S_s is the area, whereto the users of the tagged SBS arrive.

Far less obvious are the transition probabilities from the state i to state $i-1$, which are determined by the distribution of the actual service rate, i.e., the file size and the instantaneous data rate given by the Shannon's formula. For example, if i active users would have equal data rate of $\frac{r}{n}$, then the transition $(i) \rightarrow (i-1)$ would be $b_i = i\frac{1}{\theta i/r} = \frac{r}{\theta}$. In case of different instantaneous rates, $r_i \neq const$, the transitions of the aggregated process are defined by the following proposition.

Proposition 2. *The transition rate b_i for the aggregated process $N(t)$ does not depend on state for an equal resource allocation procedure and may be obtained as:*

$$b_i = \tilde{\delta}_{w/l}\left[\int\limits_0^{\infty} x\left(\int\limits_{r_R}^{r_{\lim}} \tfrac{r}{\theta} e^{-\frac{r}{\theta}x} f_r(r)dr - C\tfrac{r_{\lim}}{\theta} e^{-\frac{r_{\lim}x}{\theta}}\right)dx\right]^{-1}, \tag{4}$$

where $C = \int\limits_{r_R}^{r_{\lim}} f_r(r)dr$ and $\tilde{\delta}_{w/l} \leq \delta_{w/l} \leq 1$ is a parameter corresponding to the available resource share which is calculated later, θ is the average file size, $f_r(r), r \in [r_R, r_{\lim}]$ is the instantaneous rate distribution, r_R is the minimum instantaneous rate (i.e., that at the border of the coverage area, $w\log(1+\frac{pG}{N_0 R^k})$), and r_{lim} is an upper bound on the achievable rate.

Proof. The proof is given in the Appendix. ■

The distribution of rates $f_r(r)$ for both communication technologies may further be derived using Shannon's formula as:

$$f_r(r) = f_d\left(d(r)\right)|d'_r|,$$

where f_d is the distribution of distances between the SBS and its users; $d(r)$ and its derivative d'_r are given by:

$$d = \left(\frac{pG}{(e^{\frac{r}{w}}-1)N_0}\right)^{\frac{1}{\kappa}}, \quad |d'_r| = \frac{1}{\kappa w}e^{\frac{r}{w}}\left(\frac{pG}{N_0}\right)^{\frac{1}{\kappa}}(e^{\frac{r}{w}}-1)^{-\frac{1}{\kappa}-1},$$

where $p_{\max}$ is replaced by p for the sake of brevity, *SINR* may be replaced by *SNR* due to the assumed interference coordination process (see previous section). For the sake of clarity, we closely connect the rate limitation $r_{\lim}$ with the corresponding distance $d_{\lim}$ and virtually place users for which $w\log(1+\mathit{SNR}) \geq r_{\lim}$ at the same distance $d_{\lim} = d(r_{\lim})$. Further, we refer to the distribution $f_d(d)$, $d \in [d_{\lim}, R]$ as to that describing area or rate limitations.

The above expressions comprehensively describe the steady-state distribution and, hence, define the average number of active users, their average time spent in service, and the average effective rate per a served user:

$$E_t[n] = \sum_{i=0}^{\infty} i\pi_i, \quad E_t[T] = \frac{E[n]D^2}{\lambda S_s}, \quad E_t[r] = E_s[r]\frac{\sum\limits_{i=0}^{\infty}\frac{1}{i}\pi_i}{1-\pi_0}, \tag{5}$$

where E_t/E_s denote averaging across time and space, respectively, and the spatially-averaged instantaneous rate may be obtained as:

$$E_s[r] = \int\limits_{r_R}^{r_{\lim}} r f_d(d(r))|d'_r(r)|dr + r_{\lim}F_d(d_{\lim}).$$

It implies, in turn, that for the state-independent a_i and b_i, we may rearrange the above expressions as:

$$E_t[n] = \frac{a_i/b_i}{1-a_i/b_i}, E_t[r] = E_s[r]\frac{\log\frac{1}{1-a_i/b_i}}{a_i/b_i}, \tag{6}$$

where $a_i/b_i < 1$ is the system load.

The rest of this text provides additional information regarding the calculation of coefficients a_i and b_i, continued by considering a *tagged* random small cell polygon.

C. Underlying WiFi-related derivations

Below, we concentrate on WiFi operations, with specific features illustrated in Figure 5 at first within the area interest to determine the available resource per cell, and then in the tagged small cell. We argue that the behavior of WiFi can be expressed exactly in the form of the above-described general process, if the available WiFi resources in the tagged cell are known. To model WiFi DCF, we assume that every active WiFi link spawns two intercepting circles, defined by the individual coverage areas of the transmitter and the receiver (as both data and acknowledgments are sent).

During the operation of such WiFi link, no other user or SBS can transmit within thus excluded area. Geometrically, this translates into two overlapping circles of smaller radius $R/2$ composing the *link coverage area*, such that the distinctive link coverages cannot intersect. Further, we characterize the average distance to the serving SBS as the forming parameter. For the sake of tractability, we then approximate the link coverage area by the circle of area $\overline{S}$ (calculated in Appendix).

Due to our assumption that all the SBSs are controlled by the operator, the respective WiFi transmissions can be coordinated according to a particular scheduling (puncturing) procedure in addition to purely random access [42]. The latter implies that the link coverage areas, in the limit, correspond to the dense packing within the area of interest. Then, the number of simultaneously activated links may be derived basing on the dense packing with the coefficient $\frac{\pi}{2\sqrt{3}}$.

Proposition 3. *The share of available resource per a single SBS in the ultra-dense deployment may be obtained as:*

$$\tilde{\delta}_w = \delta_w \frac{\pi}{2\sqrt{3}\mu\overline{S}}, \tag{7}$$

where δ_w is the actually available resource, excluding all overheads and signaling, and the area of link coverage S may be calculated using (33) from Appendix as $\frac{1}{2}\pi R_2^2 - 2\overline{s}$.

Proof. The average number of SBSs per one simultaneously available resource is naturally the product of their total number

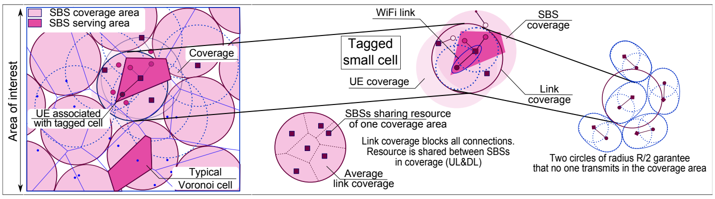

SBS coverage area
SBS serving area
Area of interest
Coverage
UE associated with tagged cell
Typical Voronoi cell
WiFi link
Tagged small cell
UE coverage
SBS coverage
Link coverage
SBSs sharing resource of one coverage area
Link coverage blocks all connections. Resource is shared between SBSs in coverage (UL&DL)
Average link coverage
Two circles of radius R/2 garantee that no one transmits in the coverage area

Fig. 5. Voronoi diagram and illustration of WiFi operation: area of interest (left) and separated small cell with the link and SBS coverage area (right).

in the area of interest μD^2 and the number of link coverage circles as:

$$\frac{\mu D^2}{\frac{\pi}{2\sqrt{3}} D^2/\overline{S}} = \frac{2\sqrt{3}\mu}{\pi}\overline{S}, \tag{8}$$

which are sharing the total WiFi resource in time. The share of the available resource per an SBS is reciprocal to the expression (8) multiplied by the given parameter δ_w. ■

The distribution of the distances between the SBS and the uniformly-distributed users in its service area (polygon) leads to the well-known distribution, which we additionally restrict by the coverage radius R and the rate-limiting distance d_{lim}:

$$f_d(d) = \begin{cases} C_1 \cdot 2\pi\mu d e^{-\pi\mu d^2}, d \in (d_{\lim}, R], \\ C_1 \cdot (1 - e^{-\pi\mu d_{\lim}^2}), d = d_{\lim}, \end{cases} \tag{9}$$

where $C_1 = \left(1 - e^{-\pi\mu R^2}\right)^{-1}$ is the normalization coefficient. Basing on (9), the rate distribution may then be produced by the expression:

$$f_r(r) = f_d\left(d(r)\right)|d'_r|, \tag{10}$$

where $d(r)$ and $d'(r)$ have been given above.

In summary, the transitions b_i may be obtained from the rate distribution $f_r(r)$. The parameters a_i depend on S_s, which here is the service area with the average relative area $1/\mu$. Therefore, for the average area, we have:

$$a_i = \frac{\lambda}{\mu D^2}. \tag{11}$$

D. Underlying LTE-related derivations

Below, we focus on LTE operation as illustrated in Figure 6 and calculate the stationary distribution for $N(t)$. It defines the state transitions of the underlying Markov chain by focusing on the tagged Voronoi cell by analogy to the previous WiFi-related discussion. In addition, all of the SBS users, which belong to the coverage area, but have associations with other (neighboring) SBS, may cause interference. We assume that these users are known to the tagged SBS and it excludes their occupied resources from its own resource pool.

We treat such a foreign-cell user equivalently to admitting a new *phantom* user on our tagged cell. The phantom user has normal traffic pattern, zero effective throughput, and the data rate corresponding to the distance to its serving SBS. Geometrically, it means that the users located within the coverage, but outside the service polygon, are "mirrored" inside the polygon against its edge. Further, and by contrast to WiFi operation, there is no need to explicitly split system resource between several SBSs. In fact, one unit of resource is divided between the users served by this SBS and its phantom users. The following proposition quantifies the share of the available resource per an SBS.

Proposition 4. *The share of the available resource per a single SBS is determined by the value of $\tilde{\delta}_l = \delta_l$. However, this resource is shared by the served users as well as the phantom users.*

Further, let us obtain the distribution of user distances inside the coverage area within the average polygon and outside of it. The distribution of distances between the SBS and all the users, for which this SBS is the closest, is identical to that for WiFi operation above, see equation (9). We note that due to our assumed SNR-target power control, the maximum data rate $r_{tar} \leq r_{\lim}$ is defined by the target SNR η and remains fixed. We also emphasize that r_{tar} as well as the distance d_{tar} correspond to the target SNR η and for simplicity all of the users with the target SNR match the distance d_{tar}. Consequently, the distribution of distances for both groups within the average polygon and outside of it is given as:

$$f_u(d) = \begin{cases} C_1 \cdot 2\pi\mu d e^{-\pi\mu d^2}, d \in (d_{tar}, R], \\ C_1 \cdot (1 - e^{-\pi\mu d_{tar}^2}), d = d_{tar}, \end{cases} \tag{12}$$

where $C_1 = \left(1 - e^{-\pi\mu R^2}\right)^{-1}$.

The data rate distribution $f_r(r)$ obtained with the expression (10) delivers the transitions b_i, whereas the transitions a_i depend on the area S_s. Due to the presence of phantom users, S_s is characterized by the area of coverage πR^2:

$$a_i = \lambda\frac{\pi R^2}{D^2}. \tag{13}$$

In sharp contrast to WiFi-related derivations, the LTE system evolution has to be represented as a two-dimensional process with the state space $\{n_l, n_p, n_l = \overline{0,\infty}, n_p = \overline{0,\infty}\}$, where n_l and n_p are the numbers of served and phantom users, respectively. This process (see Figure 7) can be described by a system of Kolmogorov's balance equations. For an arbitrary state (i,j), such that $0 < i < \infty$, $0 < j < \infty$, we have;

$$\begin{aligned} p_{i,j}\left(aq_1 + a(1-q_1)(1-\nu) + \tfrac{i}{i+j}x + \tfrac{j}{i+j}x\right) \\ = p_{i-1,j}aq_1 + p_{i,j-1}a(1-q_1)(1-\nu) \\ + p_{i+1,j}\tfrac{i+1}{i+j+1} + p_{i,j+1}\tfrac{j+1}{i+j+1}, \end{aligned} \tag{14}$$

where ν is a share of the resource, which is not yet excluded. In other words, this is a share of proximate neighboring users,

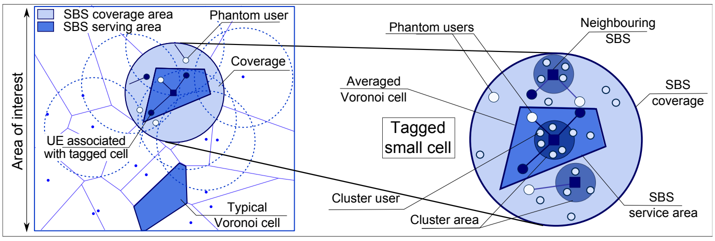


Fig. 6. Voronoi diagram and illustration of LTE operation: area of interest (left) and separated small cell with SBS coverage area and phantom users (right).

which do not become additional phantom users by occupying already blank SBS resource. The data rate for both served and phantom users is proportional to the share $\frac{1}{n_l+n_p}$, since they consume a common resource.

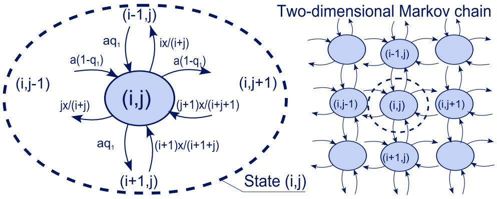


Fig. 7. A particular state of the two-dimensional Markov chain for LTE.

In light of the above, our sought solution may be obtained in the following form:

$$\begin{aligned} p_{i,j} &= p_{0,0}\frac{a^i q_1^i}{x^i}\frac{\prod_{m=1}^{i+j} m}{\prod_{m=1}^{i} m \prod_{k=1}^{i} k}\frac{a^j(1-q_1)^j(1-\nu)^j}{x^j} \\ &= p_{0,0}\frac{(i+j)!}{i!j!}\frac{a^i q_1^i}{x^i}\frac{a^j(1-q_1)^j(1-\nu)^j}{x^j}, \end{aligned} \tag{15}$$

where $p_{0,0}$ is given by appropriate normalization $\sum p_{i,j} = 1$ and q_1 is the probability to land within the coverage area, but outside the service area. The parameter a corresponds to the total arrival rate a_i and is determined by the expression (13), whereas x delivers the LTE service rate of i users sharing a unit resource only between themselves. It is given by the equation (30) in Appendix as $x = \frac{\tilde{\delta}_l}{E\left[\frac{s_j}{r_j}\right]}$.

Therefore, the average number of served users may be obtained as:

$$\begin{aligned} &\sum_{i=1}^{\infty}\sum_{j=0}^{\infty} i p_{i,j} = p_{0,0}\sum_{i=1}^{\infty} i\frac{a^i q_1^i}{x^i}\sum_{j=0}^{\infty}\frac{(i+j)!}{i!j!}\frac{a^j(1-q_1)^j(1-\nu)^j}{x^j} \\ &= p_{0,0}\frac{aq_1}{x}\sum_{n=0}^{\infty} n\frac{(1-q_1)^n(1-\nu)^n a^n}{x^n}\sum_{i=0}^{n}\frac{n!}{i!(n-i)!}\frac{q_1^i}{(1-q_1)^i(1-\nu)^i} \\ &\quad= p_{0,0}\frac{aq_1}{x}\sum_{n=0}^{\infty} n\frac{(1-q_1)^n(1-\nu)^n a^n}{x^n}\frac{((1-q_1)(1-\nu)+q_1)^n}{(1-q_1)^n(1-\nu)^n} \\ &= \frac{p_{0,0}\cdot aq_1}{x}\sum_{n=0}^{\infty} n\frac{a^n((1-q_1)(1-\nu)+q_1)^n}{x^n} = \frac{p_{0,0}\cdot aq_1}{x\left(1-a\frac{1-\nu+q_1\nu}{x}\right)}. \end{aligned} \tag{16}$$

The average number of phantom users as well as the total number of users in the system may be derived in a similar way. However, when LTE and WiFi operations are not independent (as is in the case of simultaneous transmission on both radio interfaces), we could further simplify the process and replace the operation within the coverage area by that within the service area for LTE. By that, we unify both schemes and can always refer to the service area.

Theorem 1. *The two-dimensional process accounting for the served and the phantom users is equivalent to a one-dimensional process with the modified transitions aq_1 and $x - a(1-q_1)(1-\nu)$.*

Proof. Let us consider a one-dimensional birth-death process with the state space $\{n_l, n_l = \overline{0,\infty}\}$ denoting the number of LTE users served by the tagged SBS. The arrival rate of the served users is proportional to the average service area and hence equals aq_1, which constitutes the transitions from the state $i-1$ to the state i. Then, the steady-state distribution of the considered process may be obtained as:

$$\pi_n = \pi_0 \prod_{i=1}^{n}\frac{aq_1}{b_i}, \tag{17}$$

where b_i are the transitions between the states i and $i-1$, while π_0 may be obtained through the normalization condition.

In order to derive the transitions b_i, we recalculate the steady-state distribution basing on the original two-dimensional process:

$$\begin{aligned} \pi_i &= \left(1-\tfrac{a}{x}\right)\textstyle\sum_{j=0}^{\infty}\frac{(i+j)!}{i!j!}\frac{a^i q_1^i}{x^i}\frac{a^j(1-q_1)^j(1-\nu)^j}{x^j} \\ &= \left(1-\tfrac{a}{x}\right)\textstyle\sum_{n=i}^{\infty}\left(\frac{q_1}{(1-q_1)(1-\nu)}\right)^i\frac{n!}{i!(n-i)!}\left(\frac{a^n(1-q_1)^n(1-\nu)^n}{x^n}\right) \\ &= \left(1-\tfrac{a}{x}\right)\left(\frac{q_1}{(1-q_1)(1-\nu)}\right)^i\frac{\left(\frac{a(1-q_1)(1-\nu)}{x}\right)^i}{\left(1-\frac{a(1-q_1)(1-\nu)}{x}\right)^{i+1}} \\ &= \frac{x-a}{x-a(1-q_1)(1-\nu)}\frac{(aq_1)^i}{(x-a(1-q_1)(1-\nu))^i}. \end{aligned} \tag{18}$$

Comparing the expressions (17) and (18), we conclude that b_i may be chosen as $x - a(1-q_1)(1-\nu)$ to replace the two-dimensional process with the simpler one, which completes the proof. ∎

In summary, we confirm that in terms of *averages* the above joint process is equivalent to the process with the proportionally-reduced arrival and service rates. It leads, for example, to the expression for $E[n_l]$ that has been obtained previously in (16). In the following, we investigate the evolution of users within one service area.

E. Baseline scheme: preferential use of WiFi technology

Here, we collect the above underlying results so as to study a baseline integrated LTE-WiFi system. As a simple representative of the class of existing multi-radio network selection algorithms, we adopt the method of preferential use of one particular RAT. Without loss of generality, we concentrate on the WiFi-preferred scheme, which implies that a user switches to WiFi whenever the perceived SNR exceeds a particular threshold η_0, or otherwise transmits on LTE [19]. This approach corresponds to the operator's desire to automate the conventional network selection routine already performed by most users today, as well as reasonably balance the loading across both unlicensed and licensed bands.

The offloading threshold η_0 defines an internal zone of radius $d_0 = \min\left[\left(\frac{p_w G}{\eta_0 N_w}\right)^{1/\kappa_w}, R_w\right]$ of the service area, where WiFi users are served exclusively. In what follows, the notation "$\cdot_w$" refers to the use of WiFi technology, with the corresponding maximum transmit power denoted, for convenience, as p_w. The probability for a user to land into WiFi service area may be calculated as:

$$q = \Pr\{d < d_0\} = C_1(1 - e^{-\pi\mu d_0^2}), C_1 = \frac{1}{1-e^{-\pi\mu R^2}}. \quad (19)$$

The process of our baseline system evolution within one tagged small cell incorporates the states (n_w, n_l, n_p), where n_w, n_l, and n_p are the numbers of WiFi users, LTE users, as well as phantom LTE users from the neighboring small cells. We note that in this case, WiFi and LTE evolutions constitute independent processes. Therefore, we decompose the overall system operation into WiFi and LTE dynamics, respectively.

On the one hand, the performance dynamics of WiFi is characterized by the state space $\{n, n = \overline{0, \infty}\}$. The corresponding distribution of distances should then evolve according to the maximum link length d_0, such that this distance distribution is defined as in (9). However, R is replaced here with d_0, while $C_1 = (1 - e^{-\pi\mu d_0^2})$. On the other hand, the distance distribution $f_d(d)$ modifies the average area of the communication link:

$$\begin{aligned} \overline{S} &= 2\pi\left(\tfrac{d_0}{2}\right)^2 - \int_0^{2\frac{d_0}{2}} s(x) f_d(x) dx \\ &= \tfrac{1}{2}\pi d_0^2 - C_1 \int_0^{d_0} s(x) \cdot 2\pi\mu x e^{-\pi\mu x^2} dx, \end{aligned} \quad (20)$$

where $s(x)$ is given in Appendix by the equation (32). Accounting for thus thinned Poisson arrivals, the parameter a_i transforms into:

$$a_i = q\frac{\lambda}{\mu D^2}. \quad (21)$$

Similarly, LTE performance dynamics is characterized by the distance distribution, which is restricted by d_0 as:

$$f_d(d) = \begin{cases} C_1 \cdot 2\pi\mu d e^{-\pi\mu d^2}, d \in (d_m, R], \\ \max\left(C_1(e^{-\pi\mu d_0^2} - e^{-\pi\mu d_{tar}^2}), 0\right), d = d_m, \end{cases} \quad (22)$$

where $d_m = \max(d_{tar}, d_0)$, $C_1 = \left(e^{-\pi\mu d_0^2} - e^{-\pi\mu R^2}\right)^{-1}$.

According to Theorem above, we may substitute the two-dimensional process with the equivalent one-dimensional process having the transitions $a_i = q_1 a$, where a is the overall arrival rate on the LTE system:

$$a = (1 - q)\lambda\frac{\pi R^2}{D^2}, a_i = (1 - q)\lambda\frac{1/\mu}{D^2}. \quad (23)$$

The transitions b_i have to be chosen such that $x - a(1 - q_1)(1 - \nu)$ to replace the two-dimensional process with a simpler one. Importantly, the parameter q_1 has to also be recalculated according to the distance distribution, as well as the expression $E\left[\frac{s_j}{r_j}\right]$ depending on the rate distribution. The latter, in turn, depends on the distance distribution, which is restricted by $d > d_0$.

F. Advanced scheme: simultaneous LTE and WiFi operation

Here, we arrive at a more advanced multi-radio communication scheme, which assumes simultaneous use of both LTE and WiFi radios as enabled by the proposed architecture. This scheme is made available by the recent progress in co-located small cell technology and the rationale behind it has been outlined in Section II. Based on the underlying derivations summarized previously in this section, we may now mathematically characterize the system in question and obtain the average parameters of interest.

We further assume that the evolution of all neighboring small cells is stochastically equivalent to the performance dynamics of the tagged cell. Correspondingly, the phantom users leave the system at the same rate as the served users. This rate is defined by the aggregate of the corresponding WiFi and LTE rates.

In order to mathematically describe our system of interest, we need to analyze the two-dimensional process detailed in Figure 7, while redefining its underlying parameters. To this end, we adopt our proposed move with the process replacement. In other words, we replace the process with the states $\{n, n_p, n = \overline{0, \infty},\ n_p = \overline{0, \infty}\}$, where n is the number of the served users and n_p is the number of the phantom users, by the process $\{n, n = \overline{0, \infty}\}$. Therefore, the transition parameters corresponding to the service area may be obtained as:

$$a_i = \lambda\frac{1/\mu}{D^2}, b_i = x - a(1 - q_1)(1 - \nu), \quad (24)$$

where x is given by:

$$x = \frac{i}{E\left[\frac{s_j}{\frac{\tilde{\delta}_w r_w}{i} + \frac{\tilde{\delta}_l r_l}{i}}\right]} = \frac{1}{E\left[\frac{s_j}{\tilde{\delta}_w r_w + \tilde{\delta}_l r_l}\right]}. \quad (25)$$

The latter expression is recalculated basing on the distance distribution defined by:

$$f_d(d) = \begin{cases} C_1 \cdot 2\pi\mu d e^{-\pi\mu d^2}, d \in (d_{\lim}, R], \\ \max\left(C_1 \cdot (1 - e^{-\pi\mu d_{\lim}^2}), 0\right), d = d_{\lim}, \end{cases} \quad (26)$$

where $C_1 = \left(1 - e^{-\pi\mu R^2}\right)^{-1}$.

Due to the operation of the WiFi-preferred scheme for all SBSs in our network, the averaging approach to recalculating the share q_1 leads to the same value as discussed previously:

$$q_1 = \frac{S_{lte}}{S_{all}} = \frac{1/\mu - S_{wifi}}{\pi R_{lte}^2 - (\mu\pi R_{lte}^2)S_{wifi}} = \frac{1}{\mu\pi R_{lte}^2}. \quad (27)$$

G. Important practical extension

As our final step, we comment on how it is possible to incorporate a mixture of PPP and cluster processes of user arrivals into our methodology. We remind that the users clustered in the circle r are added to the distribution $f_d(d)$ considered above with the density $f_c(d) = \frac{2d}{r^2}$ and the coefficient $(1-\epsilon)$.

Proposition 5. *Considering the mixture of PPP/cluster processes, the resulting distribution of distances is defined as a piecewise function below:*

$$f_d(d)=\begin{cases} C_1\epsilon\cdot 2\pi\mu d e^{-\pi\mu d^2}, r<d<R, \\ C_1\epsilon\cdot 2\pi\mu d e^{-\pi\mu d^2}+(1-\epsilon)\frac{2d}{r^2}, d_{\lim}<d\leq r, \\ C_1\epsilon\cdot(1-e^{-\pi\mu d_{\lim}^2})+(1-\epsilon)\frac{d_{\lim}^2}{r^2}, d=d_{\lim}, \end{cases} \tag{28}$$

where the distance $d_{\lim}$ of the maximum rate limitation r_{lim} is taken into consideration, while $C_1 = (1-e^{-\pi\mu R^2})^{-1}$ is a term constraining the distances according to the coverage radius R.

Additionally, our proposed methodology allows derivation of, e.g., blocking probabilities when moving from the state $i-1$ to the state i for both baseline and advanced transmission schemes. This may also easily incorporate the consideration of several extra special cases when transitions depend on both state indexes i and j. This extension might be useful, should an operator decide to deny user admission if its predicted data rate would be below a certain threshold reflecting the minimum desired QoS [43].

V. Performance Characterization and Discussion

A. Supportive simulations for HetNet density limits study

While studying ultra-dense networks analytically, it is crucial to keep track of the complex scheduling algorithms employed by modern HetNets [44]. As those are largely based on engineering intuition rather than mathematical abstraction, they are not always directly tractable, except for simpler cases. To this end, below we assess with simulations how interference coordination affects the degrees of spatial reuse and thus provide comprehensive support for the analytical framework outlined above.

To illustrate the operation of the ultra-dense LTE network, we utilize our *WINTERsim* simulation framework [45], which has been verified in our past publications [42], [19] and allows to model multi-RAT networks on all levels of abstraction. For the purposes of this study, we have constructed a scenario that assumes presence of a UE in each 5x5 m square within the area of interest. The UEs are all scheduled in a round-robin fashion, with a practical interference coordination algorithm employed by all SBSs serving this area. The SBSs have been setup with the ideal receivers operating at the Shannon's limit, while all the UEs follow the SINR-target power control. The SBSs are deployed uniformly within the square area of interest with wrap-around enabled on all edges of the square.

An example of resource allocation for uniformly-deployed users and SBSs is shown in Figure 8. One can clearly see that the surface area is indeed partitioned similarly to a Voronoi diagram, except for the regions where there happens to be no coverage. What is important to note here is that even in the areas where there is a large over-population of cells, we can still provide reliable levels of service. Further, let us observe how system performance in terms of spatial reuse is connected with the network topology. Motivated by our scenario in Figure 8, we propose to connect the system's geometry with the probability ν of re-using a unit of resource that was previously blocked by other cell when performing interference coordination (see Section III above for details).

Characterizing the system geometry, we introduce a new measure, which we term network's specific density. It is defined as $D_s = \frac{4\cdot N_{cells}\cdot S_{cell}}{S_{area}}$ and can be understood as the number of coverage areas that are encountered within a coverage area of an "average" cell (including its own). For example, in the conventional cellular network, the $D_s = 7$ is a typical specific density, which results in the hexagonal grid of cells, and each cell thus having exactly 6 neighbors. Consequently, we name a particular deployment ultra-dense if $D_s > 7$, as this results in a system that is no longer conventional cellular.

We have established that the network's specific density is highly correlated with the reuse probability η, which has been introduced in the course of the above analysis. The corresponding dependence is shown in Figure 9. In our scenario, a reasonable fit has been obtained by a function $\nu = a\cdot x^b + c$ for realistic ultra-dense deployments ($7 < D_s < 100$). Outside of the ultra-dense deployments, the considered parameter D_s may still be useful, as it continues to be tightly connected with the resource reuse probability across all the observed cases.

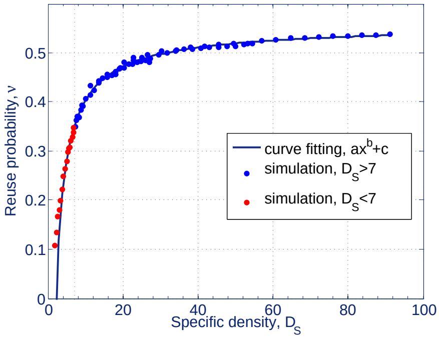


Fig. 9. Reuse probability ν as function of specific density D_s.

We continue with characterizing the performance of WiFi, which drastically differs from that of LTE due to different interference coordination. Unlike LTE, which is centered around an individual SBS, WiFi is a completely flat network with no hierarchy as far as medium access control is concerned. As a result, all links are fairly contending for the channel access, and those experiencing most contention are blocked by backoff (through DCF) and carrier sensing (CCA) functions. As a result, in a dense WiFi network, one can typically observe a limited number of simultaneous transmissions irrespective of the number of active users in the network.

Assessing WiFi operation, we again utilize the *WINTERsim* simulation framework, but this time to construct a series of most likely matchings (i.e., non-interfering sets of links) for all desired AP densities. The matchings represent potential combinations of UEs that could be active at the same time, and are generated naturally as snapshots of regular WiFi

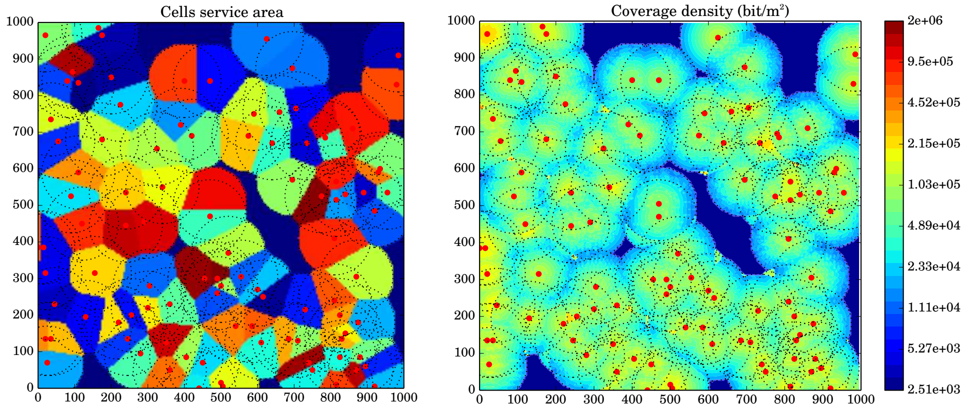


Fig. 8. A realization of HetNet resource allocation for LTE (a snapshot).

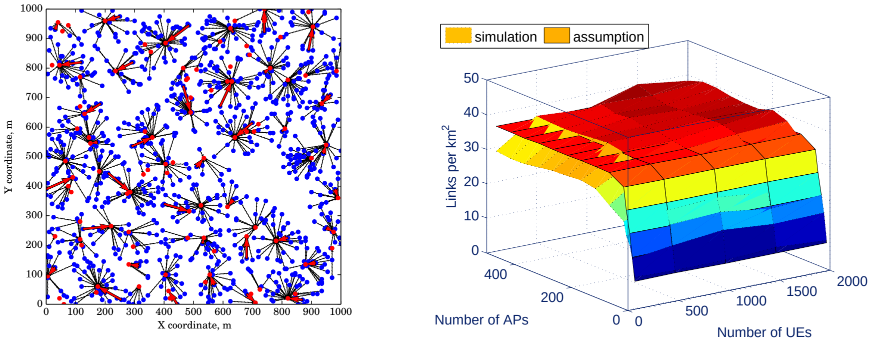


Fig. 10. WiFi matchings example (left) and deployment density effects (right).

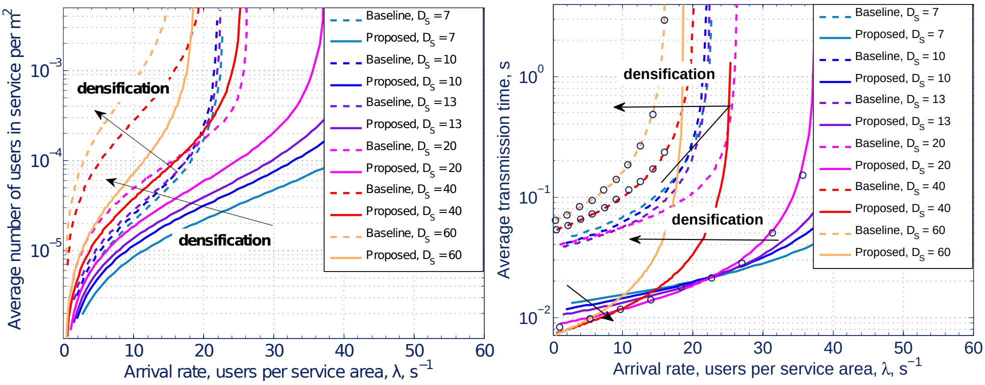


Fig. 11. Average number of users per m^2 (left) and transmission time (right).

operation. An example of how such matching looks like is offered in Figure 10 (left). What we are most interested in for the purposes of further analysis is the maximum number of links that could be activated simultaneously within the area of interest.

Our results confirm that as long as we have significantly more users than the maximum number of supported links, it does not really matter how many users there are – the system dynamics will not be affected. Conversely, having more SBSs does help, as the links end up being shorter and thus block other links less. Even though WiFi operates at the full power during admission, shorter links are still preferred as confirmed by Figure 10(right).

B. Analytical results of comparative case-study

Here, we rigorously compare the two HetNet operation schemes: the preferential use of WiFi investigated in Section IV.E and the proposed simultaneous use of LTE and WiFi analyzed in Section IV.F. The parameters of the considered scenario are similar to our simulation-based study above and are deemed typical for future ultra-dense small cell deployments in light of ongoing 3GPP discussions.

Let us first consider how our considered integrated HetNet system reacts to various user loads. In Figure 11 (right), circle markers correspond to our simulation results, which selectively verify the obtained analytical dependencies (solid curves). The figure suggests a set of dependencies corresponding to both the baseline algorithm (dotted) and the proposed simultaneous operation scheme (solid), which shift according to the arrows with increasing network densification. As Figure 11 suggests, the network has a very notable response to overloads. Essentially, the moment we reach the overload intensities, the transmission times grow exponentially, along with the number of backlogged users. Interestingly enough, the point where the system hits overload is sometimes inversely proportional to the deployment density. In essence, providing more access points than necessary may have a negative effect on network capacity.

To illustrate this important effect better, let us study what happens when each user has a small cell of its own. With a fixed SBS coverage area, the lion's share of the resources will become allocated to phantom users, thus decreasing the amount of resource actually available to a particular tagged user. On the other hand, this user's SINR will be exceptionally high. The practical limitation, however, is that the UE can only make use of around 25 dB SINR; anything above that is essentially useless. Hence, unless SBS power is reduced appropriately, over-densification may have a visible negative impact on system capacity, which calls for further research in this area.

In Figure 12, we illustrate the discussed effect once again, but under a different angle. One can clearly see that for higher user densities, the proposed transmission scheme with the specific density of 7 enjoys the best performance (which effectively corresponds to "almost" regular lattice layout). On the other hand, when system remains essentially idle, it is still beneficial to have more small cells. As this makes the UE-SBS links shorter, the effect is the greater attainable data rates at lower loads.

Contrarily to how the proposed simultaneous transmission scheme operates, the baseline WiFi-preferred system typically benefits from densification much less: at some point all of the UEs are forced to use WiFi by their RAT selection policy. This indicates, in turn, that whenever a choice of multiple alternative RATs is available, the UE should not be restricted to using either one of those, irrespective of its position relative to the SBS.

VI. Conclusions and prospective trends

Leveraging on the strengths of 3GPP cellular (mobility support, billing of wireless edges, authentication, etc.) and WiFi (local, cost-efficient, and generally available broadband access) technologies, as well as the recent trend of cellular being able to provide rates at par of WiFi systems, tightly integrated WiFi-LTE systems will proliferate in the emerging 5G RAT ecosystem. To aid the design, optimization, and deployment of said dense heterogeneous networks, this paper has significantly advanced the state of the art in the field by proposing vital architectural enablers and providing a rigorous and unprecedented analytical framework.

In more detail, going beyond currently stipulated local IP access, HeNB-GW, and Trusted-WiFi access approaches in 3GPP, we have introduced an Alien Access Gateway (AAGW), which mirrors 3GPP functionalities into a WLAN RAT (such as WiFi) and vice-versa. This, in turn, allows a UE to make maximum use of the available air interfaces **without** the need for separate mobility solutions for non-3GPP access. This brings along significant operational advantages, such as (i) enabling the truly integrated WiFi-LTE HetNet deployment with flexible flow splitting analyzed throughout the paper; (ii) significantly shortening end-to-end delays by removing detours to a very likely remote PDN GW; and (iii) dramatically lowering small cell deployment by removing requirement for direct Internet access to be provided along with operator backhaul.

In terms of the mathematical framework, we were able to capture the spatial randomness of the users' distribution jointly with their uplink data dynamics. To this end, we formulated the problem as a Markov process and then introduced suitable mathematical abstractions via state aggregation. For LTE, we proposed a novel concept of phantom users, which makes interference coordination and scheduling analytically tractable. Consequently, this allowed us to obtain the stationary distribution and transition rate of the aggregated service process, as well as the resulting resource slicings in WiFi, LTE, and joint deployments. We also analyzed the important practical case of mixed PPP-cluster user distributions. The obtained equations

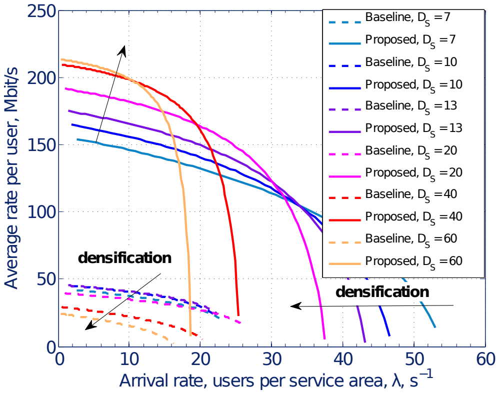


Fig. 12. Average transmission data rate per user.

are mostly in closed form, and thus easy to apply.

The mathematical model has then been verified by means of a 3GPP-compliant simulator, where we contrasted the baseline scenario to the truly integrated HetNet approach with flow splitting. The baseline refers to the case where offload to WiFi is always preferred to LTE service. The performance parameters considered were the average number of users per unit area, the average transmission time of the uplink file transmission, and the rate per user under loaded system conditions. For instance, for a fixed normalized SBS density $D_s = 20$, analysis and simulations proved performance gains of 42%, 300%, and 50% along the above parameters. Improvements were shown to be dependent on the SBS density.

We thus conclude that truly integrated WiFi-LTE HetNets will become the de facto mode of usage within the 5G technology landscape. Further paradigm changes, however, will likely be the norm in the upcoming 3GPP releases:

- **Decoupled Data & Control Channels.** A further enabler of multi-RAT HetNets is the decoupling of data and control channels. This enables an eNodeB to provide control channel coverage to handle data transfer via 3GPP and non-3GPP RATs in a completely equal and transparent fashion.
- **Decoupled Uplink & Downlink.** A trend likely to persist is to completely decouple up and downlinks which was shown to yield significant capacity gains. Again, properly designed HetNets will be core to such developments since the downlink can be provided by a 3GPP-compliant RAT whereas the uplink via a non-3GPP RAT. This enormously increases the flexibility and thus performance of the network.
- **Decoupled Addressing & Forwarding.** Finally, above will be supported by recent developments in decoupling of addressing and forwarding. Notably, flow splits – such as occurring in integrated HetNets – will be key enablers for efficient spectrum use.

We hope that the architecture vision and mathematical tools outlined in this paper will be of substantial use in the discussion and design of emerging 5G systems.

Appendix

A. Transition rate b_i for the aggregated process $N(t)$

Proof. Due to assumed equal resource sharing, when the actual data rate is $\tilde{r}_j = \frac{\delta}{i} r_j$, the transition rate b_i may be obtained as:

$$b_i = \frac{i}{E\left[\frac{s_j}{\tilde{r}_j}\right]} = \frac{1}{E\left[\frac{s_j}{\delta r_j}\right]} = \frac{\delta}{E\left[\frac{s_j}{r_j}\right]}.$$

Further, we calculate the distribution of the random variable $x = s_j/r_j$, where r_j is the maximum data rate and s_j is the exponentially-distributed random file size. Hence, the distribution of x, $x > 0$ may be established as:

$$F_x(x) = \Pr\{\frac{s_j}{r_j} < x\} = \Pr\{s_j < x r_j\}.$$

Therefore, given the independence between r and s and the borders $r \in [0, r_{lim}]$, $s \in [0, \infty]$, we may derive the probability $\Pr\{s < xr\}$ as:

$$\begin{aligned}
\Pr\{s < xr\} &= \int_0^{\infty} \frac{1}{\theta} e^{-\frac{1}{\theta}s} \left(\int_{\max(\frac{s}{x}, r_R)}^{r_{\text{lim}}} f_r(r)dr + F_r(r_{\text{lim}}) \right) ds \\
&= \int_0^{\infty} \frac{1}{\theta} e^{-\frac{1}{\theta}s} \left(1 - \int_{r_R}^{\max(\frac{s}{x}, r_R)} f_r(r)dr \right) ds \\
&= 1 - \int_0^{x r_{\text{lim}}} \frac{1}{\theta} e^{-\frac{1}{\theta}s} \int_{r_R}^{\max(\frac{s}{x}, r_R)} f_r(r) dr ds \\
&= 1 - \int_0^{x r_{\text{lim}}} \frac{1}{\theta} e^{-\frac{1}{\theta}s} F_r(\max(\tfrac{s}{x}, r_R)) ds \\
&= 1 - \int_{x r_R}^{x r_{\text{lim}}} \frac{1}{\theta} e^{-\frac{1}{\theta}s} F_r(\tfrac{s}{x}) ds = 1 - \int_{r_R}^{r_{\text{lim}}} \frac{x}{\theta} e^{-\frac{x}{\theta}r} F_r(r) dr \\
&= 1 + e^{-\frac{x}{\theta} r_{\text{lim}}} \int_{r_R}^{r_{\text{lim}}} f_r(r)dr - \int_{r_R}^{r_{\text{lim}}} e^{-\frac{x}{\theta}r} f_r(r)dr,
\end{aligned}$$

where $r_R = w \log\left(1 + \frac{p_{\max}}{(N_0+I)} \frac{G}{R^{\kappa}}\right)$ is the effective minimum level of the maximum data rate corresponding to the user at the distance R, and the rate distribution function is constructed such that $F_r(r_R) = 0$ and $\max r = r_{\text{lim}}$.

Hence, basing on the probability $\Pr\{s < xr\}$, the probability density function of the random variable x may be described as:

$$f_x(x) = \frac{dF(x)}{dx} = \int_{r_R}^{r_{\text{lim}}} \frac{r}{\theta} e^{-\frac{r}{\theta}x} f_r(r)dr - C \frac{r_{\text{lim}}}{\theta} e^{-\frac{r_{\text{lim}} x}{\theta}}, \quad (29)$$

where $C = \int_{r_R}^{r_{\text{lim}}} f_r(r)dr$.

From the expression (29), we derive the expectation of the random variable s_j/r_j and, hence,:

$$E\left[\frac{s_k}{r_k}\right] = \int_0^{r_{\text{lim}}} x \left(\int_{r_R}^{r_{\text{lim}}} \frac{r}{\theta} e^{-\frac{r}{\theta}x} f_r(r)dr - C \frac{r_{\text{lim}}}{\theta} e^{-\frac{r_{\text{lim}} x}{\theta}} \right) dx, \quad (30)$$

which is fully determined by the rate distribution $f_r(r)$. ■

B. Transition rate b_i for the advanced scheme

The transition parameters corresponding to the service area may be established as:

$$b_i = \frac{i}{E\left[\frac{s_j}{\frac{\tilde{\delta}_w r_w}{i} + \frac{\tilde{\delta}_l r_l}{i}}\right]} = \frac{1}{E\left[\frac{s_j}{\tilde{\delta}_w r_w + \tilde{\delta}_l r_l}\right]}. \quad (31)$$

C. Average union of overlapping coverage areas

Consider two overlapping circles (coverage areas) of radius R_2. The area of half of their intersection, which corresponds to the excluded system resources, is defined as:

$$s(x) = R_2^2 \arccos \frac{x}{2R_2} - \frac{1}{2} x r \sqrt{1 - \frac{x^2}{4R_2^2}}, \quad (32)$$

where x is the distance between the centers. Integrating the above by the distribution of distances $f_d(d)$, we obtain:

$$\overline{s} = \frac{1}{2} \int_0^{2R_2} s(x) f_d(x) dx. \quad (33)$$

The average common area of the two overlapping circular coverage areas may thus be calculated as:

$$2\pi R_2^2 - 2\overline{s} = 2\pi R_2^2 - \int_0^{2R_2} S(x) f_d(x) dx, \quad (34)$$

where R_2 is the coverage radius.

References

[1] M. Dohler, R. Heath, A. Lozano, C. Papadias, and R. Valenzuela, "Is the PHY layer dead?," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 159–165, 2011.

[2] D. Lopez-Perez, I. Guvenc, G. de la Roche, M. Kountouris, T. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 22–30, 2011.

[3] S. Deb, P. Monogioudis, J. Miernik, and J. Seymour, "Algorithms for Enhanced Inter-Cell Interference Coordination (eICIC) in LTE HetNets," *IEEE/ACM Transactions on Networking*, vol. 22, no. 1, pp. 137–150, 2014.

[4] H. ElSawy and E. Hossain, "Two-Tier HetNets with Cognitive Femtocells: Downlink Performance Modeling and Analysis in a Multichannel Environment," *IEEE Transactions on Mobile Computing*, vol. 13, no. 3, pp. 649–663, 2014.

[5] I. Hwang, B. Song, and S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 20–27, 2013.

[6] J. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. Reed, "Femtocells: Past, Present, and Future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, 2012.

[7] R. Baldemair, E. Dahlman, G. Fodor, G. Mildh, S. Parkvall, Y. Selen, H. Tullberg, and K. Balachandran, "Evolving Wireless Communications: Addressing the Challenges and Expectations of the Future," *IEEE Vehicular Technology Magazine*, vol. 8, no. 1, pp. 24–30, 2013.

[8] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, 2014.

[9] S. Lee and K. Huang, "Coverage and Economy of Cellular Networks with Many Base Stations," *IEEE Communications Letters*, vol. 16, no. 7, pp. 1038–1040, 2012.

[10] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?," *IEEE Journal on Selected Areas in Communications*, vol. TBD, pp. 1–17, 2014.

[11] Q. Li, G. Wu, and R. Hu, "Analytical study on network spectrum efficiency of ultra dense networks," in *Proc. of the IEEE Symposium on Personal, Indoor, and Mobile Radio Communications*, pp. 2764–2768, 2013.

[12] A. Gotsis and A. Alexiou, "On Coordinating Ultra-Dense Wireless Access Networks: Optimization Modeling, Algorithms and Insights," *ArXiv preprint:* `http://arxiv.org/abs/1312.1577`, vol. TBD, pp. 1–13, 2013.

[13] A. Gotsis, S. Stefanatos, and A. Alexiou, "Spatial Coordination Strategies in Future Ultra-Dense Wireless Networks," *ArXiv preprint:* `http://arxiv.org/abs/1405.2576`, vol. TBD, pp. 1–9, 2014.

[14] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, and Y. Selen, "5G radio access," *Ericsson Review*, vol. 6, pp. 2–7, 2014.

[15] J. Park, S.-L. Kim, and J. Zander, "Asymptotic Behavior of Ultra-Dense Cellular Networks and Its Economic Impact," *ArXiv preprint:* `http://arxiv.org/abs/1404.1547`, vol. TBD, pp. 1–7, 2014.

[16] M. Bennis, M. Simsek, A. Czylwik, W. Saad, S. Valentin, and M. Debbah, "When Cellular Meets WiFi in Wireless Small Cell Networks," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 44–50, 2013.

[17] S. Singh, H. Dhillon, and J. Andrews, "Offloading in Heterogeneous Networks: Modeling, Analysis, and Design Insights," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2484–2497, 2013.

[18] H. Dhillon, R. Ganti, and J. Andrews, "Load-Aware Modeling and Analysis of Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 4, pp. 1666–1677, 2013.

[19] S. Andreev, M. Gerasimenko, O. Galinina, Y. Koucheryavy, N. Himayat, S.-P. Yeh, and S. Talwar, "Intelligent Access Network Selection in Converged Multi-Radio Heterogeneous Networks," *IEEE Wireless Communications Magazine, to appear*, vol. TBD, pp. 1–10, 2014.

[20] N. Himayat, S.-P. Yeh, A. Panah, S. Talwar, M. Gerasimenko, S. Andreev, and Y. Koucheryavy, "Multi-radio Heterogeneous Networks: Architectures and performance," in *Proc. of the IEEE International Conference on Computing, Networking and Communications*, pp. 252–258, 2014.

[21] J. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, 2013.

[22] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, 2014.

[23] J. Xu, J. Wang, Y. Zhu, Y. Yang, X. Zheng, S. Wang, L. Liu, K. Horneman, and Y. Teng, "Cooperative distributed optimization for the hyper-dense small cell deployment," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 61–67, 2014.

[24] Q. Li, H. Niu, A. Papathanassiou, and G. Wu, "5G Network Capacity: Key Elements and Technologies," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 71–78, 2014.

[25] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic Geometry for Modeling, Analysis, and Design of Multi-Tier and Cognitive Cellular Wireless Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 996–1019, 2013.

[26] D. Ohmann, A. Fehske, and G. Fettweis, "Transient flow level models for interference-coupled cellular networks," in *Proc. of the Annual Allerton Conference on Communication, Control, and Computing*, pp. 723–730, 2013.

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

[28] S. Singh and J. Andrews, "Joint Resource Partitioning and Offloading in Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 888–901, 2014.

[29] *3GPP TS 24.302. Access to the 3GPP Evolved Packet Core (EPC) via non-3GPP access networks (Release 12)*, June 2014.

[30] W. Ni and I. Collings, "A New Adaptive Small-Cell Architecture," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 5, pp. 829–839, 2013.

[31] D. Astely, E. Dahlman, G. Fodor, S. Parkvall, and J. Sachs, "LTE Release 12 and beyond," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 154–160, 2013.

[32] Y. Kojima, J. Suga, T. Kawasaki, M. Okuda, and R. Takechi, "LTE-WiFi Link Aggregation at Femtocell Base Station," in *Proc. of the World Telecommunications Congress*, pp. 1–6, 2014.

[33] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, "Modeling and Analysis of K-Tier Downlink Heterogeneous Cellular Networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 550–560, 2012.

[34] O. Galinina, S. Andreev, M. Gerasimenko, Y. Koucheryavy, N. Himayat, S. ping Yeh, and S. Talwar, "Capturing spatial randomness of heterogeneous cellular/WLAN deployments with dynamic traffic," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1083–1099, 2014.

[35] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP Heterogeneous Networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, 2011.

[36] D. Lopez-Perez, X. Chu, and I. Guvenc, "On the Expanded Region of Picocells in Heterogeneous Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 3, pp. 281–294, 2012.

[37] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution Towards 5G Multi-tier Cellular Wireless Networks: An Interference Management Perspective," *IEEE Wireless Communications Magazine, to appear*, vol. TBD, pp. 1–10, 2014.

[38] M. Bennis, S. Perlaza, P. Blasco, Z. Han, and V. Poor, "Self-Organization in Small Cell Networks: A Reinforcement Learning Approach," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3202–3212, 2013.

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

[40] S. Singh, F. Baccelli, and J. Andrews, "On Association Cells in Random Heterogeneous Networks," *IEEE Wireless Communications Letters*, vol. 3, no. 1, pp. 70–73, 2014.

[41] V. Kalashnikov, *Mathematical Methods in Queuing Theory*. Springer Science & Business Media, 1993.

[42] S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, and Y. Koucheryavy, "Cellular traffic offloading onto network-assisted device-to-device connections," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 20–31, 2014.

[43] C. de Lima, M. Bennis, and M. Latva-aho, "Statistical Analysis of Self-Organizing Networks with Biased Cell Association and Interference Avoidance," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 1950–1961, 2013.

[44] S. Borst, S. Hanly, and P. Whiting, "Optimal resource allocation in HetNets," in *Proc. of the IEEE International Conference on Communications*, pp. 5437–5441, 2013.

[45] *WINTERsim tool,* `http://winter-group.net/downloads/`.

AUTHORS' BIOGRAPHIES

Olga Galinina is a Ph.D. Candidate in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. She received her B.Sc. and M.Sc. degrees in Applied Mathematics from Department of Applied Mathematics, Faculty of Mechanics and Physics, St. Petersburg State Polytechnical University, Russia. Her research interests include applied mathematics and statistics, queuing theory and its applications; wireless networking and energy efficient systems, machine-to-machine and device-to-device communication.

Alexander Pyattaev is a Ph.D. Candidate in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. He received his B.Sc. degree from St. Petersburg State University of Telecommunications, Russia, and his M.Sc. degree from Tampere University of Technology. Alexander has publications on a variety of networking-related topics in internationally recognized venues, as well as several technology patents. His primary research interest lies in the area of future wireless networks: shared spectrum access, smart RAT selection and flexible, adaptive topologies.

Sergey Andreev is a Senior Research Scientist in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. He received the Specialist degree (2006) and the Cand.Sc. degree (2009) both from St. Petersburg State University of Aerospace Instrumentation, St. Petersburg, Russia, as well as the Ph.D. degree (2012) from Tampere University of Technology. Sergey (co-)authored more than 90 published research works on wireless communications, energy efficiency, heterogeneous networking, cooperative communications, and machine-to-machine applications.

Mischa Dohler is Chair Professor in Wireless Communications at King's College London, Director of the Centre for Telecommunications Research, co-founder and member of the Board of Directors of the smart city pioneer Worldsensing, Fellow (2014) and Distinguished Lecturer of the IEEE, and Editor-in-Chief of the Transactions on Emerging Telecommunications Technologies. He is a frequent keynote, panel and tutorial speaker. He has pioneered several research fields, contributed to numerous wireless broadband, IoT/M2M and cyber security standards, holds a dozen patents, organized and chaired numerous conferences, has more than 200 publications, and authored several books. He has a citation h-index of 39 (top 1%). He acts as policy, technology and entrepreneurship adviser, examples being Richard Branson's Carbon War Room, House of Lords UK, UK Ministry BIS, EPSRC ICT Strategy Advisory Team, European Commission, ISO Smart City working group, and various start-ups. He is also an entrepreneur, angel investor, passionate pianist and fluent in 6 languages. He has talked at TEDx. He had coverage by national and international TV & radio; and his contributions have featured on BBC News and the Wall Street Journal.

Yevgeni Koucheryavy is a Full Professor and Lab Director at the Department of Electronics and Communications Engineering of Tampere University of Technology (TUT), Finland. He received his Ph.D. degree (2004) from TUT. He is the author of numerous publications in the field of advanced wired and wireless networking and communications. His current research interests include various aspects in heterogeneous wireless communication networks and systems, the Internet of Things and its standardization, as well as nanocommunications. He is Associate Technical Editor of IEEE Communications Magazine and Editor of IEEE Communications Surveys and Tutorials.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3564-2
ISSN 1459-2045