TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Sergey Andreev

Energy Efficient and Cooperative Solutions for Next-Generation Wireless Networks

Julkaisu 1060 • Publication 1060

Tampere 2012

Tampereen teknillinen yliopisto. Julkaisu 1060
Tampere University of Technology. Publication 1060

Sergey Andreev

Energy Efficient and Cooperative Solutions for Next-Generation Wireless Networks

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 21st of August 2012, at 12 noon.

Tampereen teknillinen yliopisto – Tampere University of Technology
Tampere 2012

Supervisor:

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

Co-supervisor:

Andrey Turlikov, Ph.D., Professor
Department of Information Systems Security
St. Petersburg State University of Aerospace Instrumentation
St. Petersburg, Russia

Pre-examiners:

Timo Hämäläinen, Ph.D., Professor
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland

Boris Bellalta, Ph.D., Assistant Professor
Department of Information and Communication Technologies
Universitat Pompeu Fabra
Barcelona, Spain

Opponent:

Anthony Ephremides, Ph.D., Distinguished University Professor
Department of Electrical and Computer Engineering
University of Maryland
College Park, Maryland, USA

ISBN 978-952-15-2881-1 (printed)
ISBN 978-952-15-2899-6 (PDF)
ISSN 1459-2045

ABSTRACT

Energy efficiency is increasingly important for next-generation wireless systems due to the limited battery resources of mobile clients. While fourth generation cellular standards emphasize low client battery consumption, existing techniques do not explicitly focus on reducing power that is consumed when a client is actively communicating with the network. Based on high data rate demands of modern multimedia applications, active mode power consumption is expected to become a critical consideration for the development and deployment of future wireless technologies.

Another reason for focusing more attention on energy efficient studies is given by the relatively slow progress in battery technology and the growing quality of service requirements of multimedia applications. The disproportion between demanded and available battery capacity is becoming especially significant for small-scale mobile client devices, where wireless power consumption dominates within the total device power budget. To compensate for this growing gap, aggressive improvements in all aspects of wireless system design are necessary.

Recent work in this area indicates that joint link adaptation and resource allocation techniques optimizing energy efficient metrics can provide a considerable gain in client power consumption. Consequently, it is crucial to adapt state-of-the-art energy efficient approaches for practical use, as well as to illustrate the pros and cons associated with applying power-bandwidth optimization to improve client energy efficiency and develop insights for future research in this area. This constitutes the *first objective* of the present research.

Together with energy efficiency, next-generation cellular technologies are emphasizing stronger support for heterogeneous multimedia applications. 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 design principles of such networks. However, as wireless resources are limited and shared by clients, service integration may become challenging. A key element in such systems is the packet scheduler, which typically helps ensure that the individual quality of service requirements of wireless clients are satisfied.

In contrastingly different distributed wireless environments, random multiple access protocols are beginning to provide mechanisms for statistical quality of service assurance. However, there is currently a lack of comprehensive analytical frameworks which allow reliable control of the quality of service parameters for both cellular and local area networks. Providing such frameworks is therefore the *second objective* of this thesis. Additionally, the study addresses the simultaneous operation of a cellular and a local area network in spectrally intense metropolitan deployments and solves some related problems.

Further improving the performance of battery-driven mobile clients, cooperative communications are sought as a promising and practical concept. In particular, they are capable of mitigating the negative effects of fading in a wireless channel and are thus expected to enhance next-generation cellular networks in terms of client spectral and energy efficiencies. At the cell edges or in areas missing any supportive relaying infrastructure, client-based cooperative techniques are becom-

ing even more important. As such, a mobile client with poor channel quality may take advantage of neighboring clients which would relay data on its behalf.

The key idea behind the concept of client relay is to provide flexible and distributed control over cooperative communications by the wireless clients themselves. By contrast to fully centralized control, this is expected to minimize overhead protocol signaling and hence ensure simpler implementation. Compared to infrastructure relay, client relay will also be cheaper to deploy. Developing the novel concept of client relay, proposing simple and feasible cooperation protocols, and analyzing the basic trade-offs behind client relay functionality become the *third objective* of this research.

Envisioning the evolution of cellular technologies beyond their fourth generation, it appears important to study a wireless network capable of supporting machine-to-machine applications. Recent standardization documents cover a plethora of machine-to-machine use cases, as they also outline the respective technical requirements and features according to the application or network environment. As follows from this activity, a smart grid is one of the primary machine-to-machine use cases that involves meters autonomously reporting usage and alarm information to the grid infrastructure to help reduce operational cost, as well as regulate a customer's utility usage.

The preliminary analysis of the reference smart grid scenario indicates weak system architecture components. For instance, the large population of machine-to-machine devices may connect nearly simultaneously to the wireless infrastructure and, consequently, suffer from excessive network entry delays. Another concern is the performance of cell-edge machine-to-machine devices with weak wireless links. Therefore, mitigating the above architecture vulnerabilities and improving the performance of future smart grid deployments is the *fourth objective* of this thesis.

Summarizing, this thesis is generally aimed at the improvement of energy efficient properties of mobile devices in next-generation wireless networks. The related research also embraces a novel cooperation technique where clients may assist each other to increase per-client and network-wide performance. Applying the proposed solutions, the operation time of mobile clients without recharging may be increased dramatically. Our approach incorporates both analytical and simulation components to evaluate complex interactions between the studied objectives. It brings important conclusions about energy efficient and cooperative client behaviors, which is crucial for further development of wireless communications technologies.

Preface

To my teachers...

The research work summarized in this thesis has been carried out in the Department of Communications Engineering at Tampere University of Technology (Finland) over the years 2010-2012. This manuscript would not have been possible in its current form without the support of many wonderful people, who are gratefully acknowledged here, without the intention of forgetting anyone.

First and foremost, I have been extremely fortunate to work under the supervision of Prof. Yevgeni Koucheryavy, who has improved my research capabilities significantly. His unconditional concern for intellectual and personal growth inside his group, as well as his strong leadership skills, will always be the role model of a successful mentor to me. Also I would like to express my deepest appreciation to Prof. Andrey Turlikov from St. Petersburg State University of Aerospace Instrumentation (Russia). As a co-supervisor of this thesis, he initialized my interest in wireless communications and shaped my research capabilities. Beyond his insight, intuition, and intelligence, he is really patient in advising students so that they could develop an ability for independent thinking.

It has been a real privilege to work next door to Prof. Markku Renfors and Prof. Jarmo Harju, who were always ready to share their wisdom and time.

Needless to say, I would like to express my gratitude to the reviewers of this thesis, Prof. Timo Hämäläinen from University of Jyväskylä (Finland) and Assistant Prof. Boris Bellalta from Universitat Pompeu Fabra (Spain) 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 Distinguished University Prof. Anthony Ephremides from University of Maryland (USA) 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 Tampere Graduate School in Information Science and Engineering (TISE) for the initial three years of my research work in Tampere. The financial contribution by Nokia Foundation and HPY Research Foundation in the form of personal grants is also greatly appreciated.

I am proud of having an unparalleled opportunity to collaborate all these years with many brilliant experts and outstanding professionals. Dr. Zsolt Saffer from Budapest University of Technology and Economics (Hungary) has my sincere gratitude for inspiring technical discussions that proved to be invaluable and fruitful for my research. I am very happy to keep in touch with Dr. Nageen Himayat and Dr. Kerstin Johnsson from Wireless Communications Laboratory, Intel Corporation (USA). Their intellectual curiosity has always been a source of inspiration for me. Last but definitely not least, I am especially indebted to Alexey Vinel, who continually encourages me to pursue high-quality academic writings.

Remembering my time at the Department of Communications Engineering, I would like to thank my colleagues and co-authors for making this place vivid, warm, and attractive. It was a pleasure doing research side by side with Dmitri Moltchanov, Eugeniy Belyaev, Olga Galinina, Alexander Pyattaev, Vitaly Petrov, Mikhail Gerasimenko, and Luís de Sousa. I am also grateful to Alexey Anisimov and Eugeny Pustovalov, my co-authors in St. Petersburg.

Taking this opportunity, let me also acknowledge the kind support by the wonderful SMACS Research Group of Ghent University (Belgium) and particularly mention Prof. Herwig Bruneel, Dr. Dieter Fiems, Koen de Turck, and Thomas Demoor, who have first showed me what top-level European research looks like. To all of them I am most deeply indebted.

Naturally, I would like to extend my appreciation to Ulla Siltaloppi, Tarja Erälaukko, Daria Ilina, Sari Kinnari, Elina Orava, and Matti Tiainen for their responsiveness, prompt assistance with practical matters, and friendly support.

With the sincerest gratitude, I would like to finally thank my parents for their everlasting love, encouragement, support, and understanding, as well as my friends for the enjoyable moments we had together.

Sergey D. Andreev

August 1, 2012, Tampere, Finland

Table of Contents

List of Publications

This thesis is mainly based on the following publications:

[P1] 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, no. 14, pp. 1662–1671, 2012.

[P2] S. Andreev, Z. Saffer, A. Turlikov, and A. Vinel, “Upper bound on overall delay in wireless broadband networks with non real-time traffic,” in *Proc. of the 17th International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA)*, pp. 262–276, 2010.

[P3] S. Andreev, Z. Saffer, and A. Turlikov, “Delay analysis of wireless broadband networks with non real-time traffic,” in *Proc. of the 4th International Workshop on Multiple Access Communications (MACOM)*, pp. 206–217, 2011.

[P4] S. Andreev, Y. Koucheryavy, and L. de Sousa, “Calculation of transmission probability in heterogeneous ad hoc networks,” in *Proc. of the Baltic Congress on Future Internet and Communications (BCFIC)*, pp. 75–82, 2011.

[P5] S. Andreev, K. Dubkov, and A. Turlikov, “IEEE 802.11 and 802.16 cooperation within multi-radio stations,” *Wireless Personal Communications Journal (WIRE)*, vol. 58, no. 3, pp. 525–543, 2011.

[P6] S. Andreev, O. Galinina, and A. Vinel, “Performance evaluation of a three node client relay system,” *International Journal of Wireless Networks and Broadband Technologies (IJWNBT)*, vol. 1, no. 1, pp. 73–84, 2011.

[P7] S. Andreev, E. Pustovalov, and A. Turlikov, “A practical tree algorithm with successive interference cancellation for delay reduction in IEEE 802.16 networks,” in *Proc. of the 18th International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA)*, pp. 301–315, 2011.

[P8] S. Andreev, O. Galinina, and Y. Koucheryavy, “Energy-efficient client relay scheme for machine-to-machine communication,” in *Proc. of the 54th IEEE Global Communications Conference (GLOBECOM)*, 2011.

List of Abbreviations

3/4G	Third/Fourth Generation
3GPP	3G Partnership Project
ACK	Acknowledgment
AP	Access Point
BE	Best Effort
BEB	Binary Exponential Backoff
BS	Base Station
CA	Collision Avoidance
CCA	Clear Channel Assessment
CCI	Co-Channel Interference
CDMA	Code Division Multiple Access
CSI	Channel State Information
CSMA	Carrier Sense Multiple Access
CTS	Clear To Send
D2D	Device-to-Device
DCF	Distributed Coordination Function
DL	DownLink
DRX	Discontinuous Reception
EDCA	Enhanced Distributed Channel Access
ertPS	extended real-time Polling Service
FDD	Frequency Division Duplex
HTTP	HyperText Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things

LTE	Long Term Evolution
LTE-A	LTE-Advanced
M2M	Machine-to-Machine
MAC	Medium Access Control
MANET	Mobile Ad-hoc Network
MIMO	Multiple-Input and Multiple-Output
MR	Multi-Radio
NAV	Network Allocation Vector
nrtPS	non real-time Polling Service
OFDM	Orthogonal Frequency-Division Multiplexing
OFDMA	OFDM Access
PHY	Physical
QoS	Quality of Service
RF	Radio Frequency
R-SICTA	Robust SICTA
rtPS	real-time Polling Service
RTS	Request To Send
SIC	Successive Interference Cancellation
SICTA	SIC and a Tree Algorithm
SINR	Signal to Interference-plus-Noise Ratio
SS	Subscriber Station
TDD	Time Division Duplex
TXOP	Transmission Opportunity
UGS	Unsolicited Grant Service
UL	UpLink
VoIP	Voice over Internet Protocol
WLAN	Wireless Local Area Network
WPAN	Wireless Personal Area Network
WWAN	Wireless Wide Area Network

List of Figures

Chapter 1

Introduction

1.1 RESEARCH MOTIVATION

The rapid expansion of wireless cellular systems over the last decades has introduced fundamental changes to “anytime, anywhere” mobile Internet access, as well as posed new challenges for the research community. The *fourth generation* (4G) of broadband communication standards targets aggressive improvements in all aspects of wireless system design, including system capacity, energy efficiency, and *quality of service* (QoS).

Primary next-generation challenges include, for example, energy efficient communications, multi-radio coexistence, client cooperation, and support for advanced services (see Figure 1.1). These key research directions are insufficiently addressed by the conventional simulation methodology and existing analytical models. Moreover, known approaches fail to account for many important performance factors, such as realistic network architectures, practical QoS frameworks, wireless channel degradation factors, etc. As the result, the output of these models provides inadequate insight into the performance of a real-world wireless network.

The main target of this thesis is the development of the comprehensive system models and evaluation methodologies that may be used for the performance assessment of next-generation (4G and beyond) communication technologies. Throughout this work, we particularly emphasize the need for energy efficient system behavior to save power for wireless devices with a tight battery budget. Further, we address the most crucial QoS aspects of contemporary *wireless wide area networks* (WWANs) and *wireless local area networks* (WLANs), first separately, and then in the more practical scenarios where a WWAN and a WLAN are co-located. We also adopt the promising concept of cooperative communications for WWAN environments, where wireless clients may assist each other when transmitting data. Analyzing various flavors of client relay, we propose and detail several practical solutions for next-generation networks. Finally, we consider emerging machine-to-machine communications that are predicted to support the Internet of Things. In particular, we

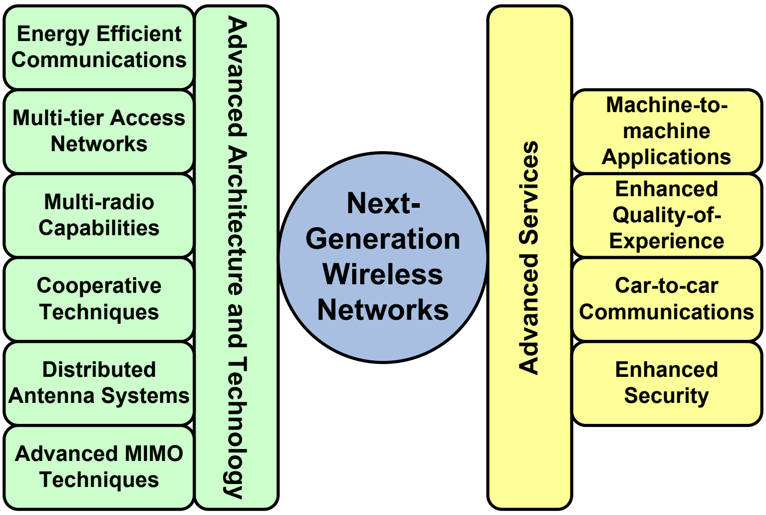

Energy Efficient Communications
Multi-tier Access Networks
Multi-radio Capabilities
Cooperative Techniques
Distributed Antenna Systems
Advanced MIMO Techniques
Advanced Architecture and Technology
Next-Generation Wireless Networks
Advanced Services
Machine-to-machine Applications
Enhanced Quality-of-Experience
Car-to-car Communications
Enhanced Security

Figure 1.1 Next-generation network challenges.

review the state-of-the-art technology enablers and then focus critical design challenges, such as handling a large number of devices and improving the performance of devices with weak links.

Consequently, our approach involves deep and systematic study of energy efficient and cooperative techniques in modern and future wireless networks. It not only proposes new system architectures and corresponding communication algorithms that substantially improve spectral efficiency, energy efficiency, and effectively satisfy diverse performance requirements of heterogeneous traffic flows, but also provides a deep understanding of fundamental mechanisms in advanced wireless resource management. The joint study of energy efficiency, QoS, heterogeneity, and cooperation is expected to facilitate the deployment of next-generation wireless multimedia networks that support converged client objectives in complex heterogeneous wireless environments.

1.2 SCOPE OF THE THESIS

Adoption of wireless technology has become increasingly widespread as new high data rate communication standards emerge, allowing for improved access to services and applications previously only supported through fixed broadband systems. The increasing importance of energy efficiency for wireless systems is given by the relatively slow progress in battery technology and the growing quality of service requirements of multimedia applications. The disproportion between demanded and available battery capacity is becoming especially significant for small-scale mobile client devices, where wireless power consumption dominates within the total device power budget. To compensate for this growing gap, aggressive improvements in all aspects of wireless system design are necessary.

Wireless client power saving has been an important consideration in defining WWAN standards where special protocols have been developed to reduce the energy consumption of a mobile device. Consequently, wireless technologies support reduction in client power consumption through maximizing the clients' sleep and idle periods. However, they do not explicitly address the energy expenses when a client is active, that is, communicating with the network. Given the battery-limited power budget of mobile devices and the high data rate demands of multimedia applications, *active mode* power consumption becomes an important consideration for wireless system design and standards development.

However, little research addresses active mode power consumption. Some theoretical frameworks apply optimization theory to propose joint link adaptation and resource allocation strategies. While effective solutions have indeed been obtained, they have never been tested in a realistic wireless cellular environment. Therefore, concluding upon the feasibility of such energy efficient schemes is an important consideration in this thesis. Additionally, it is practically valuable to understand how close the power saving techniques proposed by competitor 4G technologies are. As such, we also target the comparison of power saving mechanisms within alternative next-generation standards.

Whereas energy efficiency is accentuated by the need of extending client device operation time without recharging, the support for higher QoS is dictated by the ubiquitous wireless multimedia applications. Contemporary WWAN standards provide a variety of QoS features, but effective mechanisms to control those features are typically left out of scope. In this thesis, we review the most advanced QoS schemes in contemporary WWAN deployments and then demonstrate how to apply them efficiently to improve client performance. Additionally, we address QoS provisioning in the state-of-the-art WLANs and build a comprehensive framework to control performance across diverse client requirements.

Currently, WWAN, WLAN, and *wireless personal area network* (WPAN) technologies as well as supportive network architectures are evolving toward more advanced and complex *converged* networks (see Figure 1.2). Hence, it is highly relevant to consider client operation in areas where different wireless communication systems are co-located. On the other hand, consumer electronics is spawning a huge explosion in number and variety of *multi-radio* devices, driven by the user demand for "anytime, anywhere" connectivity. The problem of interworking between disjoint wireless technologies within a client multi-radio device is therefore addressed in this thesis in order to develop provably efficient coordination protocols that allow for significant performance improvement in heterogeneous wireless environments.

Aiming at even higher performance improvement, we recall that various *diversity* techniques are known to mitigate the negative effects of multipath channel fading and thus increase the reliability of a wireless link. Whereas much research effort has been invested into time and frequency domains, spacial domain diversity is only beginning to come to attention. Consequently, one of the most promising approaches for next-generation mobile systems is spatial transmit diversity that exploits two or more transmit antennas to enhance the link quality. However, mobile terminals with multiple transmit antennas may be costly to produce due to their size and/or hardware limitations. For this reason, a concept of *cooperative communications* has

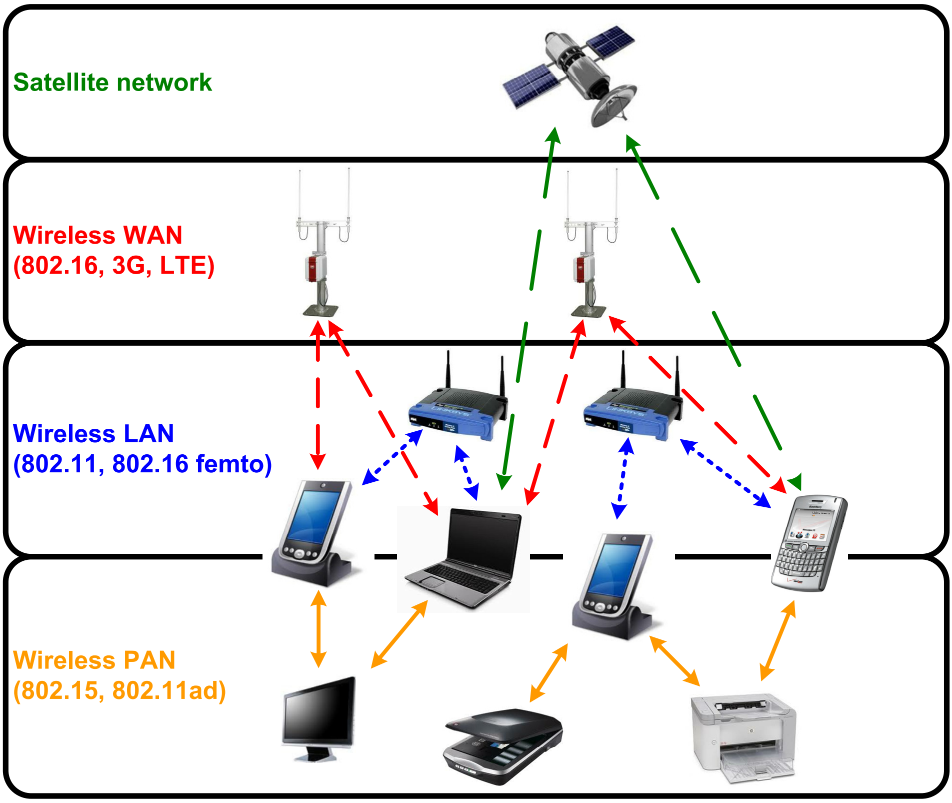


Figure 1.2 Converged communication technologies.

been introduced allowing single-antenna mobile devices to take advantage of spatial diversity gain and provide so-called cooperative diversity.

With cooperative communications, neighboring clients across a WWAN deployment may assist each other by relaying traffic opportunistically. We may further differentiate between the two flavors of this technique, which we term *client relay*, in what follows. In homogeneous client relay, all wireless links are in-band, whereas heterogeneous client relay allows for out-of-band offloading. The latter option may use WLAN technology for client-to-client links to save WWAN resources meaning, in turn, that cooperating clients should have multi-radio capabilities. Client relay schemes are becoming increasingly attractive as they may take advantage of both distributed and cellular assisted control functions, thus providing a flexible mechanism for improving system spectral and energy efficiencies. As homogeneous client relay is expected to lead to simpler algorithms, we focus on this alternative in the course of this thesis to propose efficient solutions and verify them in realistic WWAN environments. We also couple cooperation with power saving client behavior under a more general system model.

Finally, we argue that advanced services, such as *machine-to-machine communications*, are about to reshape the Internet as we know it today. The paradigm of the *Internet of Things* is currently generating a lot of attention while beyond-4G technologies are targeting decisive improvements to support the tight requirements of emerging machine-to-machine applications. Given our experience in standardization, we indicate vulnerable system design components and critical needs, e.g., supporting a large number of devices accessing the network nearly simultaneously and recovering the performance of devices with weak wireless links. Here, we propose the use of advanced signal processing techniques at the receiver and couple these with an improved multi-access algorithm. Additionally, we tailor our homogeneous client relay scheme to the reference machine-to-machine scenario in order to support devices with poor channel quality and improve their spectral and energy efficiencies.

Summarizing, this thesis targets various aspects of energy efficiency, heterogeneous QoS assurance, and cooperative communications in the context of next-generation wireless networks. We not only propose effective communication algorithms, but also test them in practical wireless environments where applicable. The ultimate goals of this research are to propose novel and verify existing state-of-the-art solutions, to extend the respective evaluation methodologies, as well as to predict the final gains within the realistic wireless system deployments. In order to achieve the objectives of the study, we extensively combine advanced analytical and simulation techniques. Concluding, the obtained results are expected to reduce power consumption of wireless mobile devices and improve their performance, thus providing a significant contribution to the next-generation networking community.

1.3 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 publications [1], [2], [3], [4], [5], [6], [7], and [8], which are summarized and seamlessly integrated into the body of the manuscript. In order to make this text accessible by a more general audience, we begin with the fundamental issues and core trade-offs related to spectral and energy efficiencies. We then gradually shift toward particular protocols, architectures, and algorithms to provide the reader with additional important details. Finally, we consider selected key scenarios and features which require specific knowledge of the state-of-the-art communication systems. Facilitating the flow of thought, the material given in the initial chapters is reused by the subsequent chapters. As such, the focus of the narration tends to transfer from more general to more detailed problem formulations and related research.

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 energy efficient communications and provide the reader with the related background. We then discuss energy efficient features in different modes of client operation, focusing more on the active mode behavior. Conducting a system-level performance evaluation, we adapt several advanced solutions for practical use within a reference 4G system. Our power optimization research indicates significant system-wide reductions in energy consumption due to enhanced link adaptation and resource allocation mechanisms.

In Chapter 3, various aspects related to QoS in contemporary WWANs and WLANs are discussed. We propose efficient QoS assurance solutions in terms of algorithms, architectures, and performance evaluation frameworks. Accounting for realistic cooperation between a WWAN and a WLAN, we address coordination algorithms within a multi-radio device to mitigate the effect of radio-to-radio interference. Our results report significant performance improvement when enabling coordination. In Chapter 4, we review client cooperation schemes and argue that they may be efficiently used in future networking design to improve spectral efficiency, energy efficiency, and QoS perception of wireless clients. In particular, we extensively analyze the technique of homogeneous client relay and predict the expected gains within the system-level context. We also update the reader on state-of-the-art advances for heterogeneous client relay.

In Chapter 5, the novel concept of machine-to-machine communications is addressed and recent progress in respective standardization is summarized including our own contributions. We then indicate weak architectural components that fail to satisfy target performance requirements, as well as recover performance by considering alternative techniques. In particular, we propose the usage of successive interference cancellation to reduce network entry delay in case of large device population and exploit a type of client relay scheme to increase the performance of devices with weak links. Our results confirm that the proposed solutions are successful in bridging across the indicated system vulnerabilities. Chapter 6 concludes the introductory part and outlines some interesting directions for future work. Finally, Chapter 7 summarizes the publications constituting the second part of this thesis and highlights the author's contribution to them.

Chapter 2

Energy Efficient Wireless Systems

2.1 INTRODUCTION AND MOTIVATION

2.1.1 General background

Wireless networks demonstrate worldwide proliferation, which has further advanced recently with the introduction of novel communication technologies [9], [10]. However, the future success of wireless communications significantly depends on the solution to overcome the disproportion between the demanded QoS and limited network resources. The overview of such candidate solutions captured in the following sub-sections is largely based on the comprehensive surveys in [11], [12], and [13] as well as the references therein. Where necessary, it is extended and updated with more recent results reflecting cutting-edge developments in the field.

Over the years, wireless spectrum has become one of the most valuable natural resources. Therefore, the importance of its efficient usage accentuates the need for *spectral efficiency*. However, *energy efficiency* is also becoming increasingly important primarily for small form-factor mobile devices [12]. This is due to the growing gap between the available and the required battery capacity, which is demanded by the ubiquitous multimedia applications [14], [15].

For the above reasons, efficient resource allocation and management becomes critical for technologies where multiple clients share the limited wireless spectrum [16]. Currently, the layered principle dominates in networking design and each system layer is operated independently to maintain architectural transparency [17]. Among conventional layers, the *physical* layer is responsible for the raw-bit transmission, whereas the *medium access control* layer arbitrates the access of clients to the shared wireless channel [11].

We reiterate the fact that wireless channels are commonly known to suffer from multipath fading. Furthermore, the statistical channel properties may vary significantly across different clients [18]. Therefore, the traditional layer-wise architecture turns out to lack flexibility and thus results in inefficient wireless resource utilization. As such, an integrated and adaptive design involving adjacent layers is required to overcome this limitation. Consequently, *cross-layer* optimization across both physical and medium access control layers is desired for efficient wireless resource allocation and data packet scheduling [19].

Enabling cross-layer optimization, *channel-aware* approaches have been introduced and developed recently to explicitly take into account wireless *channel state information* (CSI). Typically, a channel-aware technique flexibly adapts data transmission and dynamically controls resources to ensure that a client with more favorable channel conditions transmits its packet [20]. Taking advantage of independent channel variation across multiple clients, channel-aware approaches were shown to substantially improve network performance through multiuser diversity, whose gain increases with the growing client population [21].

2.1.2 Physical layer

The *physical* (PHY) layer is of primary importance in wireless communications due to the challenging nature of the underlying medium. It concentrates on raw-bit transmission over the wireless channel and incorporates *radio frequency* (RF) circuits, modulation and coding schemes, power control algorithms, and other key system elements [18]. Conventional wireless technologies are typically built to communicate data on a fixed set of operating points [22] by sacrificing flexible power adaptation for design simplicity. This often causes excessive energy consumption or pessimistic transmission rates selected for peak channel conditions [23]. Hence, PHY 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.

As wireless medium is shared, the communication efficiency and the energy consumption are affected not only by the performance of a point-to-point wireless link, but also by the interaction between the individual links across the entire network [17]. This necessitates a more complex system-level approach. Importantly, *orthogonal frequency division multiplexing* (OFDM) is becoming the primary modulation scheme for next-generation wireless standards [9], [10]. From a resource management perspective, multiple channels in OFDM-based systems have the potential for more efficient medium access control design since sub-carriers may be dynamically assigned to different clients [24], [25]. To further improve performance, adaptive power allocation on each sub-carrier may be applied [26], [27].

2.1.3 Medium access control layer

The *medium access control* (MAC) layer should maintain individual client QoS requirements and at the same time ensure that wireless resources are efficiently allocated maximizing network-wide performance metrics. Therefore, MAC strategies that manage resources pessimistically to guarantee worst-case QoS may often degrade total network spectral and energy efficiencies [12].

Typically, MAC schemes can be either distributed or centralized. For distributed access, MAC is expected to minimize the number of wasted transmissions that are corrupted by the interference from other network clients, whereas for centralized access efficient scheduling algorithms are necessary to exploit the variations across clients to maximize the overall network performance [28]. The MAC layer manages resources on behalf of the PHY layer and they together define the general wireless network operation.

2.1.4 Cross-layer approaches

Summarizing, spectral and energy efficiencies are influenced by every component of the wireless system architecture, ranging from RF circuits to user applications (see Figure 2.1 and [29]). As mentioned above, the conventional layer-wise architecture implies independent design of different layers and may result in sub-optimal network performance. By contrast, cross-layer approaches leverage the interactions between different system layers and may considerably improve performance in terms of adaptability to service, traffic, and environment dynamics [11], [12], [13]. Throughput optimization via cross-layer approaches has long been an attractive research direction [30], [31]. However, as wireless clients become increasingly mobile, the focus of recent efforts tends to shift toward energy consumption at all layers of communication systems, from architectures [32] to algorithms [33].

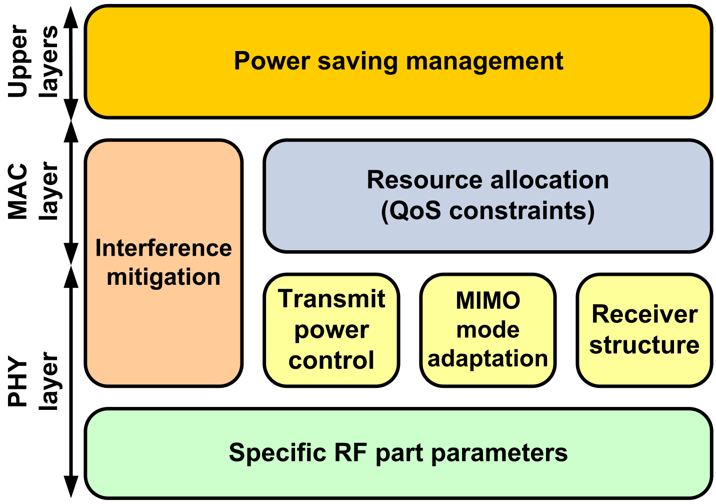


Figure 2.1 Energy efficient system components.

Given that wireless channels are shared and highly dynamic, efficient resource management is believed to be the most challenging element in the channel-aware system design [13], [34]. A scheduler to perform adaptive resource control should thus account for at least three primary performance metrics: *system capacity* (or spectral efficiency), *energy consumption* (or energy efficiency) of wireless clients, and their *quality of experience* (or QoS). It is also highly desirable to flexibly control the trade-offs associated with these metrics [35]. In the course of this thesis, we consider each of these important metrics and the related trade-offs in more detail.

2.2 SPECTRAL EFFICIENCY

2.2.1 Background

Because of fading, the characteristics of a wireless channel vary significantly with time, frequency, and client. Previously, we emphasized that as any wireless system relies on the shared medium, its communication performance depends on individual links and, more importantly, their interaction across the entire network. Accounting for this fact, channel-aware MAC schemes have been introduced to adaptively communicate data and dynamically manage wireless resources based on CSI [23], [36]. With these schemes, wireless network spectral efficiency, which refers to the data rate that can be transmitted over a given bandwidth in a particular communication system (typically, in $bit/s/Hz$), may be substantially improved [28].

The main principle behind cross-layer channel-aware MAC is to schedule a client with more favorable channel conditions to transmit with optimized link adaptation according to CSI [25], [27]. As mentioned above, MAC schemes can be either distributed or centralized and we consider each option separately.

2.2.2 Distributed medium access

Random multiple access algorithms allow clients to share network resources subject to distributed control. Conventional contention-based methods include pure, slotted, and reservation Aloha solutions, as well as *carrier sense multiple access* (CSMA) and CSMA with *collision avoidance* (CSMA/CA) schemes, multiple access with collision avoidance for wireless [37] technique, and many others [12]. However, these MAC approaches do not use CSI explicitly. Hence, when a client decides to transmit a packet, its wireless link may experience a deep fade [12]. By contrast, when the client link is in a favorable condition the transmission may be deferred, which is a waste of channel resources.

Recently, so-called *opportunistic* random multiple access schemes have been investigated in [38], [39], [40], and [41] to exploit CSI for performance improvement. With opportunistic random access, each client is made aware of its CSI and accounts for it during the contention behavior. Thus, clients with better channel qualities have higher contention probabilities and enjoy more frequent transmissions. It is important to emphasize that the majority of known opportunistic approaches consider wireless networks where clients transmit to a common receiver, e.g., an access point. However, this well-explored scenario does not include many practical wireless system setups which are typical for sensor [42], ad-hoc [43], [44], and mesh networks [45], [46]. Those may require separate attention.

The distributed random-access techniques and associated challenges are discussed in more detail in Chapter 3.

2.2.3 Centralized medium access

With assistance of a central controller, the highest performance is known to be obtained by scheduling the client with the best channel conditions [25], [27]. However, the extent of CSI feedback to dynamically determine such a client may sometimes incur excessive overheads, especially for densely-populated mobile networks, which negatively impacts network scalability [12]. To reduce the required CSI feedback, distributed approaches are often preferable. However, the operation of a distributed MAC protocol may sometimes be prohibited by the network topology.

Recently, the principles of MAC design have evolved from traditional point-to-point models to more advanced multiuser approaches (see Figure 2.2, [47], and [48]). Special attention has been paid to the fact that time-varying fading is a unique property of a wireless channel [11]. Previously, with adaptive modulation and coding, the client could transmit at higher data rates as long as the channel condition remained satisfactory [49]. However, spectral efficiency degraded dramatically during periods of deep fade. Consequently, exploiting multiuser diversity has become increasingly attractive and channel-aware scheduling was tailored originally for *code division multiple access* (CDMA) systems [50].

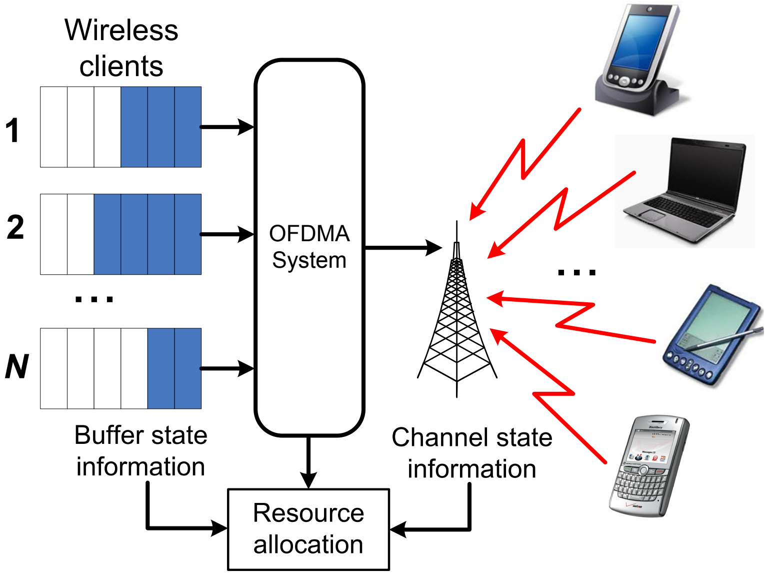


Figure 2.2 Multiuser wireless system.

The initial results with respect to multiuser diversity indicated that the use of a simple channel-aware scheduler alone can significantly recover network spectral efficiency [20]. Naturally, multiuser diversity gain follows from the independent channel variation for different clients. With the growing client population, the packets are more likely to be communicated at higher data rates since different clients experience independent fading fluctuations [11]. From a client perspective, if the system is enhanced with a channel-aware scheduler then the data is transmitted stochastically, which is sometimes referred to as opportunistic communications [51].

2.2.4 Interference-limited scenarios

Next-generation wireless networks, particularly those with cellular topology [9], [10], are becoming increasingly interference-limited as more clients share the same spectrum to receive high-rate multimedia service (see Figure 2.3). In modern cellular systems, *co-channel interference* (CCI) is often the dominant limiting factor that affects performance, especially as these systems shift toward deployments with more aggressive frequency reuse [9], [10]. Whereas the total spectral efficiency may indeed improve with aggressive frequency reuse, the performance of the *cell-edge* clients degrades dramatically.

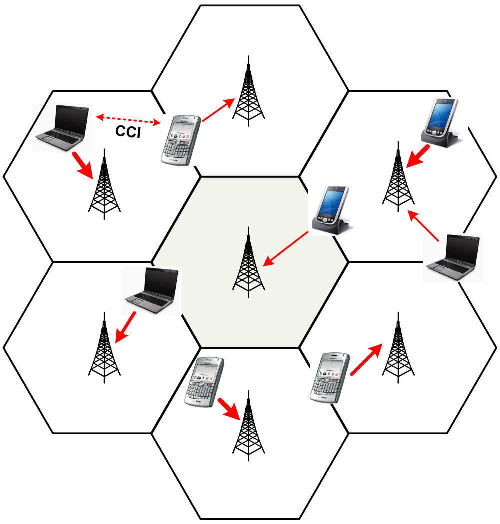


Figure 2.3 Multi-cell wireless network.

A popular CCI mitigation technique is to provide neighboring cells with non-overlapping sets of channels [52] and we refer to [53] for a good summary of channel assignment schemes. In particular, a relatively novel approach to reducing cell-edge interference is through *fractional frequency reuse* [54]. Hence, partial frequency reuse is applied for clients at cell edges, whereas full frequency reuse is specified for those at cell centers. Consequently, the throughput of cell-edge clients improves as they experience lower levels of interference.

Targeting a further increase in spectral efficiency with frequency reuse, CCI can be combated by applying advanced signal processing schemes, such as *interference cancellation* [55]. However, interference cancellation techniques are typically complex (see Chapter 5 for more discussion) and therefore may result in prohibitive implementation costs for mobile client devices. For *downlink* (DL) transmission, CCI can be reduced by joint encoding schemes across several base stations [56], or nearly avoided by using cooperative scheduling [57], both of which require an exchange of extra instantaneous feedback. Alternatively, contention-based techniques have also been introduced for CCI mitigation together with advanced channel-sensing MAC strategies [58].

2.3 ENERGY EFFICIENCY

2.3.1 Background

Energy efficiency is becoming increasingly important for contemporary wireless networks due to the limited battery lifetime of mobile clients. For maximizing energy efficiency, so-called “bits-per-Joule” [59] or “throughput-per-Joule” [28] metrics are often considered. As such, a measure of average energy efficiency of the client n in the time frame t may be the total data size sent by this client by the time t ($D_n[t]$) divided by the total consumed energy ($E_n[t]$):

$$u_n[t] = \frac{D_n[t]}{E_n[t]}. \tag{2.1}$$

Due to the fact that the radio frames typically have equal size, the equation (2.1) could be rewritten as:

$$u_n[t] = \frac{T_n[t]}{P_n[t]}, \tag{2.2}$$

where $T_n[t]$ is the throughput of the client n, $P_n[t]$ is the total consumed power. The $T_n[t]$ and $P_n[t]$ may be calculated recursively [60] by:

$$T_n[t] = \left(1 - \frac{1}{w}\right) T_n[t-1] + \frac{1}{w} \cdot r_n[t] \quad \text{and} \tag{2.3}$$

$$P_n[t] = \left(1 - \frac{1}{w}\right) P_n[t-1] + \frac{1}{w} \cdot p_n[t], \tag{2.4}$$

where $r_n[t]$ is the data rate of the client n in the frame t, $p_n[t]$ is the consumed power by the client n in the frame t, and w is the sample window length, $w \gg 1$. Thus, energy efficiency shows how many data bits are sent by a client per Joule of consumed energy (bit/J or bpJ).

Several approaches are known to focus energy efficiency, which may include water-filling power allocation techniques that optimize throughput with respect to the fixed total transmit power limitation [25], [27], as well as adaptation of both the total transmit power and its allocation according to the CSI [23], [36].

Again, we emphasize the important fact that energy efficiency of a wireless client is affected not only by the performance of its point-to-point communication link, but also by that of the other links in the system. Therefore, a cross-layer approach is required, including transmission adaptation together with multiuser resource assignment [11]. Moreover, energy-efficient schemes are expected to provide benefits to other co-channel clients by reducing the levels of interference.

Energy-efficient transmission has first been addressed within the framework of information theory more than two decades ago [61]. Summarizing the theoretical efforts, for any communication rate below the *capacity per unit energy* [62], the probability of an error decreases exponentially with increasing total energy [28]. In particular, it was demonstrated that the capacity per unit energy may be reached only with increasing bandwidth [63] or by extending the transmission time [64]. As both are difficult to achieve in a real-world wireless network, more practical approaches to improving energy efficiency are addressed below.

2.3.2 Link adaptation

As wireless channel quality varies with time, frequency, and client, link adaptation can be applied to improve communication performance. Link adaptation typically operates by adjusting modulation order, coding rate, and transmit power with respect to CSI [17]. Earlier research on link adaptation exploited power allocation to improve individual channel capacity [65], whereas state-of-the-art approaches highlight the importance of joint link adaptation and resource allocation [28].

More specifically, since channel frequency responses differ significantly across frequencies and clients, transmission rate adaptation for individual sub-carriers, dynamic sub-carrier selection, and flexible power adjustment may significantly improve the performance of OFDM-based networks [11]. With data rate adaptation, a client can enjoy a higher rate and lower power consumption over the sub-carriers in better condition so as to improve its throughput while ensuring an acceptable bit-error rate at every sub-carrier [49]. However, regardless of such adaptation, deep fading on particular sub-carriers may still degrade the channel capacity.

The vast majority of the information-theoretic findings (see e.g., [63] and [64]) account only for the transmit power when investigating energy consumption over the link. Typically, a client device will consume extra *circuit power* (see Figure 2.4), which is incurred independently of the transmission rate [66], [67]. As such, the circuit power consumption should be considered explicitly when optimizing energy efficiency [13]. Consequently, the known approach to maximize the transmission time may not be attractive anymore since circuit energy consumption grows with transmission duration. With the emphasis on circuit power, the challenge shifts toward using optimization theory for establishing energy-optimal link settings [23], [68].

Energy-efficient communications may thus be considered as a trade-off between transmit power, circuit power, and transmission time [67]. Clearly, the optimal rate that minimizes the total power consumption may be established with respect to a particular throughput constraint [69]. Despite the fundamental importance of power optimization for energy conservation and interference mitigation, surprisingly little research attention has been given to studying the joint operation of link adaptation and resource allocation. The work in [70] mentions some known solutions that focus separately on either throughput or energy efficiency in the context of power control for CDMA networks. Few other papers address this joint limitation and investigate energy-efficient power allocation for OFDM communications [23], [28], and [36].

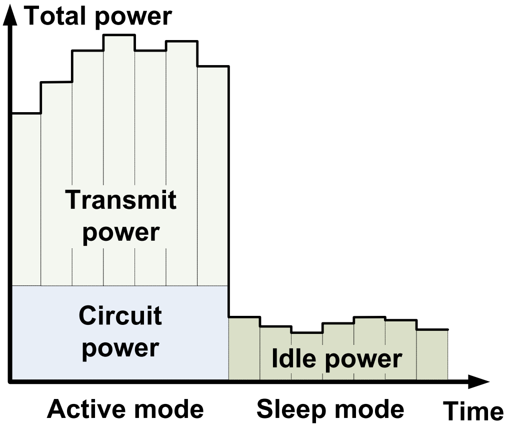


Figure 2.4 Example device power profile.

2.3.3 Resource allocation

Due to the scarcity of wireless resources, intricate performance trade-offs arise between an individual client and the entire network [12]. Exploiting diversity across clients is likely to reduce the total network energy consumption. Importantly, wireless resource management over different domains may further improve system energy efficiency. In this sub-section, we consider *time* and *frequency* domains, whereas *spatial* domain is focused on in Chapter 4.

With *time-division multiple access*, the wireless channel is shared by the clients in the time domain. Each client thus attempts to extend its transmission time to save some of its energy and consequently contradicts the respective needs of other clients [13]. As such, efficient transmission time allocation across all clients is critical for maximizing network energy efficiency. To define a practical resource management strategy, the process of scheduling may be partitioned into a design-time phase and a run-time phase [71]. In the design-time phase, the energy-performance profile of every client may be determined to capture the relevant trade-offs. In the run-time phase, low-complexity (greedy) solutions may be applied to adjust the operating points and further improve the energy efficiency.

While extensive efforts have been undertaken to optimize energy-efficient resource management in time domain, little attention has been devoted to frequency domain [12]. Here, while increasing transmission bandwidth naturally improves energy efficiency, the entire frequency resource cannot be allocated exclusively to one client. This is due to the fact that in a multiuser system the energy efficiency of other clients may suffer, as well as that of the overall network [28]. Therefore, frequency-domain resource control is crucial in optimizing the total energy efficiency of a wireless network. Frequency selectivity of broadband wireless channels emphasizes this need even further [23].

The OFDM technique is known to split the entire frequency channel into multiple orthogonal narrowband sub-channels (sub-carriers) to compensate for the frequency-selective fading and to support higher data rates. Furthermore, in an OFDM-based wireless network, different sub-carriers can be assigned to different clients to enable a flexible MAC scheme and to effectively exploit multiuser diversity. In multiuser environments, since channel properties for different clients are almost mutually independent [11], the sub-carriers suffering from a deep fade for one client may be favorable for other clients. Therefore, a particular sub-carrier could be in an attractive condition for some clients in a multiuser OFDM-based wireless network. Hence, with dynamic sub-carrier allocation, the network can gain additional performance through multiuser diversity [11].

2.4 ENERGY EFFICIENT CELLULAR NETWORKS

2.4.1 Fourth generation wireless systems

The parallel evolution of personal, local, and metropolitan area networks (see Figure 1.2 and [72]) provides end clients with a wide choice of infrastructures to use for a particular application. The *Institute of Electrical and Electronics Engineers* (IEEE) and the *3rd Generation Partnership Project* (3GPP) are currently introducing next-generation wireless technologies [9], [10]. Given the importance of power consumption for battery constrained mobile devices, client power saving and improved energy efficiency are challenging objectives for the emerging 4G standards [73].

Active mode power consumption is increasingly important for reliable *uplink* (UL) transmissions due to the significant transmit power required to overcome path loss degradation and low efficiency of contemporary RF power amplifiers. Consequently, by achieving reduced active mode power consumption, the battery lifetime of mobile clients can be extended, which is crucial for the deployment of 4G high data rate wireless networks [74].

Higher energy efficiency through reduced power consumed in the network is also becoming attractive due to environmental concerns as well as due to the operators' desire to reduce maintenance costs [75] (see e.g., 3GPP discussions on "green" radio access networks for Release 12 and beyond [76]). However, our focus in the remainder of this chapter will be on client energy efficiency. The results reported below are primarily concentrated on the IEEE 802.16 standards [9], but are equally applicable to other cellular technologies based on *OFDM access* (OFDMA), such as 3GPP *Long Term Evolution* (LTE) [10].

2.4.2 Advanced power saving operation

As repeatedly emphasized above, power saving mechanisms are becoming increasingly important for next-generation wireless networks. In order to save power and maximize the battery lifetime of small-scale mobile devices, either 4G cellular technology specifies a *power saving* technique. Improving client operation time without recharging its battery, IEEE 802.16m proposes a so-called *sleep mode*, whereas 3GPP *LTE-Advanced* (LTE-A) defines *discontinuous reception* (DRX) mode.

Studying mobile client performance in sleep or DRX mode requires the derivation of a more advanced wireless system model. Queuing theoretic methods may be used for this purpose. In particular, the behavior of a wireless client may be represented as a queue with vacations accounting for both delay and power consumption. Adequate modeling of power saving operation allows maximizing the client's sleep/DRX periods while satisfying its QoS constraints. As such, it has the potential to significantly decrease the power consumption of a battery-driven mobile device.

In this sub-section, we only mention some of our results on power saving and continue the related discussion in Chapter 4, where power saving operation is coupled with cooperative communications. In [1], we argue that existing research works are not addressing the explicit comparison between the DRX and the sleep mode. We then bridge in this gap by conducting the comparative analysis of the two power saving techniques. In particular, we focus on two polar scenarios with respect to practical traffic patterns: *voice over Internet protocol* (VoIP) and *hypertext transfer protocol* (HTTP) traffic. Our analysis indicates that both advanced power saving features generally show excellent energy efficient performance and are important for the development of the next-generation wireless cellular networks.

2.4.3 Existing energy efficient frameworks

We discussed above that modern wireless technologies support reductions in client power consumption through maximizing the client's sleep/DRX periods. However, they do not explicitly focus on *active mode* power consumption. Given the tight battery budget of mobile client devices and the increasingly high data rate demands of multimedia applications [77], active mode power consumption is also expected to become an important consideration for next-generation wireless networks.

Generally, Shannon's theorem for a point-to-point link can be considered to provide intuition on possible approaches for transmit power reduction:

$$c = b \cdot \log\left(1 + \frac{g \cdot p}{\sigma^2}\right), \tag{2.5}$$

where b is the allocated bandwidth, g is the channel gain, p is the transmit power of a client, σ^2 is the noise power, and c is the achievable capacity for a client.

The channel *capacity* is known to be the maximum rate at which reliable communication is possible in the system. Given that it is linearly related to bandwidth, but exponentially related to power, client transmit energy consumption may be reduced by one of the following.

- For the fixed data rate, if transmission bandwidth is increased, power can be exponentially decreased in the system.
- For the fixed data rate, clients experiencing good channel conditions can be scheduled.
- If the delay can be tolerated and the data rate is reduced, power can be exponentially decreased.

Therefore, the network can allocate power and bandwidth, as well as control delay across clients [78], to conserve transmit energy.

Surprisingly, very little scientific attention is paid to the problem of client energy efficiency in the active mode. Here we emphasize that the network may utilize additional techniques that go beyond simple management of the transmit power per link to improve battery consumption for its mobile clients [79]. One such solution is to develop novel resource allocation strategies at the base station that would minimize client power consumption. Joint power and resource optimization for wireless cellular systems has recently been investigated by a number of research works that are summarized below.

In [80], the problem of energy efficient transmission in OFDMA-based networks is studied for frequency-flat fading channels. Although cellular wireless channels are typically frequency-selective, the assumption of flat fading is a sufficient representation for the case when OFDMA systems exploit distributed or randomized sub-carrier sub-channelization. Consequently, the effective channel quality is relatively similar across all sub-channels and thus may be modeled through the flat fading assumption. The work in [23] extends the results of [80] for the frequency-selective wireless channels.

In [28], the essentials of the cross-layer wireless system design are summarized for energy efficient communications. Particularly, a general information-theoretic approach to the energy efficient communications is introduced. Further, several energy efficient resource allocation strategies are proposed accounting for both transmit and circuit power consumption. The paper [36] focuses energy efficient communications in interference-limited cellular systems. Low complexity solutions for energy efficient optimization are proposed in [60], which significantly reduces the computational complexity typical for earlier iterative approaches. Closed form solutions for energy efficient link adaptation and resource allocation are derived by considering time-averaged, steady state metrics.

The inherent limitation of the above power optimization research is that it is based on a simplified network model of the cellular environment. Therefore, we extend the previously reported results using a realistic system-level simulation model, which is compliant with the methodology proposed for the IEEE 802.16m standard [73]. Some of our findings are summarized below.

2.4.4 System-level performance evaluation

In [P1], we carried out an in-depth system-level performance evaluation (see Figure 2.5) of energy efficient resource and power optimization for OFDMA-based wireless cellular networks by extending our initial research in [2]. The most advanced IEEE 802.16m evaluation methodology [73] was used to investigate a reference 4G system and realistic system parameters, channel environments, and implementation considerations were addressed.

In the course of this work, we adapted several state-of-the-art approaches and associated modifications for practical use. In particular, low complexity energy efficient schemes from [60] were evaluated and shown to perform similar to near-optimal, but significantly more complex, iterative approaches. A performance comparison with existing state-of-the-art throughput efficient power optimization schemes was also considered.

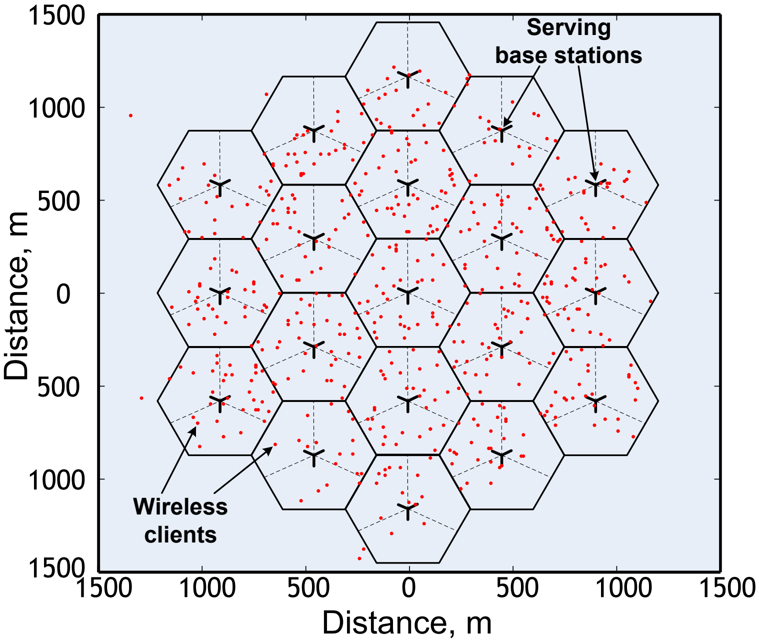


Figure 2.5 Example system-level client layout.

The contribution of [P1] is therefore the detailed investigation of important practical trade-offs, such as the dependence of the considered schemes on the circuit/idle power consumption, as well as amplifier efficiency and fairness aspects. We also focus on the important relationship between inter-cell interference and power reduction to compare the performance of energy efficient schemes against power-control based interference management. The results reported in [P1] illustrate the pros and cons associated with applying power-bandwidth optimization approaches for improving client energy efficiency and develop insights for future research in this field.

The system-level performance characterization of energy efficient wireless transmission techniques appears to be the first of its kind and indicates significant promise for this research area. Future extensions of this work need to focus on more advanced system models and algorithms [81]. In particular, the full-buffer assumption used in [P1] needs to be relaxed and the traffic models for multimedia services should be included [82]. Queuing models, arrival flows, and traffic-aware energy efficient scheduling might also be accounted for. It should also be noted that IEEE 802.16m standard now provides specific mechanisms for mobile devices to initiate their active mode power savings. The results reported in [P1] were important toward enabling this feature in the standards [83] and may be investigated for enhancing client energy efficiency in next-generation cellular system implementations.

Chapter 3

Heterogeneous Networking and QoS

3.1 INTRODUCTION AND MOTIVATION

3.1.1 General background

Modern wireless networks are constantly evolving to enable better support for heterogeneous multimedia applications [84], that is, at the very least, best effort and streaming traffic [13]. 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 [70], [85]. However, as wireless resources are limited and shared by clients, service integration may become challenging [86], [87], especially in heterogeneous networks [88], [89], [90]. A key element in such systems is the packet scheduler, which typically helps ensure that the individual QoS requirements of wireless clients are satisfied. As discussed in the previous chapter, 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 [51], [91], [92], [93], [94], and [95].

More advanced QoS-constrained opportunistic frameworks for wireless cellular networks focus *flow-level* performance and consider stochastic traffic loads [96]. 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 [97]. 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 realistic network may represent a stream of packets corresponding to a new file transfer, web-page browsing, or real-time voice/video session [13]. To mimic the flows produced by a large population of independent clients, *Poisson processes* have extensively been applied [92]. Originally, flow-level frameworks were helpful investigating flexible bandwidth allocation mechanisms in the context of wired systems [98]. Extending their applicability to wireless networks, it was concluded that the throughput experienced by a dynamic client population can substantially differ from that received by a fixed number of clients [92]. As such, studying dynamic wireless systems may require separate attention and some promising steps in this direction are discussed below.

3.1.2 Opportunistic resource management

Time- and frequency-varying channels are unique properties of a wireless client as opposed to wired communication networks. We emphasized previously that inefficient use of channel variability may potentially result in degraded system capacity. As a preventive measure, channel-aware (*opportunistic*) scheduling may be applied, where the network favors clients with better channel conditions. Opportunistic scheduling has been demonstrated to considerably improve system capacity [51], [91] and thus constitutes a promising strategy to achieve higher performance benefits.

However, in an opportunistic wireless system, more capacity may not always translate into better levels of QoS [97]. This stems from the fact that the maximum achievable capacity is constrained by the individual QoS requirements of wireless clients [99]. Specifically, to support minimum bandwidth guarantees, it is often required to sacrifice opportunism by serving clients with a relatively bad channel state. With increasing integration of diverse multimedia applications, the benefits of opportunism continue to cut down even further, which may negatively impact both system throughput and traffic delay. All these negative factors are known as the loss in opportunism due to integration [13].

In an integrated dynamic wireless network, where client population changes over time, it is important to investigate time-averaged system performance metrics, such as average throughput and traffic delay. Some initial studies, such as in [92], did not focus on the mixture of different traffic. Other efforts to investigate diverse packet flows within wireless opportunistic systems have been made later by [94]. However, these studies targeted packet-level performance, whereas the work in [96] has recently addressed flow-level dynamics, but in the context of the simplified system model. Therefore, there is currently a lack of comprehensive analytical frameworks which allow reliable control of the QoS parameters for heterogeneous wireless systems and we concentrate on bridging in this gap in what follows.

3.2 QOS ASPECTS OF WWANS

3.2.1 General QoS architecture

Contemporary WWAN deployments extensively use mobile cellular network technologies with the leading positions being held primarily by IEEE 802.16 [9] and 3GPP LTE [10]. Believed to be the most advanced wireless communication standards, IEEE 802.16 specifications detail a high speed wireless access system with support for various multimedia services. The IEEE 802.16 family of standards has been commercialized under the name of WiMAX (from “Worldwide Interoperability for Microwave Access”). Given the importance of this pioneering technology, we concentrate on IEEE 802.16 mechanisms and features in the course of this chapter.

More specifically, a series of IEEE 802.16 standards defines an air interface for broadband wireless access systems. As the result of a recent revision, the contemporary baseline document IEEE 802.16-2009 [100] consolidates the legacy IEEE 802.16-2004 [101] protocol with several amendments (see Figure 3.1). Further, IEEE 802.16m-2011 [9], also known as Mobile WiMAX Release 2, is an advanced air interface complying with the ITU-R IMT-Advanced requirements on 4G systems. The high data rate technology specified by the IEEE 802.16 enables multimedia services in both *line-of-sight* (LOS) and *non line-of-sight* (NLOS) conditions as well as provides support for diverse traffic flows to satisfy a wide range of QoS requirements for its clients.

Generally, the QoS is a broad concept and refers to the network ability of assigning different priority to diverse multimedia applications, clients, or their data flows, as well as to guarantee certain levels of performance to a particular data flow. QoS metrics may include received throughput, data packet delay, jitter, bit error rate, packet drop probability, and others. In this thesis, we, however, concentrate on the primary QoS characteristics, such as client and system throughputs, as well as the mean packet delay.

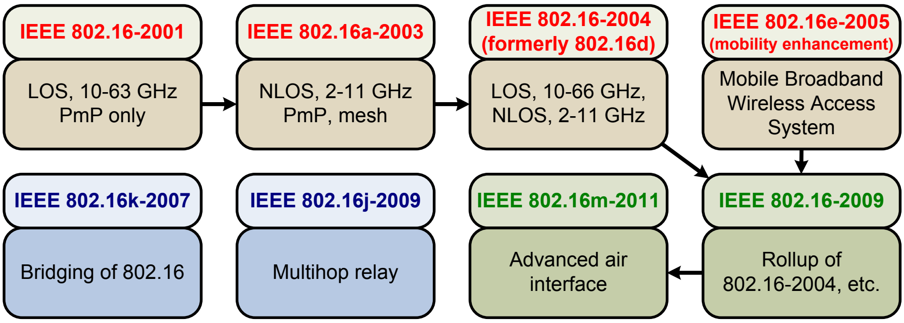


Figure 3.1 IEEE 802.16 standards evolution.

The core point-to-multipoint (PmP) IEEE 802.16 architecture comprises a *base station* (BS) and a set of clients or *subscriber stations* (SSs) in its vicinity (see Figure 3.2). BS performs polling of its SSs and schedules their transmissions such that the QoS guarantees of each data flow at every SS are satisfied [102]. The BS and the associated SSs exchange packets through disjoint communication channels. In the DL channel, the BS broadcasts data to its SSs, whereas in the UL channel the transmissions from the SSs are multiplexed. IEEE 802.16 provides two duplexing schemes for the aforementioned DL and UL channels. With *time division duplex* (TDD), a time frame is split into DL and UL sub-frames, respectively. With *frequency division duplex* (FDD), the channel frequency range is divided into non-overlapping sub-ranges to avoid cross-interference.

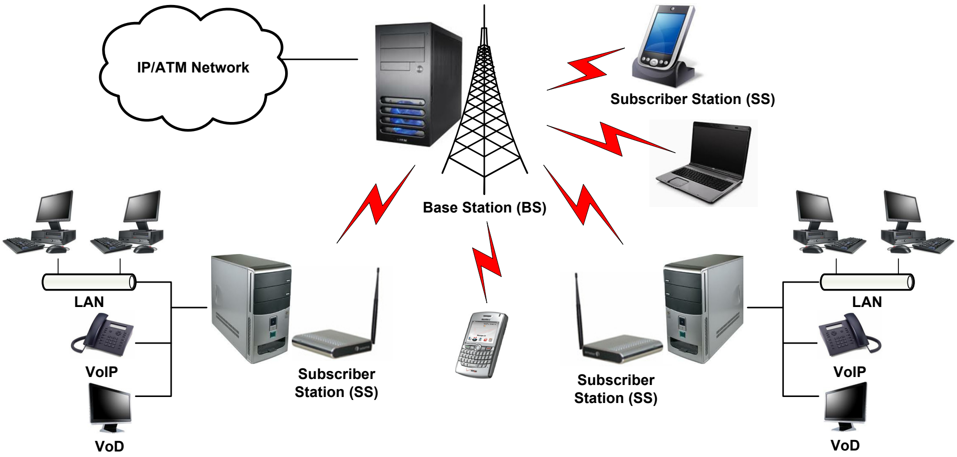


Figure 3.2 Core WWAN architecture.

The simplified TDD frame structure is demonstrated in Figure 3.3. Importantly, together with the data packets, the BS also communicates the relevant scheduling information for both DL and UL channels. The UL sub-frame schedule is incorporated into the UL map (UL-MAP) management message within the DL sub-frame header and is used by the SSs to determine the start time of their transmission in the UL. Enabling entry of new SSs into the system, a *ranging interval* is provided. Similarly, in order to allow the SSs to indicate their bandwidth demand to the BS, a *reservation interval* is provided. The SSs can send their ranging or bandwidth requests during transmission opportunities comprising the respective intervals. These requests are then processed by the BS.

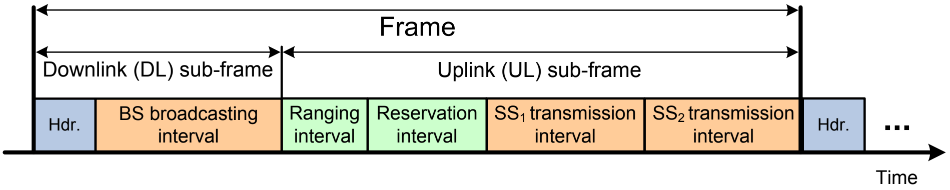


Figure 3.3 Simplified TDD frame structure.

IEEE 802.16 is known to successfully manage various multimedia connections [103]. It is equally suitable for both high data rate (VoIP, audio, and video) and low data rate (web browsing) applications. The protocol also supports bursty data flows and delay-sensitive traffic. In order to ensure that the diverse QoS requirements of all these applications are satisfied, IEEE 802.16 standard introduces five *QoS profiles*. Consequently, a data flow with a dedicated identifier is mapped onto one of the following profiles:

1. *Unsolicited grant service* (UGS). Used for real-time data sources with constant bit-rate (VoIP traffic without silence suppression).

2. *Real-time polling service* (rtPS). Used for real-time data sources with variable bit-rate (MPEG traffic).

3. *Extended real-time polling service* (ertPS). Used for real-time data sources with variable bit-rate, which may require more strict delay and throughput guarantees (VoIP traffic with silence suppression). This profile was introduced in one of the later specifications, IEEE 802.16e-2005.

4. *Non real-time polling service* (nrtPS). Used for non real-time data sources with variable packet length (FTP traffic).

5. *Best effort* (BE). Used for non real-time data sources, which do not require delay and throughput guarantees (HTTP traffic). This profile utilizes the residual bandwidth after the scheduling of all the above profiles has been completed.

The standardization of WWAN technologies is an ongoing activity by the IEEE 802.16 Working Group under the support of WiMAX Forum [104]. The UL data packet scheduler, which is out of scope of the IEEE 802.16 standard (see Figure 3.4), has a major impact on the efficiency of ensuring QoS requirements for the end clients. As a consequence, numerous research papers address the challenging problem of effective scheduling, such as [105], [106], and [107], where various frameworks are built and analyzed to provide a specified level of QoS. In particular, the work in [108] proposes a novel QoS architecture based on priority scheduling and dynamic bandwidth allocation. The paper [109] compares and contrasts the performance of various reservation disciplines in the framework of the simplified wireless system model. For a good summary on QoS in the context of IEEE 802.16 networks, we refer to the on-line document [110].

As follows from Figure 3.4, a particular *admission control* procedure is also undefined by the IEEE 802.16 specification. Designing efficient admission control methods is a separate research problem (see e.g., [111] and [112]) and is not addressed in this thesis.

In the scientific literature, the majority of known analytical frameworks do not account for the bandwidth *reservation* and packet *scheduling* components of the packet delay. However, the importance of considering both components to evaluate the overall delay of access-control systems was emphasized by a pioneering fundamental work [113], as well as by the earlier paper of the author [114]. For a more practical approach, we refer to [115], where the realistic performance measures of the IEEE 802.16 system are studied by various techniques. Some alternative polling techniques are investigated in [116]. We continue below with our own approach to jointly analyze the reservation and scheduling delays.

3.2.2 Overall packet delay analysis

Below we consider the two major components of the overall packet delay at the reservation and scheduling stages respectively. We define the *overall delay* (W_i) of the tagged packet as the time interval elapsed from its arrival into the outgoing

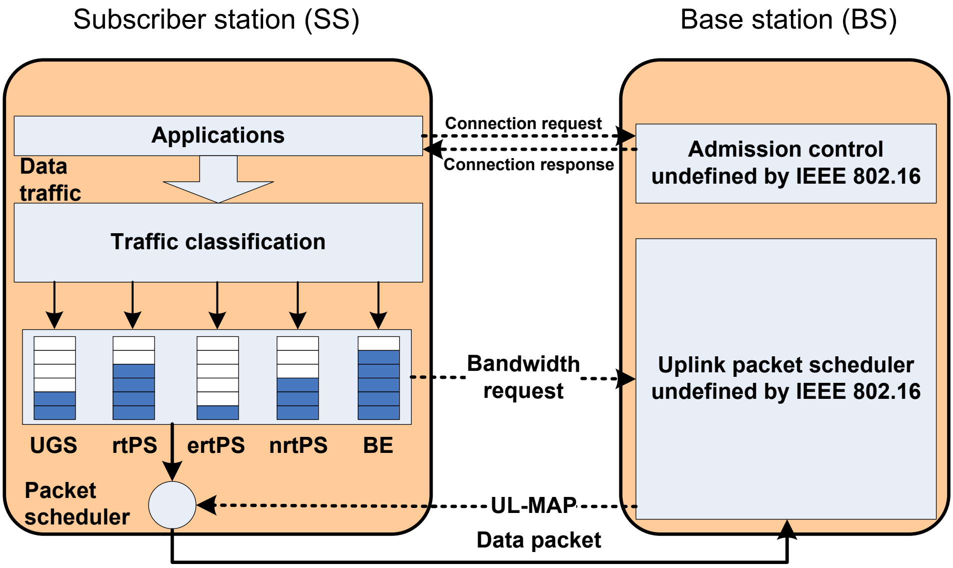


Figure 3.4 Simplified QoS framework.

buffer of the SS number i until the end of its successful transmission in the UL. It may be shown to be composed of several parts:

$$W_i = W_i^r + \alpha + W_i^s + W_i^t + \tau. \tag{3.1}$$

Here, W_i^r is the *reservation* delay, which is the time interval from the packet arrival at the SS i to the start of the corresponding bandwidth request transmission to the BS. Further, α is the duration of a successful bandwidth request transmission. We define the *grant time* of the tagged packet as the scheduling epoch of the frame preceding the one where the tagged packet has been transmitted. As such, W_i^s is the *scheduling* delay, which is the time interval from the end of the bandwidth request transmission corresponding to the tagged packet to its grant time. W_i^t is the *transmission* delay, which is the time interval from the grant time of the tagged packet to the start of its successful transmission in the UL sub-frame. Finally, τ is the data packet transmission time.

Our analysis in [P2] assumes that each SS has an individual buffer at the BS and a dedicated data packet transmission period in the UL sub-frame. As such, we demonstrate that the statistical behavior of an SS is independent of the operation by other SSs. Consequently, to establish the overall packet delay of the tagged SS it is sufficient to model its behavior separately from the rest of the system.

Accordingly, we study the entire wireless broadband system from the point of view of the tagged SS. In [P2], we construct a Markov chain embedded at the sequence of start times of the consecutive reservation intervals. The state of the chain corresponds to the number of packets in the SS and BS buffers. In particular, we assume that there exist *three* buffers for the data packets (see Figure 3.5). The first buffer is the one at the tagged SS where the packet is queued during the reservation delay. Further, at the beginning of the corresponding reservation interval the packet is immediately transferred to the virtual buffer. The *virtual* buffer accounts for the fact that a packet cannot be transmitted in the current frame, that is, experiences the delay of at least one frame. After this additional delay the packet enters the individual BS buffer of the tagged SS. There, the packet is queued until the end of the scheduling delay. Finally, the packet is transmitted.

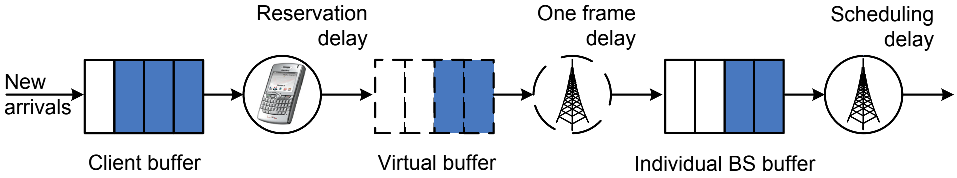


Figure 3.5 Overall packet delay evaluation model.

We then continue with obtaining the steady-state expressions for the mean numbers of packets in all the considered buffers. However, due to intricate internal correlations between the studied variables, in [P2] we only establish a closed-form *upper bound* on the overall data packet delay in the IEEE 802.16 network. As such, [P3] follows the same methodology and derives the *exact value* of the delay, but with a more complex numerical solution. Therefore, [P2] and [P3] together constitute a complete framework for the overall packet delay evaluation.

3.2.3 Advanced model for dynamic capacity allocation

Leveraging our experience on the analysis of IEEE 802.16 system accumulated by [P2] and [P3], we propose an extended analytical model [3] to perform efficient *dynamic* capacity allocation. Particularly, the non real-time (delay-tolerant) traffic of each SS can utilize a portion of the spare bandwidth remaining after the capacity allocation for the real-time (delay-critical) traffic flows at every SS. Thus, the model incorporates the effect of the capacity allocation for the rtPS, the ertPS, and the UGS flows on the overall delay of the nrtPS flow.

More specifically, the analytical model in [3] is applied to the performance evaluation of the UL nrtPS traffic in the IEEE 802.16-based network. Providing comprehensive numerical examples, we study the influence of the real-time traffic on the delay of the nrtPS service flow. We discuss how to take into account an upper bound on the mean delay of the nrtPS service flow at the SSs in determining the maximum sum of the real-time capacities at every SS. Finally, we introduce a cost model which considers the QoS constraints on some important parameters, including packet delay and real-time capacity. The various aspects of our performance analysis have potential applications in network control, since they facilitate the proper selection of the capacity parameters with respect to the requirements of the actual application scenario.

Based on our performance evaluation in [3], the following general conclusions can be made:

- The influence of the mean nrtPS packet delay on the total UGS capacity and on the maximum (e)rtPS capacity for the *uniform* distribution follow similar trends.

- The distribution of the (e)rtPS capacity has essential impact on the mean nrtPS packet delay.

The presented analytical model also enables enforcing specific upper bounds on the mean nrtPS packet delays at every SS in a given range of loads. In this case, the optimal value of the total real-time capacity can be determined.

3.3 QOS ASPECTS OF WLANS

3.3.1 General QoS architecture

The WLANs based on a series of IEEE 802.11 standards (or "The Standard for Wireless Fidelity," WiFi) is a rapidly growing technology worldwide. Given that this type of network is cheaper to deploy, it provides an effective low-cost solution to achieve ubiquitous wireless connectivity across various devices. There are important special cases of a WLAN, such as a *mobile ad-hoc network* (MANET) [117]. Generally, an *ad-hoc* network is a self-configuring network of devices connected wirelessly. The emerging IEEE 802.11 technologies (such as 802.11ac, 802.11ad, 802.11ah, etc.) motivate the need to further evaluate the network performance in order to support increasing client population and exploit scarce wireless resources even more efficiently. However, the lack of infrastructure complicates the MAC-level analysis of an ad-hoc network.

Typically, WLAN applications embrace two kinds of traffic: *unicast* traffic of point-to-point connections and *broadcast* traffic of point-to-multipoint connections. As these traffic types serve different purposes and typically coexist within a network, neither of them should be neglected at the performance evaluation stage. It is commonly believed that the overall population of contemporary WLANs is growing, whereas not all the network clients have similar necessities with respect to the channel resources. For this reason, traffic differentiation is crucial in modern WLANs. As such, a group of privileged clients (generating e.g., real-time traffic) might need to enjoy higher channel access probabilities than the other clients.

The contemporary IEEE 802.11-2012 [118] standard has merged several important amendments (including IEEE 802.11n-2009 for higher data rates) together with its previous version IEEE 802.11-2007 to specify the up-to-date features of the PHY and the MAC layer. At MAC, a contention-based protocol allows wireless clients to share channel resources through the use of a *carrier sense multiple access with collision avoidance* (CSMA/CA, see Chapter 2) mechanism based on the *binary exponential backoff* (BEB) collision resolution procedure. The carrier sense is performed by physical and virtual means. The *virtual* carrier sense scheme is typically implemented as an explicit exchange of two dedicated reservation frames, *request to send* (RTS) and *clear to send* (CTS) prior to the actual transmission of the data frames. These control frames specify a duration field within the frame header that contains the prediction of the ongoing busy period duration. The latter includes sending data, as well as receiving the respective *acknowledgment* (ACK) or *block ACK* (BA) frame.

The duration field is important to prevent the hidden users from accessing the channel while a transmission is in progress. A *hidden user* is a wireless client that cannot sense the sender but may influence the receiver if it starts communicating its data frame. In wireless environments, active clients are expected to constantly monitor RTS/CTS frames and use the duration field information therein to update their *network allocation vectors* (NAVs).

The *physical* carrier sense is known as *clear channel assessment* (CCA) and exploits the *signal to interference-plus-noise ratio* (SINR) to determine whether the medium is idle or not. Generally, a client shall refrain from transmission when either the NAV timer is active or the CCA senses the medium busy. However, if both carrier sense mechanisms indicate an idle channel, a *transmission opportunity* (TXOP) is obtained (see Figure 3.6) by the client. A TXOP may be regarded as a bounded time interval during which a sequence of packets encapsulated into the frames is transmitted by the source, while only service messages are received from the destination.

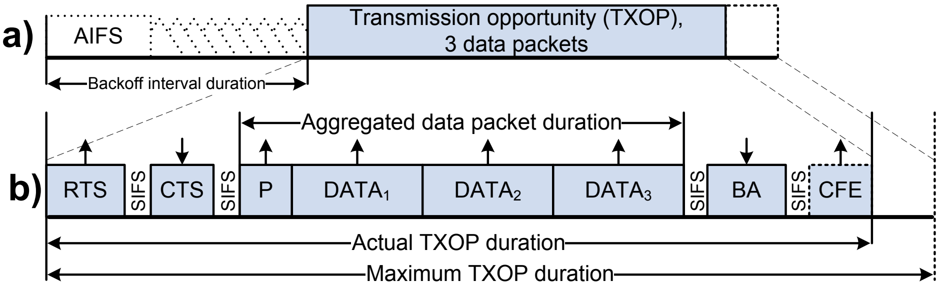


Figure 3.6 Typical TXOP structure: a) general, b) detailed.

Enabling traffic differentiation, *enhanced distributed channel access* (EDCA) replaced the legacy *distributed coordination function* (DCF) of IEEE 802.11 in order to provide improved QoS to wireless clients (see Figure 3.7). The QoS is handled via the *four* dedicated queues for different packet types. Every one of the queues, or *access categories*, has a separate BEB instance to control the medium access. The four different access categories are: *voice* (VO), *video* (VI), *best effort* (BE), and *background* (BK). Typically, the multimedia categories (voice and video) access the channel with higher priority. However, our literature survey in [P4] indicates that the influence of broadcast traffic on the performance of diverse (heterogeneous) client groups has never been studied. Therefore, it is important to propose a novel comprehensive analytical model that captures the heterogeneous IEEE 802.11 MAC saturation performance with both unicast and broadcast traffic.

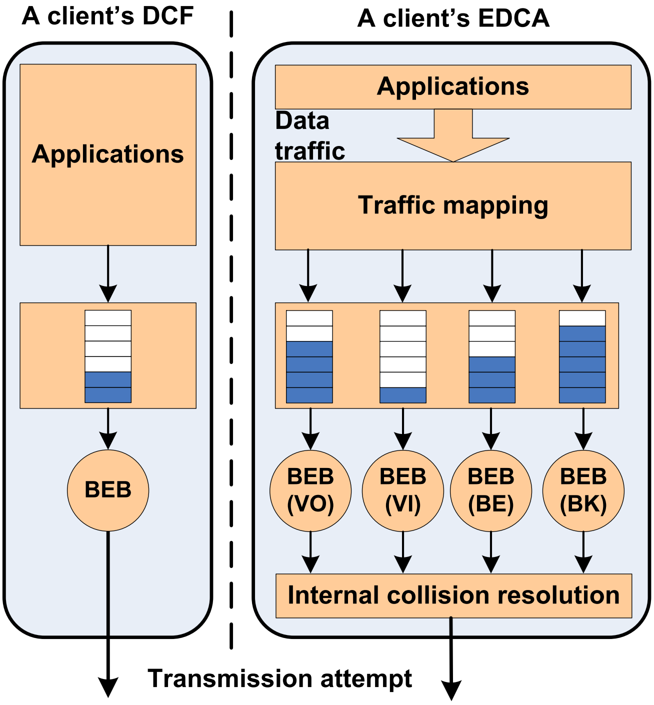


Figure 3.7 EDCA vs. DCF comparison.

3.3.2 Advanced model for a saturated WLAN

With an advanced model in [P4], we target realistic IEEE 802.11 performance evaluation that accounts for the mixture of traffic, heterogeneous groups of clients with different QoS requirements, and limited number of retries after a failed transmission. Our analytical approach is based on the concept of *regeneration cycles* and accounts for the average number of packet transmission attempts during such a cycle. Hence, a method to calculate the packet transmission probability p_t is as follows:

$$p_t = \lim_{n \to \infty} \frac{\sum\limits_{i=1}^{n} B^{(i)}}{\sum\limits_{i=1}^{n} D^{(i)}} = \frac{E[B]}{E[D]}, \tag{3.2}$$

where $B^{(i)}$ is the number of transmissions in a cycle and $D^{(i)}$ is the mean number of contention time slots for the i^{th} transmission attempt. Notice that $B^{(i)}$ and $D^{(i)}$ are independent and identically distributed with respect to i, but both are dependent on the conditional collision probability p_c. Here, $E[B]$ stands for the average number of transmissions in a cycle and $E[D]$ is the average number of slots that a client shall backoff before it starts its last transmission.

Further, considering a realistic *lossy* model when after k unsuccessful transmission attempts the packet is discarded and accounting for a limiting number of m BEB stages, the equation above modifies as:

$$p_t = \frac{E[B]}{E[D]} = \begin{cases} E[B] \cdot \Psi_1^{-1}, & k \le m+1 \\ E[B] \cdot \Psi_2^{-1}, & k > m+1 \end{cases} . \tag{3.3}$$

In [P4], we establish analytical expressions for the two components of $E[D]$, Ψ_1 and Ψ_2. We also assume that the network population is composed of G different groups with respective channel access probabilities. These groups are heterogeneous, whereas all clients belonging to a particular group j share similar parameters. Therefore, each group j has a respective transmission probability $p_t^{(j)}$ and, consequently, observes different conditional collision probability $\mathrm{p}_c^{(j)}$:

$$\mathrm{p}_c^{(j)} = 1 - \left(1 - p_t^{(j)}\right)^{M_j - 1} \prod_{i=1;i\neq j}^{G} \left(1 - p_t^{(i)}\right)^{M_i}. \tag{3.4}$$

The equation (3.4) explains how the conditional collision probability may be calculated for a generic group j. Note that a collision can be intergroup, intragroup, or both at the same time. An *intergroup* is a collision between the clients belonging to different groups. An *intragroup* collision is between the clients of the same group. The transmission probability $p_t^{(j)}$ is calculated as per (3.3) according to the properties of group j:

$$p_t^{(j)}(W_0 = W_0^{(j)}, m = m^{(j)}, k = k^{(j)}, p_b = p_b^{(j)}, p_c = \mathrm{p}_c^{(j)}), \tag{3.5}$$

where the core parameters of the model are summarized in Table 3.1.

Table 3.1 Saturation model parameters

Parameter	Description
M	Total number of clients in the system (M_j is the size of group j)
G	Total number of different client groups
W_0	Initial contention window value for the BEB protocol
k	Maximum number of transmissions that a unicast packet can tolerate. However, a broadcast frame is only sent once
m	The BEB protocol stage, which is the maximum number of possible increases for W_0
p_b	Probability of a broadcast packet generation
p_u	Probability of a unicast packet generation. Hence, it is complementary to p_b

Notice that (3.4) and (3.5) constitute a system of *non-linear* equations with a numerical solution. Once the transmission probability for each group is established, the system-wide probabilities can be calculated in order to obtain the overall saturation throughput. Therefore, in [P4] we proposed a simple and accurate approach to evaluate the main performance metrics, whereas currently more complicated techniques are typically used. The model was validated theoretically by proving the backwards compatibility to previous well-known models e.g., [119], as well as via extensive simulations.

3.4 COEXISTENCE OF WWAN AND WLAN

3.4.1 Possibilities for interworking

Contemporary metropolitan-scale wireless deployments often include areas where different communication networks are co-located [120], [121]. Following the trend for universality, the concept of the multi-radio device was introduced in [122] to allow for the simultaneous operation of different technologies. A multi-radio unit may operate in several wireless networks at the same time in accordance with the inbuilt protocols. However, several research problems arise to enable efficient heterogeneous functioning [123].

The problems caused by the multi-protocol operation at the MAC layer have not yet received much attention in the scientific literature. Some papers (see, for example, [124], [125], and [126]) cover IEEE 802.11 and IEEE 802.15.1 (Bluetooth) coexistence issues. Another case is IEEE 802.16 and IEEE 802.11 cooperation. Although these communication standards adopt drastically different MAC protocols, the capability of IEEE 802.11 reuse by IEEE 802.16 in the mesh mode was demonstrated in [127]. In [128], the general coexistence evaluation approach was shown and [129], particularly, addressed IEEE 802.16 and IEEE 802.11e interworking, where a concept of the base station hybrid coordinator was introduced. The use of such a coordinator is possible when the base station of IEEE 802.16 and the hybrid coordinator of IEEE 802.11e are co-located.

In this section, we also address the case of cooperation between IEEE 802.16 and IEEE 802.11 technologies (see Figure 3.8). But by contrast to the approach of [129], we consider a more realistic scenario without a central coordinating node in the communication system [P5]. Instead, the problem of the *MAC coordination* within a client *multi-radio* (MR) station itself is addressed, thus avoiding any restriction on the network topology.

Currently, IEEE 802.16 and IEEE 802.11 wireless networks operate in non-overlapping frequency bands [121]. Therefore, they may coexist simultaneously without any significant performance degradation. However, this is the case only when each client device supports the functionality of exactly one technology. When the capabilities of two or more standards are co-located within a single MR station, the client performance degrades dramatically [130]. This is explained by the fact that the radio parts of a device are in close proximity and the ongoing transmission in one network prohibits the reception by another one. Preventive solutions in a form of MAC coordination algorithms require a novel evaluation framework that takes into account mutual interference between wireless systems at a MR station.

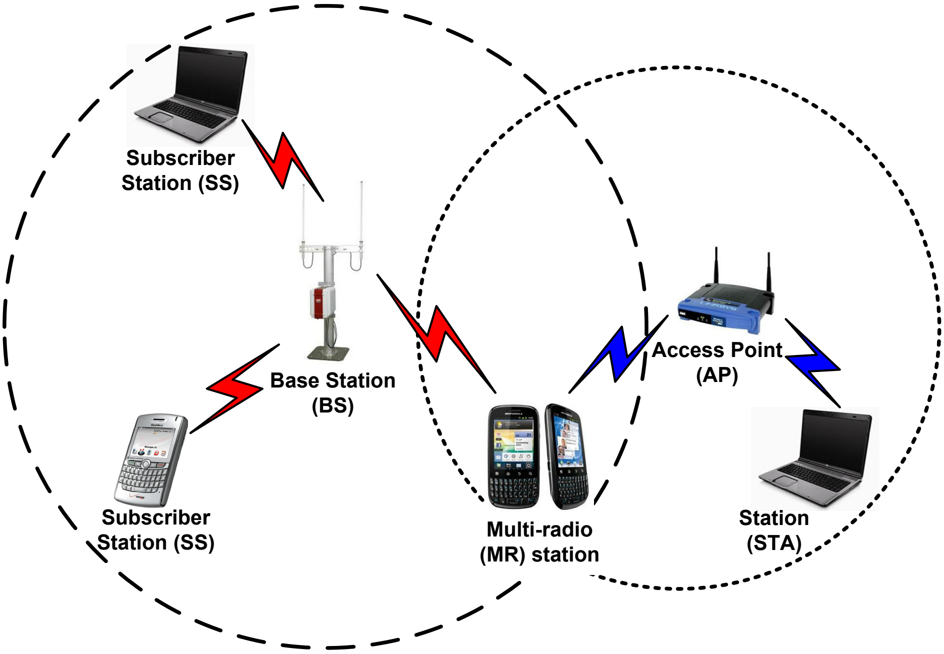


Figure 3.8 Coexistence between a WWAN and a WLAN.

3.4.2 MAC coordination solutions

Due to the multi-radio interference, when a station is receiving some data an overlapping transmission via the co-located technology prevents its successful reception. It was shown in [130] that 802.16 and 802.11 radio-to-radio interference severely degrades the performance and requires significant isolation of at least 55 dB. Increasing isolation is costly, large in size, and highly platform dependent. An alternative solution may be pursued in the time domain (see Figure 3.9). To mitigate the indicated effect, a special module on top of the respective MAC layers may be implemented for the purposes of the MAC coordination. This solution is known to be universal, effective, and medium independent.

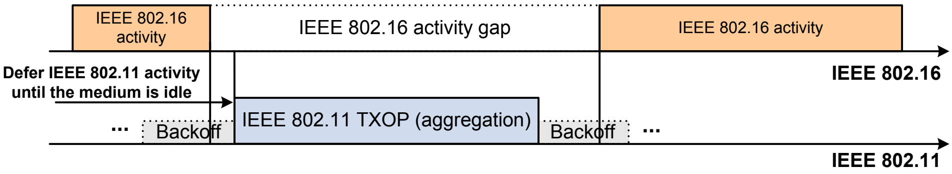


Figure 3.9 Principle of MAC coordination.

The MAC coordination module controls scheduling of both network activities within a MR station and thus enables the simultaneous operation of 802.16 and 802.11. As 802.16 system is schedule-based, the MAC coordination module only monitors its transmit and receive activity and allows/denies the channel access of the 802.11 part depending on the current 802.16 schedule. In [P5], we propose a range of coordination algorithms, which demonstrate performance-complexity trade-off and thoroughly analyze them. Given that the details of the introduced algorithms are covered by [P5], below we only list their main pros and cons denoted by "+" and "–" respectively.

1. Basic coordination algorithm:
   + Simple implementation.
   + Workability in case of both *shared* and *separate* antennas.
   – Constant *atomic* operation, some resource waste.
   – Usage of activity gaps only, non-maximum performance.

2. Enhanced coordination algorithm:
   + Dynamic atomic operation, enhanced performance.
   + Workability in case of both shared and separate antennas.
   – Higher computational and implementation complexity.
   – Usage of activity gaps only, non-maximum performance.

3. Suppressing enhanced coordination algorithm:
   + Simultaneous operation in both networks, better resource utilization.
   + Dynamic atomic operation, enhanced performance.
   – Workability in case of separate antennas only.
   – CCA suppression, highest computational and implementation complexity.

Our subsequent performance analysis of the introduced coordination algorithms [P5] is carried out within the framework of a simplified system model. We study the operation of the MAC coordination solutions and establish the respective *MAC goodput* of a MR station, which is defined as the portion of 802.11 PHY layer data rate available for the information transmission at its MAC layer.

For instance, the MAC goodput of the Basic coordination algorithm in case of a single TXOP per 802.16 frame G_1^B may be calculated as:

$$G_1^B = \frac{LQ_{max}}{T_{frame}} \cdot \Pr\{\tilde{T}_{tail} \leq T\}, \tag{3.6}$$

where the important parameters are summarized in Table 3.2 and T is the threshold value of the backoff interval *remainder* duration $\tilde{T}_{tail}$ that still results in one TXOP per 802.16 frame.

Generalizing, the MAC goodput of the Basic coordination algorithm in case of not more than two TXOPs per 802.16 frame G_2^B may be calculated as:

$$G_2^B = \frac{LQ_{max}}{T_{frame}} \cdot (1 + \Pr\{\tilde{T}_{tail} + \tilde{T}_{BO} \leq T\}), \tag{3.7}$$

where T is the threshold value of the backoff interval remainder duration that now results in two TXOPs per 802.16 frame.

Table 3.2 Coordination model parameters

Parameter	Description
T_{frame}	IEEE 802.16 frame duration
Q_{max}	Maximum number of packets within IEEE 802.11 TXOP
Q_{mod}	Maximum number of packets within modified IEEE 802.11 TXOP
L	IEEE 802.11 data packet length
$\tilde{T}_{tail}$	IEEE 802.11 tagged backoff interval duration that avoids overlapping with IEEE 802.16 header
$\tilde{T}_{BO}$	IEEE 802.11 backoff interval duration
$\tilde{Q}_{last}$	Number of packets within the last IEEE 802.11 TXOP per IEEE 802.16 frame

Correspondingly, the MAC goodput of the Enhanced coordination algorithm in case of not more than two TXOPs per 802.16 frame G_2^E may be calculated as:

$$G_2^E = \frac{L}{T_{frame}} \cdot (Q_{max} + E[\tilde{Q}_{last}]). \tag{3.8}$$

Finally, the MAC goodput of the Suppressing enhanced coordination algorithm in case of not more than three TXOPs per 802.16 frame G_3^S may be calculated as:

$$G_3^S = \frac{L}{T_{frame}} \cdot (Q_{max} + E[\tilde{Q}_{last}] + Q_{mod}) = G_2^E + \frac{LQ_{mod}}{T_{frame}}. \tag{3.9}$$

Summarizing, in [P5] we presented an efficient approach to enable the simultaneous operation of WWAN and WLAN technologies within a multi-radio client device. The MAC coordination concept was introduced and three various coordination algorithms were discussed that demonstrate the performance-complexity trade-off. Our analysis of the MAC goodput indicates significant performance gains of the proposed solutions. Moreover, some aspects of coexistence with other MR stations, as well as the mean packet delay simulations, were reported in [P5].

Chapter 4

Client Cooperation Techniques

4.1 INTRODUCTION AND MOTIVATION

4.1.1 General background

As more wireless clients are sharing limited spectrum in modern broadband communication networks and future cellular technologies are shifting toward aggressive full-frequency reuse [9], [10], the performance of wireless networks is becoming heavily impaired by interference. Since wireless communication is inherently broadcast, the transmission of a particular client may interfere with that of neighboring clients and thus reduce the energy efficiency. However, clients can recover their energy efficiency if cooperation among clients in close proximity is allowed. Therefore, *spatial* domain resource management is important to control the operation of wireless clients at different geographical locations [12]. However, client cooperation may require extra signaling overhead and additional energy consumption.

Previously, it was shown that with either transmitter or receiver cooperation significant energy savings may be achieved [131]. Furthermore, it has also been demonstrated that cooperation has the potential to reduce packet delay within certain transmission ranges due to the fact that sometimes it enables higher order modulation and increased data rate. Similarly, considerable energy savings may be obtained when receiver cooperation is properly exploited [132].

Since the energy required for reliable communication is known to grow exponentially with increasing distance between the participants [53], it is typically more energy efficient to transmit data over several shorter intermediate hops than via a longer hop [133]. However, client cooperation may incur excess energy consumption of the assisting relay nodes and, in some scenarios, it may still be preferable to keep using longer hops [134]. As such, the *relay selection* problem may be considered as a compromise between the performance of the data originator and the expenses of the assisting cooperator with respect to the overall system energy efficiency [12]. Below we mention some practical implications of this trade-off.

4.1.2 Cooperative communications

We remind you of the fact that wireless communication channels are known to suffer significantly from fading, which practically means that the attenuation of a signal varies over the transmission duration. An efficient solution to mitigate the negative effects of channel fading is via transmitting several copies of the initial signal independently, thus creating *diversity*. Particularly, by transmitting from different locations, so-called spatial diversity is created. It has the potential to improve signal reception quality as different copies of the same signal fade independently from each other before arriving at the receiver [135]. *Cooperative communications* receive increasing attention from the research community as a novel and practical way to create spatial diversity [136].

Historically, the core ideas behind *data relaying* were first introduced in the fundamental work [137], 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 [138]. 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 sometimes refer to as the “classical” relay channel.

Unlike the approach of [137] and [138], in cooperative communications both the data originator and the relay may act as information sources, which was considered in [139]. Accounting for practical decode-and-forward signaling, the paper [139] inspired many contemporary works on cooperation. With the recent proliferation of smart phones and machine-to-machine communications, wireless technology is steadily evolving toward 4G mobile systems. As such, a renewed surge of interest has arrived with rapidly expanding literature on cooperative communications. A good overview of this area may be found in [140].

The originator of a data packet may practically have more than one cooperator to partner with. The problem of relay selection is a rich topic and has been elaborated upon in [141] and [142], which assume the availability of a centralized controller in the network. Hence, the concept of cooperation was brought into the scope of wireless cellular networks with a base station controlling client activity. Further, in [143] a scenario is considered where the direct link from the originator is not available due to, for example, its poor quality. As such, the originator can only communicate its packets to the destination through the relays and [143] proposes an efficient protocol for the relay selection.

In [144], it is argued that relay-enhanced cellular architecture is a cost-effective way to achieve higher performance for end clients. More importantly, [144] suggests that wireless clients themselves may relay data for other clients thus acting as *mobile* relay stations [145]. Developing the novel concept of client relay, we seek to propose simple and feasible cooperation protocols, as well as to analyze the basic trade-offs behind client relay functionality.

4.1.3 Client relay in WWANs

While infrastructure relay technology [146], [147] has been standardized in e.g., IEEE 802.16j-2009 [148] and IEEE 802.16m-2011 [9], limited attention has been paid across WWAN protocols to the cooperation between wireless clients themselves. *Client relay* is more formally defined as the process by which one or more clients forward traffic between the originating client and the BS. As such, client relay has the potential to improve WWAN system performance by providing higher data rates while moderately impacting infrastructure complexity. Compared to infrastructure relay, client relay is also expected to be cheaper to deploy.

Client relay can technically be implemented in a homogeneous or heterogeneous manner. In *homogeneous* client relay (see Figure 4.1), all transmissions (i.e., client-to-BS, client-to-client) occur on the licensed bands (i.e., IEEE 802.16, 3GPP LTE). More specifically, homogeneous client relays forward traffic over the same DL and UL channels as their respective data originators. In *heterogeneous* client relay (see Figure 4.2), client-to-BS transmissions occur on the licensed bands, whereas client-to-client transmissions are carried out over the unlicensed bands (i.e., IEEE 802.11, Bluetooth). Although both flavors of client relay require the same basic functionality, the corresponding communication protocols are expected to be substantially different due to the centralized and distributed nature of the licensed and unlicensed bands respectively.

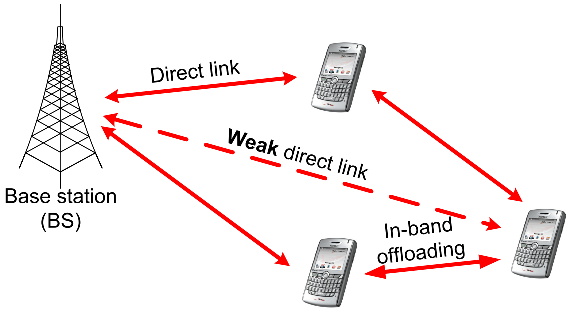


Figure 4.1 Homogeneous client relay.

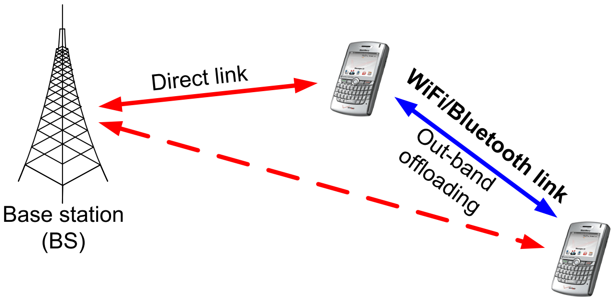


Figure 4.2 Heterogeneous client relay.

With the client relay technique, both DL and UL traffic can be relayed. However, given that clients typically operate at much lower powers than BSs, we expected that the majority of gains will be achieved for the UL traffic. Therefore, we only concentrate on the UL channel in what follows. Consequently, the UL client relay performance evaluation requires the derivation of an adequate wireless system model. As the integral research is complex, it might be divided into two components.

Firstly, analytical techniques can be used to derive a coarse system model. Queuing theoretic methods may be widely applied for this purpose. However, complex interactions between the queues during cooperation prohibit the direct use of known results for basic queuing models. Thus, an advanced model should be formulated to obtain throughput, energy efficiency, and packet delay.

Secondly, simulation may be exploited to account for numerous factors that influence the practical performance of client relays, such as realistic traffic arrival flows, predefined QoS parameters, wireless channel degradation factors, etc. In special cases, the simulation results should converge to those obtained with analytical modeling, which ensures the adequateness of the constructed simulation model. Simulation data would allow for many interesting insights into the cooperative wireless networking and bring important conclusions about the desirable application area of relay-enhanced technology.

4.2 HOMOGENEOUS CLIENT RELAY

4.2.1 Baseline triangle model

Motivated by the scenario from [143] and other related research works, we consider a wireless cellular network enhanced with client relay functionality in [P6]. Similarly to [149], we concentrate on the simplest network topology (see Figure 4.3) comprising two source nodes and one sink node. The node A is termed the originator and generates its own data packets with the mean arrival rate λ_A. The node R is termed the relay and generates its own data packets with the mean arrival rate λ_R. Additionally, the relay is capable of eavesdropping on the transmissions from the originator and may temporarily store its packets for the subsequent retransmission. The node B is termed the base station and receives data packets from both the originator and the relay.

The channel is error-prone and is based on the multipacket reception channel model from [150] and [151]. Once transmitted, a packet is corrupted with a constant probability which depends only on the link type and the number of transmitters. The basic parameters of the model are:

- $p_{AB} \triangleq \Pr\{\text{packet from } A \text{ is received at } B \mid \text{only A transmits}\}$
- $p_{RB} \triangleq \Pr\{\text{packet from } R \text{ is received at } B \mid \text{only R transmits}\}$
- $p_{AR} \triangleq \Pr\{\text{packet from } A \text{ is received at } R \mid \text{only A transmits}\}$
- $p_{CB} \triangleq \Pr\{\text{packet from } A \text{ is received at } B \mid \text{A and R cooperate}\}$

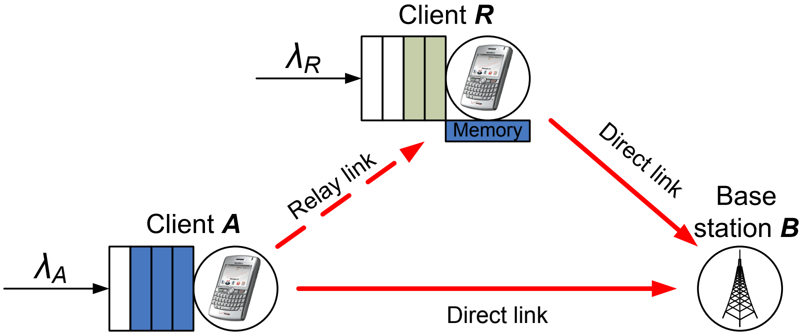


Figure 4.3 Triangle client relay model.

If a packet is corrupted, it is retransmitted by the source. The maximum allowed number of retry attempts is infinite. The nodes are equipped with single transceivers and thus cannot transmit and receive at the same time.

Upon the first transmission from the originator, the relay successfully eavesdrops on the packet with $p_{AR} > p_{AB}$. If the base station fails to receive this packet from the originator in the current slot with $1 - p_{AB}$ the relay stores it in the memory location for the eavesdropped packet.

Upon any retransmission from the originator, the relay performs one of the following operations. If the packet being retransmitted by the originator is already stored in the memory location, the relay transmits this packet *simultaneously* with the source [152] and the base station successfully receives the packet with $p_{CB} > p_{AB}$ due to the better quality of the relay link. Otherwise, the relay eavesdrops again on the retransmission of the originator and successfully receives the packet with p_{AR}. Once the packet from the originator is received successfully by the base station, the relay empties the memory location for the eavesdropped packets.

The proposed analytical approach to the performance evaluation of the considered three node client relay system is based on the notion of the packet service time. The *service time* of the tagged packet from a particular node starts when this packet becomes the first one in the queue of this node and ends when its successful transmission finishes. We denote the service time of a packet from node A as $T_{AR}(\lambda_A, \lambda_R) \triangleq T_{AR}$. Additionally, we introduce the mean service time of a packet from node A as $\tau_{AR}(\lambda_A, \lambda_R) \triangleq \tau_{AR} = E[T_{AR}]$. Further, we denote by $\tau_{AR}(\lambda_A, 0) \triangleq \tau_{A0}$ the mean service time of a packet from node A conditioning on the fact that $\lambda_R = 0$. Symmetrically, we introduce respective characteristics T_{RA}, τ_{RA}, and τ_{R0} for node R.

We established in [P6] that for both a system with cooperation (when $p_{AR} > 0$) and a system without cooperation (when $p_{AR} = 0$), it holds that $\tau_{R0} = p_{RB}^{-1}$, whereas only for the system without cooperation it holds that $\tau_{A0} = p_{AB}^{-1}$. The derivation of τ_{A0} for the system with cooperation is a more complicated task. Finally, we denote the queue load coefficient of node A as $\rho_{AR}(\lambda_A, \lambda_R) \triangleq \rho_{AR}$. By definition, we have $\rho_{AR} = \Pr\{Q_A \neq 0\} = \lambda_A \tau_{AR}$. In particular, the queue load coefficient of node A conditioning on the fact that $\lambda_R = 0$ may be established as

$\rho_{AR}(\lambda_A, 0) \triangleq \rho_{A0} = \lambda_A \tau_{A0}$. For the system without cooperation, ρ_{A0} further simplifies to $\rho_{A0} = {}^{\lambda_A}\!/_{p_{AB}}$. The queue load coefficients ρ_{RA} and ρ_{R0} of node R are introduced similarly. For both systems with and without cooperation ρ_{R0} further simplifies to $\rho_{R0} = {}^{\lambda_R}\!/_{p_{RB}}$.

Consider now the queue at node A and set $\rho_{A0} > \rho_{R0}$ as an example. The following propositions may further be formulated.

Proposition 1. For the queue load coefficient of node A, it holds:

$$\rho_{AR} \leq \frac{\rho_{A0}}{1 - \rho_{R0}}.$$

Proposition 2. For the queue load coefficients of nodes A and R, it holds:

$$\rho_{AR} - \rho_{RA} = \rho_{A0} - \rho_{R0}.$$

Proposition 3. For the queue load coefficient of node R, it holds:

$$\rho_{RA} = \rho_{AR} - \rho_{A0} + \rho_{R0} \leq \frac{\rho_{A0}}{1 - \rho_{R0}} - \rho_{A0} + \rho_{R0}.$$

Using the above definitions and propositions, in [P6] we establish important performance metrics of the client relay system. Firstly, our approach is applicable for determining the exact mean departure rate of packets from (throughput of) nodes A and R. Secondly, we study the behavior of node A within the framework of the queuing theory. Due to the fact that the queues of nodes A and R are mutually dependent, the notorious Pollaczek-Khinchin formula does not give the exact mean queue length of node A. We, however, apply this formula to establish the approximate value of the mean queue length of node A. Additionally, we obtain the exact values of the mean energy expenditure of nodes A and R.

4.2.2 Opportunistic cooperation

The relay improves the throughput of the originator by sacrificing its own energy efficiency. Extra energy is spent by the relay on the eavesdropping, as well as on the simultaneous packet transmissions with the originator. To save some of its energy, the relay may act opportunistically. As such, in each time slot the relay may decide not to eavesdrop on the transmissions from the originator with probability $1 - p_{rx}$ and/or not to relay a packet with probability $1 - p_{tx}$. The probabilities p_{rx} and p_{tx} correspond to a particular client relay policy and may be used to trade overall system throughput for total energy expenditure.

The client relay system operation is summarized by Algorithm 1. Accordingly, a single memory location for the eavesdropped data packets at the relay suffices for the considered client relay network operation. Importantly, the originator is unaware of the cooperative help from the relay and the relay sends no explicit acknowledgments to the originator by contrast to the approach in [151]. This enables tailoring the proposed client relay model to the contemporary cellular technologies [9], [10].

```
Generate new packets for $A$ and $R$ with $\lambda_A$ and $\lambda_R$
{Fair stochastic round-robin scheduling}
if both queues at $A$ and $R$ are not empty then
    Slot is given to either $A$ or $R$ with probability 0.5
else if queue at $A$ is not empty then
    Slot is given to $A$
else if queue at $R$ is not empty then
    Slot is given to $R$
else
    Slot is idle
{Packet transmission}
if slot is given to $A$ then
    if current packet from $A$ is not stored at $R$ then
        Packet from $A$ is successful at $B$ with $p_{AB}$
        if $R$ has decided to eavesdrop with $p_{rx}$ then
            Relay eavesdrops on the packet from $A$
            Eavesdropping is successful with $p_{AR}$
        else
            Relay stays idle
        if packet from $A$ is successful at $R$ then
            Packet from $A$ is stored at $R$
    else
        if $R$ has decided to cooperate with $p_{tx}$ then
            Relay transmits the stored packet
            Packet from $A$ is successful at $B$ with $p_{CB}$
        else
            Relay stays idle
            Packet from $A$ is successful at $B$ with $p_{AB}$
    if packet from $A$ is successful at $B$ then
        Relay empties its single memory location
    else
        Originator retransmits in the next available slot
else if slot is given to $R$ then
    Packet from $R$ is successful at $B$ with $p_{RB}$
else
    Slot is idle
```

Algorithm 1: Client relay network operation.

In [4], we extend the analysis of the baseline model from [P6] to take into account the opportunistic client relay behavior. As before, we focus on the primary performance metrics, including throughput, mean packet delay, and energy efficiency. Importantly, our model enables both opportunistic reception and transmission of relay packets, thus allowing for many useful insights into realistic network performance. We conclude that opportunistic behavior is a flexible tool to balance system spectral and energy efficiencies.

4.2.3 Coupling cooperation with power saving mode

Another important practical aspect of cellular networking discussed in Chapter 2 is that mobile devices are typically equipped with limited battery power. During time intervals when no packets are being sent or received, a wireless unit may switch to a power saving mode by shutting down its transmit and receive activities and thus save some of its power. Novel communication standards [9] and [10] support reductions in client power consumption through introducing various power saving mechanisms. Over the recent years, an extensive literature on power saving operation has accumulated.

The vast majority of research papers consider the *sleep mode* schemes as per IEEE 802.16 technology. The sleep mode within the legacy version of the standard, IEEE 802.16e-2004 [101], has been thoroughly studied in [153] and [154]. The approach therein is based on a queuing model with vacations to propose an optimization with respect to the packet loss probability. Another example of the legacy sleep mode performance evaluation is given by [155], where another queuing model with variably-distributed vacations is considered. The work also introduces a set of optimization solutions depending on which system parameters are known.

Conventionally, the arrival process of new data packets into the system is assumed to be Poisson. Therefore, the consideration of non-Poisson traffic may be of separate interest for researchers. In particular, the work in [156] concentrates on discrete time batch Markovian arrival processes and conducts a performance analysis of DL packet delay and energy consumption for the sleep mode. A more complicated operation for both DL and UL traffic was considered in [157] and has some contribution by the author of this thesis. Furthermore, the core trade-off between packet delay and energy efficiency was analyzed in [158]. Both papers adopt discrete-time queuing models to conclude on the efficiency of the sleep mode operation, whereas in [159] a continuous-time queuing model was exploited for similar purposes.

A more practical approach to the performance evaluation of sleep mode for the IEEE 802.16e-2004 standard and a comparison with its improved version in IEEE 802.16m-2011 standard may be found in [160], where the author of this thesis has also contributed significantly. Various QoS aspects associated with IEEE 802.16m sleep mode operation were addressed in [161] and [162] focusing on non real-time and VoIP traffic respectively. It was concluded that the novel sleep mode scheme may result in significant power savings for a wireless client. Later, this version of the sleep mode mechanism was also described analytically by [163] using a queuing model.

In 3GPP LTE and LTE-A, the equivalent power saving scheme is known as *discontinuous reception* (DRX, see Chapter 2). The operation of DRX has been subject to fewer research. In particular, [164] outlines a strategy to optimize the DRX parameters achieving higher power saving and resource utilization. The influence of the DRX parameters on client energy consumption and mean packet delay was studied in [165]. Some analytical results for DRX in case of bi-directional traffic were also presented in [166].

Our work in [5] couples power saving and cooperative client behavior. As the range of applications is very broad, rather than modeling one of the protocols in detail, we construct a high-level model capturing joint features of cooperation and power saving. As such, the developed model is intentionally broad-scoped, hence capturing only the key aspects of the considered system. Our model can be tailored to either the IEEE 802.16 sleep mode, or the 3GPP DRX mode operation. It allows for accurate performance evaluation and enables us to unveil less trivial trade-offs across all the potential scenarios, including homogeneous client relay.

4.2.4 System-level performance evaluation

Growing demand for bandwidth dictates the use of smaller wireless cells, which inevitably results in increased inter-cell interference. In most contemporary cellular systems, the clients at the cell edge typically use higher transmission power to compensate for increased path loss and fading and thus generate the most interference. As discussed previously, client relay is believed to be a promising technique to enhance the performance of cell-edge clients by allowing them to exploit other clients as relay nodes and thus transmit with less power.

In [6], we conduct an in-depth system-level evaluation of homogeneous client relay for the state-of-the-art wireless cellular networks. Several important features are considered, including realistic client relay operation and practical channel models. In particular, we address opportunistic client relay behavior within the context of interference, capacity, and energy efficient resource management. It is demonstrated that client cooperation may considerably improve system performance in terms of cell-edge spectral efficiency for the cost of some increase in cell-center energy consumption.

Studying the relaying strategies in [6], we have designed a simple and elegant approach that regards the cooperative network as a “virtual” multiuser *multiple-input and multiple-output* (MIMO) system [167], where the inputs are transmitting mobile terminals, and the outputs are BS antennas and other terminals. Wireless nodes may eavesdrop on the transmissions from other nodes and then communicate the previously intercepted packets together with their originators upon a retransmission, thus improving the packet reception probability. Since the clients are well-spaced and their channels are uncorrelated, there is a distinct transmit diversity gain which could be also converted into a MIMO gain by using multiple antennas at the BS. Naturally, if more nodes join the relaying the gain is higher.

The overall system model in [6] is based on IEEE 802.16m evaluation methodology [73], but some simplifications are adopted to reduce the computational complexity. The nodes are assumed to be randomly roaming around the associated BS. Generally, a comprehensive wireless channel model should take into account many real-world effects: distance, power, bandwidth, landscape, random fluctuations of the link parameters over time, etc. In our system-level evaluations, we use a widely-accepted path loss model [73] and also develop an empirical model [168] to enable the eavesdropping capability by mobile terminals. Fast fading is described by a thoroughly validated Rayleigh fading process. Slow fading process, however, is non-trivial due to extra client-to-client links. Therefore, we extend the well-known model [73] using the approach in [169].

Our research in [6] allows us to conclude that the proposed client relay technique is feasible within the IEEE 802.16m signaling without significant modification of the baseline protocol. Considerable gains in throughput and delay performance for the cell-edge clients as well as for the entire cell can be observed. Trade-offs between throughput, delay, and energy efficiency make our client relay approach a promising concept for improving performance of next-generation wireless networks. Since our methodology is built upon OFDM, it may be applicable for alternative OFDM-based technologies, including 3GPP LTE-A.

4.3 HETEROGENEOUS CLIENT RELAY

4.3.1 WWAN traffic offloading

According to many predictions, the proportion of traffic transmitted over wireless broadband networks is expected to grow considerably in the very near future [170]. Consequently, the currently deployed WWAN technologies are very likely to face serious overloads [171] resulting in a dramatic degradation in the levels of quality of experience for their end clients.

One solution to mitigate the increasing disproportion between the client QoS and the available wireless resources might be by deploying additional serving BSs [172]. However, introducing a higher density of BSs may not solely be sufficient to bridge this gap primarily for the following reasons [173]. Firstly, the inter-cell interference can grow substantially for aggressive frequency reuse patterns, thus preventing reliable communication. Secondly, due to necessary extra equipment, the rental fees are likely to increase, yielding respective difficulties in obtaining permission from regulatory authorities. Finally, the required operator maintenance costs are predicted to skyrocket, which consequently burdens the subscriber with the associated expenses.

In light of the above, it may be feasible to offload some of the WWAN traffic by enabling wireless clients (or peers) to communicate *directly* without changing much of the core network topology [174]. Where appropriate, client-to-client links are believed to become an effective solution that would relieve congestion in next-generation mobile networks [175]. By contrast to conventional data transmission over a WWAN technology (e.g., IEEE 802.16, 3GPP LTE), alternative direct communication between two or more wireless devices in close proximity is becoming attractive across the increasing number of potential use cases and scenarios. The latter promises considerable savings in terms of occupied radio resources [176] (both in DL and UL), thus providing higher QoS levels for end clients.

4.3.2 Device-to-device technology

Currently, a plethora of short-range wireless technologies exists to allow direct *device-to-device* (D2D) connectivity primarily in the form of *unlicensed band* solutions, such as WiFi Direct, Bluetooth, etc. There are, however, several serious drawbacks that limit potential benefits associated with this type of D2D communications [177]. Firstly, shorter transmission ranges may preclude the interacting peers from reliable communication, particularly when they are moving apart at pedestrian speeds. Secondly, existent WLAN/WPAN transceivers are expected to require an excess amount of energy, which may be prohibitive for small-scale battery-powered devices especially when the D2D capability is enabled for longer periods of time. Finally, device authentication procedures may become a concern, sometimes even requiring personal subscriber intervention.

The heterogeneous client relay technique introduced in the course of this chapter may prove to be really useful by mitigating the above limitations and providing a solid foundation for the reliable D2D technology. By seamlessly offloading WWAN traffic onto unlicensed band D2D links, we expect significant improvements in overall system capacity, as well as in client throughput and energy efficiency [178]. Importantly, the network may assist its clients by controlling the process of data offloading to maximize the achievable gains [177]. Furthermore, modification of existing standards (e.g., subsequent releases of 3GPP LTE-A) may be pursued to capture the basic client relay protocol mechanisms and some initial steps in this direction are mentioned in the remainder of this text.

4.3.3 Market potential and standardization

As discussed above, the D2D functionality is currently available only through the range of conventional unlicensed band technologies. Despite their huge commercial success, these solutions may suffer from session continuity limitations, excessive power consumption, and manual security procedures. Furthermore, the QoS performance of uncoordinated short-range technologies is limited by the lack of centralized resource management, which could otherwise facilitate peer discovery and selection [178]. To leverage the available gains and minimize the time to market, several companies are pushing to launch their proprietary *licensed band* technologies. One recent example of the in-band D2D solution is the FlashLinQ technology by Qualcomm [179], [180].

Understanding the significant market potential behind a standardized cellular-assisted D2D technology, 3GPP is becoming increasingly active on this topic. An important first step toward ubiquitous D2D connectivity is to indicate promising scenarios and usage models, as well as to formulate common requirements across the selected use cases. Consequently, there have been fruitful discussions inside the “Services” group of 3GPP within the respective Study Item on proximity-based applications [181]. When the preliminary work is completed, the “Radio Layer 1” group is expected to take over and initialize a follow-up Study/Work Item with respect to potential technology solutions for the future releases of the 3GPP LTE-A standard. Therefore, heterogeneous client relay appears to become a hot topic for beyond 4G wireless systems, building on the current 4G communication technologies, such as 3GPP LTE-A and IEEE 802.16m.

Chapter 5

Machine-to-machine Communications

5.1 INTRODUCTION AND MOTIVATION

5.1.1 General background

According to [182], *machine-to-machine* (M2M) communications may be defined as information exchange between a *subscriber* and a *server* in the core network through a base station which may be carried out without any human interaction. As such, M2M communications is a very distinct capability that enables the implementation of the *Internet of Things* (IoT). Industry reports indicate the considerable potential of this market, with millions of devices connected within the following years resulting in predicted revenues of $300 billion [183].

Generally, the IoT refers to the technology trend where things (e.g., everyday objects, locations, vehicles) are extended with sensors, RF identifiers, actuators, or processors, made discoverable and enabled to communicate with, and are closely integrated with future Internet infrastructure and services [184]. According to recent predictions, there will be on the order of 7 trillion such connected electronic devices for 7 billion people by 2020 [185], which would amount to around a thousand devices for every human.

Due to its huge market potential, several cellular technologies are now focusing on developing air interface enhancements to support M2M communications [186]. For example, emerging IEEE 802.16p [187] proposals address enhancements for the IEEE 802.16m standard to support M2M applications. 3GPP LTE also has several work items defined on M2M communications [188], primarily with respect to *overload control* [189], [190].

5.1.2 Core M2M usage models

The IEEE 802.16p M2M study report [182] covers several primary M2M use cases. Below we review these use cases according to [182].

1. **Secured Access and Surveillance.** This category features M2M applications that help prevent the theft of vehicles and insecure physical access into buildings. Equipping buildings and vehicles with M2M devices enables forwarding data to the M2M server in real time if movement is detected. Whenever car tampering or building intrusion has occurred, an alert signal is sent to the M2M subscriber.

2. **Tracking, Tracing, and Recovery.** The respective use cases are mainly related to services that rely on location-tracking information. In particular, vehicles are equipped with M2M devices that send status data (e.g., location, velocity, local traffic) periodically or on-demand to the M2M server. The collected information is then analyzed by the M2M server and provided to M2M subscribers via the cellular network. The examples of the emerging vehicular tracking services are navigation, traffic information, road tolling, automatic emergency call, pay as you drive, etc. The distinct feature of these scenarios is that the M2M application server needs to monitor the status and position of an individual vehicle or group of vehicles.

3. **Public Safety.** This category includes emergency response, public surveillance systems, and monitoring the environment. Depending on the requirements, M2M devices in relevant scenarios may report information to the M2M server either periodically or on-demand. In emergency response systems enhanced with WWAN M2M connectivity, public surveillance equipment may transmit real-time video to mobile police and fire teams. Furthermore, it can also be used by incoming ambulances to inform the receiving hospital staff. Finally, WWAN M2M has the potential to secure individuals in remote or high risk areas, as well as offenders under parole.

4. **Payment.** WWAN M2M communications enable a higher degree of flexibility in deploying point-of-sale/ATM terminals, parking meters, vending machines, ticketing machines, etc. Whereas providing better functionality, faster service, and simpler management, M2M also allows payment facilities to overcome a lack of wired infrastructure.

5. **Healthcare.** These applications are meant to improve both patient monitoring/tracking and doctor responsiveness. In particular, healthcare M2M services help patients with advanced age, chronic diseases, or complicated physical conditions to live independently. With more accurate and faster reporting of changes in patients' physical condition, they also considerably improve patient care. For instance, M2M devices may communicate with the healthcare management system in a hospital or a care facility and forward the patient's health information at regular periods or on demand. Furthermore, the M2M-capable healthcare management system may provide remote patient monitoring.

6. **Remote Maintenance and Control.** This category includes applications primarily used in oil, gas, water, waste, power, and heavy equipment industries. WWAN M2M services inform owners/companies of how their equipment is running and if there are signs of trouble. Providing timely information, automatic alarms, notification of consumption, and secure remote service access, M2M-enabled infrastructure may significantly improve the efficiency of remote maintenance and control services.

7. **Metering.** These services meter gas, electricity, or water consumption and send the remote meter readings to the customer. Smart metering may also improve the customer's energy/utility efficiency by regulating home appliance usage according to time-varying unit price. In particular, an M2M-enabled smart meter may collect utility usage information from home appliances and send it to the M2M server. Alternatively, a smart meter may communicate to an M2M device, which aggregates the information from many smart meters in the area and forwards the aggregated information to the M2M server.

8. **Consumer Devices.** In this market, WWAN M2M connectivity allows personal navigation, automatic e-reader updates, remote photo storage for digital cameras, as well as various netbook services. Furthermore, M2M technology facilitates content and data sharing among devices via user-friendly interfaces.

9. **Retail.** In this category, a WWAN M2M use case with considerable market potential is digital signage, which includes applications such as digital billboards along roads and highways. These billboards may receive new display information updates from the M2M server according to the M2M service consumer needs.

5.1.3 Requirements and features for M2M

In what follows, we report features described by [182] that are common to one or more M2M use cases listed above. It shall be possible to subscribe to different M2M requirements or features independently according to the application or network environment.

- *Extremely low power consumption:* the M2M devices should consume very low operational power over long periods of time. This feature is critical for battery-limited M2M devices.

- *High reliability:* whenever and wherever M2M communications are required or triggered, the connection and reliable transmission (in terms of extremely low packet error rate) between the M2M device and the M2M server should be guaranteed regardless of the operating environment (e.g., mobility and channel quality).

- *Enhanced access priority:* the M2M device may be given priority over other network nodes when competing for network access. Priority access is necessary to efficiently communicate alarms, emergency situations or any other events that require immediate attention.

- *Handling transmission attempts from an extremely large number of devices:* simultaneous or near simultaneous network entry attempts may introduce a surge at the serving BS.

- *Addressing an extremely large number of devices:* the system may need to address large numbers of devices individually.

- *Group control:* the system may support group addressing and handling of M2M devices.

- *Security:* the system should support security functions for M2M service traffic, including integrity protection and confidentiality. A WWAN M2M system should also ensure an appropriate level of authentication for the M2M devices to provide secure access to the authorized M2M devices. Furthermore, the system should guarantee verification and validation of the exchanged data.

- *Small burst transmission:* transmitted data bursts may be extremely small in size. The system should support efficient transmission of small data bursts with very low overhead.

- *Extremely low/no mobility:* the M2M device may be stationary for very long periods of time, perhaps throughout its entire lifetime, or move only within a certain region. The system can simplify or optimize the mobility-related operations for specific M2M applications with a fixed location.

- *Time-controlled operation:* the absence of ad-hoc packet transmissions. Consequently, the system can support an operation mode in which the M2M device transmits or receives data only at a predefined period of time.

- *Time-tolerant operation:* the system can provide a lower access priority to or defer the data transmission of time-tolerant M2M devices.

- *One-way data traffic:* data transmission may only be one-way, i.e., only device-originated data or only device-terminated data.

- *Extremely low latency:* the significantly reduced network access (data transmission) latency for specific M2M devices may be required.

- *Infrequent traffic:* M2M transmissions may be infrequent with large amounts of time between transmissions.

- *Extremely long range access:* a single WWAN M2M-enabled base station should serve M2M devices over a very long range. This is not necessarily a feature of any use case, but of some potential market cases that require extremely low-cost deployments.

As a summary, Table 5.1 matches the aforementioned requirements and features with the relevant M2M usage models. Whereas several important features are necessary for every category, some requirements may be relaxed for particular applications.

Table 5.1 Requirements matched with usage models

Requirement	Secured Access	Tracking, Tracing	Public Safety	Payment	Healthcare	Remote Maintenance	Metering	Consumer Devices	Retail
Low power	+	+	+				+		
High reliability	+		+	+	+	+	+		
Access priority	+		+		+	+			
Large number of devices	+	+	+		+	+	+		
Addressing	+	+	+	+	+	+	+	+	+
Group control	+	+	+	+	+	+	+	+	+
Security	+	+	+	+	+	+	+	+	+
Small bursts	+	+	+	+	+	+	+	+	+
Low mobility	+		+	+		+	+		+
Time-controlled	+	+	+	+	+	+	+	+	+
Time-tolerant	+	+	+	+	+	+	+	+	+
One-way traffic			+				+	+	+
Low latency	+	+	+	+	+	+			
Infrequent traffic	+	+	+	+	+	+	+	+	+

5.2 METERING USE CASE ANALYSIS

5.2.1 Smart metering and smart grid

As follows from the above description, smart metering is a key M2M use case that involves meters autonomously reporting *usage* and *alarm* information to grid infrastructure to help reduce operational cost, as well as to regulate a customer's utility use based on load-dependent pricing signals received from the grid [7]. Providing certain utilities (such as electricity, power, and gas) with communication capability may help automate grid operation and thus improve its efficiency, cost, robustness, and security. We therefore expect that WWAN technologies, such as IEEE 802.16 and 3GPP LTE, will play a pivotal role in enabling smart metering applications [191].

Currently, there are several ongoing efforts directed at outlining the scope and the requirements of emerging *Smart Grid* applications [7]. Together with big industry players, there is growing attention from governmental organizations worldwide primarily due to increased environmental concerns, as well as associated security challenges [186]. Targeting improved interoperability and ubiquitous coverage of Smart Grid services, the urge to standardize the relevant Smart Grid interfaces is really strong. The overview in [7] reveals a significant interest from many utilities and equipment manufacturers, governmental institutions, as well as research and standards bodies, which include the US Department of Energy, National Institute of Standards & Technology, Electric Power Research Institute, UCA International Users Group, Open SG, and other international standards development organizations, such as IEEE P2030, ETSI, etc.

Not limited by its importance for a critical industry, the Smart Grid use case also serves as a valuable reference M2M scenario [192], [193] covering many M2M features described in the IEEE 802.16 M2M study report [182] and summarized above. In particular, the report covers several M2M use cases under the broader categories of Metering, Secured Access and Surveillance, Remote Maintenance and Control, and to a limited extent under Tracking, Tracing, and Recovery. We contributed the informative text in [7] to supplement the feature description in [182], as well as to add some details on specific areas where IEEE 802.16 is applicable for Smart Grid, on the traffic characteristics across Smart Grid applications, and on the key challenges in supporting Smart Grid applications with IEEE 802.16 protocols.

5.2.2 Supporting large population of M2M devices

As mentioned in our contribution [7], it is important to estimate the typical number of M2M devices across a Smart Grid deployment, as a key challenge for WWAN networks will be supporting access from a large number of smart meters [194]. Whereas the number of devices per sector depends both on smart meter density and cell size, a range of estimates may be obtained by considering several example deployment options [7]. Given our estimates, there are several critical scenarios that may result in near simultaneous network entry attempts from a large number of metering devices, such as:

- Alarm reporting by a large number of smart meters when they access the network in an uncoordinated manner.

- A large number of devices deactivating after a massive power outage event. In particular, the smart meters are typically required to transmit a “last gasp” alarm informing the network that they have run out of power.

- A surge in network access requests when devices attempt to reestablish their connection after a power outage event.

- Periodic usage reporting by a large population of devices when short reporting intervals are used. Typically, regular communication to and from meters is infrequent, but more aggressive rates may be applicable in the future.

The above scenarios imply that it is of high importance to conduct a performance analysis of a WWAN deployment where a large population of M2M devices connects nearly simultaneously to the wireless infrastructure. In the following section, we summarize our results in this direction.

5.3 NETWORK ENTRY BY LARGE NUMBER OF DEVICES

5.3.1 Access success rate and latency analysis

In what follows, we address a typical smart metering M2M application scenario in the context of an IEEE 802.16 (3GPP LTE) wireless cellular system which features a large number of devices connecting to the network. To highlight an important issue of near simultaneous network entry by many devices, our contribution [7] calculates the number of initial *ranging* opportunities supported by the IEEE 802.16m technology [9] to process initial network entry requests. Summarizing, we conclude that a significantly higher number of transmission opportunities is required if near simultaneous access by a large number of smart meters needs to be supported.

In [8], a more detailed analysis of network entry success rates with the IEEE 802.16m initial ranging protocol is given. The analysis therein is extended to capture the binary exponential backoff procedure applied for connection establishment in IEEE 802.16 systems. Performance evaluation is conducted under various assumptions on the number of devices and the arrival rate of random access attempts. The results are then analyzed to refine target requirements for the emerging IEEE 802.16p amendment.

Generally, our findings indicate that *access success rates* can be dramatically lowered and *access latency* is increased substantially when there is a surge in near simultaneous connection requests by a large population of smart meters [8]. Therefore, the system should ensure that the QoS levels of high-priority traffic, as well as the performance of non-M2M devices are not adversely impacted by a large number of uncoordinated network entry attempts.

5.3.2 Network entry delay recovery

Our initial performance analysis indicates that network entry delays of M2M devices may be prohibitively high when there is a surge in near simultaneous network connection attempts by a large number of metering devices [8]. Such a surge in network access attempts may occur, for example, in a *power outage* scenario where a large number of smart meters attempts to connect to the network reporting the outage event and when they reconnect again upon the restoration of power. Our recent contributions to IEEE 802.16p [7], [8] characterize the *network overload* resulting from such alarm events. Therefore, some preventive measures are required to recover network entry delay values and thus improve network performance for a large population of M2M devices.

To propose an effective solution, [P7] thoroughly studies a combination of a *successive interference cancellation and a tree algorithm* (SICTA). In the course of our analysis, we focus on the algorithm performance and account for a single signal memory location, as well as for *cancellation errors* of different types. The resulting scheme is provably robust to imperfect interference cancellation. Therefore, we propose to consider this scheme as an alternative collision resolution algorithm replacing binary exponential backoff (BEB, see Chapter 3) protocol within IEEE 802.16.

Generally, the combination of *successive interference cancellation* (SIC) at PHY and tree algorithms at MAC constitutes a promising direction in improving the contemporary communication protocols. Currently, the family of SIC-based algorithms is known, where the baseline SICTA demonstrates the highest achievable throughput [195]. However, SICTA requires unbounded signal memory at the receiver side, which is practically infeasible. Moreover, its performance degrades significantly due to the imperfect interference cancellation.

In [P7], we propose a practical SICTA modification that mitigates the limitations of the baseline SICTA and reaches attractive throughput levels in case of no cancellation errors. Moreover, our *robust* SICTA (R-SICTA) is workable even in case of high cancellation error probability and demonstrates graceful performance degradation. These features make R-SICTA suitable for the UL bandwidth requesting and/or initial network entry in the IEEE 802.16 standard (see Chapter 3). Remember that bandwidth requesting and network entry procedures are both contention-based and IEEE 802.16 adopts the BEB scheme to resolve arising collisions. Replacing BEB with a SIC-based algorithm, network entry delay may be significantly decreased, which is critical for a large number of M2M devices accessing the wireless system.

We evaluate the gain after such replacement in [P7]. As expected, the baseline SICTA algorithm with unbounded signal memory demonstrates the lowest possible delay. However, our proposed R-SICTA algorithm with single signal memory performs closely to SICTA in case of no cancellation errors. Overall, our findings indicate significant delay gains after the replacement of the standardized BEB algorithm with the more practical R-SICTA. As such, the proposed solution is attractive to improve the performance of future M2M-aware cellular networks.

5.3.3 Some features of SIC-based algorithms

Conventionally, when collision resolution algorithms are considered, it is typically assumed that whenever packets collide (or interfere wirelessly), the receiver extracts no meaningful information. Recent advances at the PHY layer (see Chapter 2) enable the application of successive interference cancellation techniques to improve performance. SIC may be naturally used in the UL channel of contemporary cellular networks (IEEE 802.16m, 3GPP LTE-A), as a common receiver (e.g., base station) facilitates its operation.

A novel approach that tailors SIC to a tree algorithm was first proposed in [195]. Summarizing, SIC processes the previously stored collision packets (signals) and takes advantage of unbounded signal memory. Following [196], we show how SIC may improve the performance of a tree algorithm in Figure 5.1. Assume for simplicity that the channel is error-free. Denote by y_s the signal received by the end of slot s. Similarly, denote by x_A and x_B the signals corresponding to packets A and B respectively. Let two clients transmit their packets A and B in the first slot and collide. As such, the receiver acquires the combined signal $y_1 = x_A + x_B$ and decides that a collision occurred. The initial combined signal y_1 is then stored in the signal memory of the receiver.

After acquiring the signal $y_2 = x_A$ at the end of slot 2, the receiver successfully extracts signal x_A and decodes packet A. Further, SIC procedure processes signal

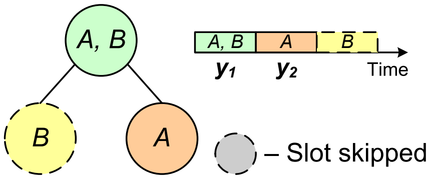


Figure 5.1 Example SIC operation.

y_1 and cancels the extracted signal x_A from the stored combination, that is, $\tilde{y}_1 = y_1 - x_A$. Then it is also possible to extract signal $x_B = \tilde{y}_1$ and to decode packet B. Therefore, the subsequent collision resolution is not necessary. In the considered example, the duration of the collision resolution interval is one slot less than for any known tree algorithm.

Maximum stable throughput is one of the most important performance parameters of tree algorithms. It may be defined as the highest arrival rate (typically, *normalized* with respect to the slot/frame duration), which still results in the bounded value of the mean packet delay. Generalizing, SICTA throughput is given by:

$$R_S \approx \ln 2 \approx 0.693, \tag{5.1}$$

which gives the previously known result from [195]. However, our approach in [P7] is far simpler and has reduced computational burden.

Whereas it is practically infeasible, the availability of infinite memory allows the original SICTA to double the performance of the standard tree algorithm (with the throughput of 0.346) proposed in [197] and [198] independently. This promising idea was taken further and a variety of SICTA modifications were proposed which may be classified into two categories. Firstly, there are algorithms that assume perfect SIC operation and therefore are susceptible to cancellation errors falling into a deadlock. Secondly, there are algorithms that are robust to imperfect SIC operation, but at the same time are unstable when the number of clients grows unboundedly. Our R-SICTA algorithm in [P7] tolerates cancellation errors and demonstrates nonzero performance even when the client population approaches infinity.

Analyzing the performance of R-SICTA, we differentiate between three probabilities of imperfect interference cancellation: q_{ce}, q_{cs}, and q_{ss}. We note that these probabilities are the parameters of SIC and depend on its implementation. However, we obtain an approximation for the throughput of the proposed algorithm depending on these probabilities as:

$$R_{RS} \approx \frac{2\ln 2}{2 + q_{ce} + \ln 2(1 + q_{cs} - q_{ce} + 2\gamma(q_{ss} - q_{cs}))}, \tag{5.2}$$

where $\gamma = 0.721$ and was calculated in [P7].

In particular, when $q_{ce} = q_{ss} = q_{cs} = 0$, that is, when there are no cancellation errors, $R_{RS} \approx 0.515$. We conclude that the proposed algorithm is an attractive and feasible solution to be exploited in next-generation wireless networks with a common receiver, in which usability may go far beyond decreasing network entry delay for M2M applications.

5.4 ENERGY-EFFICIENT CLIENT RELAY SCHEME FOR M2M

5.4.1 Advanced M2M architecture

According to the recent IEEE 802.16p M2M study report [182] reviewed above, a wireless M2M device may act as an aggregation point and communicate data packets on behalf of the other M2M devices which may lack a cellular interface or have a weak communication link to the network. In [P8], we adopt a client relay scheme presented in Chapter 4 to improve the link reliability and energy efficiency for devices with weak links. In particular, the proposed client relay scheme can help ensure that the performance of other cellular devices is not seriously impacted by uncontrolled network access attempts from M2M devices.

Figure 5.2 captures the considered system architecture which is derived from the IEEE 802.16-based M2M communications architecture shown in [182]. The *direct* M2M device is an IEEE 802.16 SS with M2M functionality. The *M2M Server* is an entity that communicates to one or more direct M2M devices through an IEEE 802.16 BS. It has an interface which can be accessed by an M2M service consumer (e.g., utility company). Note that the M2M system architecture allows a direct M2M device to act as an *aggregation point* for *indirect* M2M devices (sensors, actuators, smart meters, and others) without a cellular interface. These indirect M2M devices may use different radio interfaces, such as IEEE 802.11, IEEE 802.15, etc.

Importantly, a direct M2M device can also act as a cooperator for another direct M2M device. That is, a direct M2M device R may relay traffic on behalf of e.g., device A with a weak communication link and thus improve its performance. In this case, air interface changes to IEEE 802.16 may be expected to handle the *client relay* functionality. Particularly, the operation of R should enable an eavesdrop mode to capture traffic from A. In Chapter 4, we summarized a simple client relay protocol that may be used in cellular networks. In [P8], we tailor our client relay scheme to the considered M2M use case.

5.4.2 Performance evaluation framework

The performance of the proposed scheme is evaluated through analysis and simulation across several metrics covering client throughput, latency, and energy consumption. Our analytical approach is a novel queuing model that captures realistic (non-Poisson) traffic arrival patterns borrowed from the evaluation methodology. In [P8], we assume the aggregated M2M traffic at A according to [199]. Interestingly, when the population of meters is sufficiently high, such traffic demonstrates strong *self-similar* properties. We, therefore, account for the Hurst parameter of the self-similar process at node A.

We conclude that the proposed client relay scheme may save power for devices with weak communication links. It is expected that the novel scheme would become an important consideration for the future development of emerging IEEE 802.16p technology. In turn, its success is beneficial for smart metering market supported by international governmental organizations, utility companies, and equipment manufacturers.

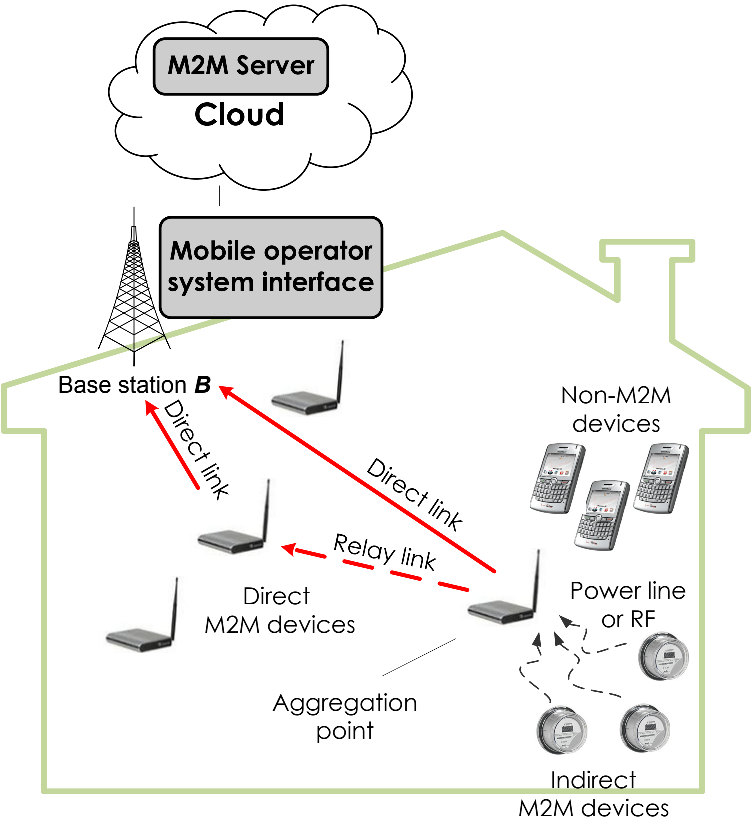


Figure 5.2 Advanced machine-to-machine architecture.

Although we study the client relay approach in the context of smart metering M2M applications, the concepts are equally applicable to other M2M use cases, such as in-building sensors or surveillance equipment that may have a weak connection to the network. Also, the general conclusions in [P8] are applicable to other M2M-enhanced wireless systems [200], such as 3GPP LTE-A.

Chapter 6

Conclusions

6.1 THESIS SUMMARY

Throughout this thesis, we covered a wide range of problems associated with next-generation wireless networks: energy efficient operation, QoS assessment and heterogeneity, client cooperation techniques, and advanced machine-to-machine applications. The major target of respective integral research is to develop novel energy efficient and cooperative solutions in the form of communication algorithms, system architectures, and performance evaluation frameworks to significantly improve client experience, as well as system spectral and energy efficiencies.

When discussing energy efficiency in Chapter 2, we repeatedly emphasized the need for joint link adaptation and resource allocation mechanisms that explicitly take into account client power consumption. Several low complexity solutions have been adapted for next-generation cellular networks and studied within an advanced system-level simulator. Highlighting important practical trade-offs, our results indicate significant promise for future research in the area of energy efficient networking. These results have also been important toward enabling energy efficient features in contemporary 4G wireless standards.

Whereas the focus of the research in Chapter 2 has been set on active mode power consumption, some results regarding client power saving operation were also obtained. In particular, we conducted an in-depth comparison of advanced sleep mode and discontinuous reception mode techniques to conclude that both demonstrate excellent energy efficient client performance and are important for future wireless networking.

In Chapter 3, we continued studying various QoS aspects of state-of-the-art wireless technologies. For WWANs, we argued the importance of accounting for both components of the overall packet delay, that is, the reservation part and the scheduling part. We then proposed an analytical framework for establishing the closed-form upper bound on the overall delay, as well as its exact numerical value. Extending this framework, we addressed the issue of efficient dynamic capacity allocation, where delay-tolerant traffic coexists with delay-critical packet flows. Detailing an efficient allocation mechanism and thoroughly analyzing it, we are now able to optimize the scheduler operation of a next-generation WWAN.

For WLANs, we constructed a generalized performance evaluation model in saturation conditions. Typically, full-buffer behavior corresponds to a worst-case estimate of realistic operation and thus has been a focus of numerous research works. Our model, however, is the most general and accounts for the mixture of traffic, diverse client groups, practical number of packet retransmission attempts, and other parameters while providing saturation throughput and key system-wide probabilities. Importantly, the proposed model incorporates other known models as its special cases. Finally, we considered the problem of simultaneous WWAN and WLAN operation within a multi-radio device. Several coordination solutions have been proposed to recover the performance degradation due to over-the-air interference. Studying these solutions, we conclude that the performance of a multi-radio device in a heterogeneous wireless environment can be successfully recovered.

Improving client performance at the cell edges in Chapter 4, we develop a concept of client relay where neighboring clients assist each other in sending UL traffic. Client relay technology has many beneficial features, including cell-edge spectral and energy efficiency improvement, packet delay reductions, co-channel interference mitigation, etc. We thoroughly study the case of homogeneous client relay, first analytically, and then with an advanced system-level simulator. We also introduce opportunistic client relay behavior, where the cooperator decides independently whether to relay traffic or not. This option is important to provide potential cooperators with a flexible mechanism to trade their individual energy efficiency for the system-wide gain. It is expected that the proposed algorithms, as well as the entire novel system architecture, will be important toward enabling client relay in wireless standards beyond 4G.

Bridging across Chapter 4 and Chapter 2, we also introduce a general analytical framework that couples client cooperation with power saving mode. The proposed model is intentionally broad-scoped to capture only the high-level features of both mechanisms. As such, it is suitable for a plethora of practical applications and its performance evaluation capabilities are really strong.

In Chapter 5, we focus on the advanced machine-to-machine applications within the context of next-generation WWANs. Firstly, we review the recent standardization documents, including our own proposals, with respect to the expected usage models, target requirements, and desired features. We then analyze the comprehensive smart metering scenario and highlight weak system architecture components. In particular, a large population of M2M devices is likely to suffer from increased network entry delays, especially during a surge in near simultaneous network access attempts. In order to recover access latency, we develop a combination of a successive cancellation receiver and a tree multi-access algorithm and thoroughly study its performance. Our analysis confirms that otherwise large delay values may be reduced considerably. Finally, we notice that benefiting the performance of cell-edge M2M devices is sometimes more critical than that of conventional wireless clients. Therefore, we tailor the client relay scheme from Chapter 4 to accommodate M2M devices with weak links yielding improved performance.

The complex research summarized in this thesis results in both theoretical innovations and practical applications, as the topic itself may lead to rethinking the architecture of contemporary multimedia-over-wireless networks. We expect that the proposed solutions and their future extensions will become of significant importance toward further development of wireless communication technologies. These solutions are primarily intended for, but not limited to, cellular operators, telecommunication research companies, equipment vendors, and mobile software companies.

6.2 FUTURE WORK

Even though the energy efficient and cooperative solutions presented in this thesis constitute a solid and integrated research, there exist many opportunities to extend and improve every particular component. For instance, low complexity energy efficient algorithms evaluated in Chapter 2 with a system-level simulator do not explicitly take into account the effect of inter-cell interference. As such, they may be extended to a multi-cell scenario themselves.

Due to the highly dynamic nature of modern wireless networks, the issue of network admission control is becoming very challenging and important. Existing work on efficiency, QoS, and cooperation might benefit from accounting for this problem explicitly. Another important issue in multi-cell communication networks is how to properly associate mobile clients with serving base stations. This problem is typically known as client association.

The QoS assurance framework outlined in Chapter 3 mainly focuses on controlling the mean packet delay. Integrating with other relevant QoS metrics is an important continuation of this research. We also remind you that the proposed WLAN performance model assumes saturation, which does not fit every practical scenario. Extending the framework for the dynamic case is crucial for generalizing our model further on.

With the client relay research in Chapter 4, we barely scratched the surface of what may be called cellular-assisted peer-to-peer communications. Practically, a WWAN client may benefit from having a direct link to its neighbor, whereas currently traffic would be looped back via a base station. The ability of a client to establish a direct link either in-band or out-band is thus highly desired to relieve congestion of cellular technologies beyond 4G. Device-to-device communications are thus a hot topic in current standardization and existing research is only the first step forward. The design of efficient relay-aware schedulers is another interesting extension of this study.

As machine-to-machine communications are coming into force worldwide very recently, the next-generation wireless standards are only starting to react. We emphasize that a network may go far beyond simple mechanisms for overload control and accounting for small burst transmissions seems to be a necessary following step. Whereas contemporary wireless technologies are optimized to transmit larger portions of data, M2M devices are likely to only send several bytes. Clearly, this results in excessive overheads which should be controlled. Consequently, supporting small burst transmission is an important consideration in beyond 4G networks.

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 were used as part of any other thesis. Additionally, the thesis refers to *eight* important references [1], [2], [3], [4], [5], [6], [7], and [8] which were completed by the author in tight international collaboration with colleagues from Finland, Russia, USA, Hungary, and Belgium. Works [P1], [P5], [P6], and [3] are articles published in scientific journals, documents [7] and [8] are IEEE standardization contributions, while the rest are conference papers.

In order to facilitate navigation between the main and the related publications constituting this thesis, Figure 7.1 demonstrates the distribution of publications across the core chapters of this manuscript. More importantly, the scheme also highlights the relations between the publications by indicating explicit and implicit logical connections. With an *explicit* connection (thick line), one research paper is derived from another borrowing/extending the methodology. With an *implicit* connection (thin line), the papers follow different methodologies but are adjacent with respect to problematics.

Publications [1], [2], and [P1] belong to Chapter 2 and, as the name implies, consider energy efficient wireless systems. In particular, [1] addresses client power saving operation, whereas [2] and its extended version [P1] focus on active mode power consumption. Further, publications [P2], [P3], [3], [P4], and [P5] relate to Chapter 3 and study various aspects of heterogeneous networking and QoS. In more detail, [P2] and [P3] target performance improvement of wireless broadband networks (similarly to [2] and [P1]) and evaluate the overall packet delay by a closed-form upper bound [P2] and an exact numerical solution [P3]. This research is continued in [3] with dynamic capacity allocation issues. By contrast, [P4] deals with QoS aspects of wireless local area networks. Then, [P5] is the logical extension of work in [P3] and [P4] assuming the coexistence of a broadband cellular network and a local area network.

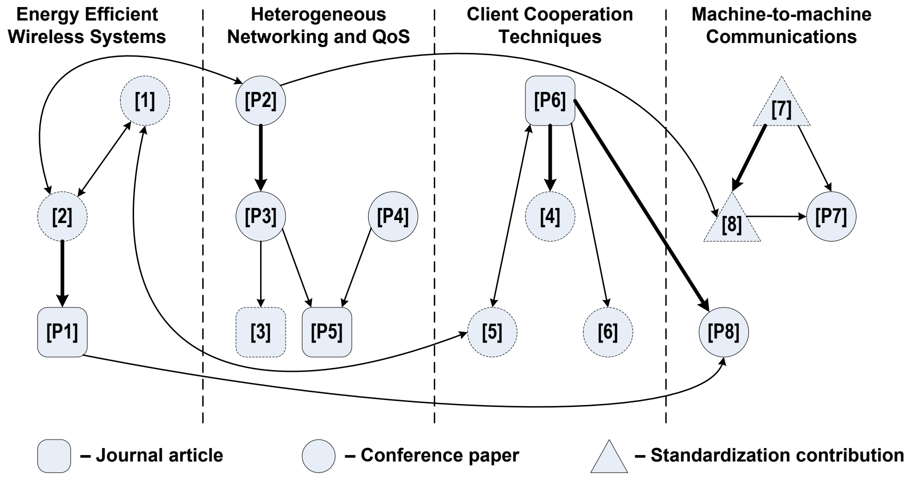

Energy Efficient Wireless Systems
Heterogeneous Networking and QoS
Client Cooperation Techniques
Machine-to-machine Communications
[1]
[2]
[P1]
[P2]
[P3]
[P4]
[3]
[P5]
[P6]
[4]
[5]
[6]
[7]
[8]
[P7]
[P8]
– Journal article
– Conference paper
– Standardization contribution

Figure 7.1 Logical connections between publications.

Regarding Chapter 4 and client cooperation technique described therein, the related papers are [P6], [4], [5], and [6]. The work in [4] extends the baseline three-node client relay model in [P6] by introducing the idea of opportunistic cooperation. Consequently, [5] seeks to couple client cooperation with power saving operation and is thus loosely connected to both [P6] and [1]. Moreover, [6] is another extension of [P6], where client relay technique is studied with system-level simulations. Finally, publications [7], [8], [P7], and [P8] are related to machine-to-machine communications and, therefore, Chapter 5. IEEE contribution [7] analyzes smart grid applications and respective vulnerabilities. Furthermore, extended contribution [8] highlights the problem of excessive network entry delays focusing on the protocol described in [P2]. Aiming to recover the prohibitive values of network entry delay as per [7] and [8], the work in [P7] proposes an efficient multi-access algorithm based on successive interference cancellation. The publication [P8] adopts client relay scheme from [P6] to improve energy efficient performance (as per [P1]) of cell-edge M2M devices with weak links.

The major contribution of each of the *main* publications is clarified below.

- **[P1]** 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, no. 14, pp. 1662–1671, 2012.

  Description

  We argue that energy efficiency is increasingly important for wireless cellular systems due to the limited battery resources of mobile clients. While modern cellular standards emphasize low client battery consumption, existing techniques do not explicitly focus on reducing power that is consumed when a client is actively communicating with the network. In [P1], we evaluate the performance of the recently introduced power-bandwidth optimization techniques using realistic cellular system simulation model, which is compliant with the methodology proposed for the IEEE 802.16m standard. The paper addresses several practical trade-offs associated with the implementation of energy efficient schemes. Our simulation results indicate that energy efficient techniques continue to provide considerable power savings, even when accounting for realistic system parameters and channel environments.

This paper is a collaborative work of the author and his supervisor with Dr. Nageen Himayat from Wireless Communications Laboratory, Intel Corporation (USA), as well as with Pavel Gonchukov and Prof. Andrey Turlikov from St. Petersburg State University of Aerospace Instrumentation (Russia).

- **[P2]** S. Andreev, Z. Saffer, A. Turlikov, and A. Vinel, "Upper bound on overall delay in wireless broadband networks with non real-time traffic," in *Proc. of the 17th International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA)*, pp. 262–276, 2010.

  Description

  In [P2], we consider the non real-time traffic in IEEE 802.16-based wireless broadband networks with contention-based bandwidth reservation mechanism. We introduce a new system model and establish an *upper bound* on the overall data packet delay. The model enables symmetric Poisson arrival flows and accounts for both the reservation and scheduling delay components. The analytical result is verified by simulation.

  This paper is a collaborative work of the author with Dr. Zsolt Saffer from Budapest University of Technology and Economics (Hungary), with Prof. Andrey Turlikov from St. Petersburg State University of Aerospace Instrumentation (Russia), as well as with Alexey Vinel from (formerly) St. Petersburg Institute for Informatics and Automation (Russia).

- **[P3]** S. Andreev, Z. Saffer, and A. Turlikov, "Delay analysis of wireless broadband networks with non real-time traffic," in *Proc. of the 4th International Workshop on Multiple Access Communications (MACOM)*, pp. 206–217, 2011.

  Description

  In [P3], we present the *exact analysis* of the mean overall packet delay of non real-time traffic in IEEE 802.16-based wireless broadband networks. As in [P2], we consider the case of contention-based bandwidth reservation. The system model accounts for both bandwidth reservation and packet transmission delay components of the overall delay. The queuing analysis is the continuation of our previous work in [P2] and is based on the description of the joint content of the outgoing subscriber station buffer and the base station grant buffer. This is achieved by means of a properly chosen bivariate embedded Markov chain. The mean overall packet delay is computed from its equilibrium solution. The analytical approach is verified by means of simulation. The corresponding analytical and simulation results show excellent agreement with each other.

This paper is a collaborative work of the author with Dr. Zsolt Saffer from Budapest University of Technology and Economics (Hungary), as well as with Prof. Andrey Turlikov from St. Petersburg State University of Aerospace Instrumentation (Russia).

- **[P4]** S. Andreev, Y. Koucheryavy, and L. de Sousa, "Calculation of transmission probability in heterogeneous ad hoc networks," in *Proc. of the Baltic Congress on Future Internet and Communications (BCFIC)*, pp. 75–82, 2011.

  Description

  In [P4], we address the problem of MAC performance evaluation of a contemporary IEEE 802.11 WLAN. The network is observed under saturation conditions and the packet transmission probability analysis is conducted with the novel regenerative approach. The proposed model accounts for collision resolution protocol parameters, packet retry limit, coexistence of unicast and broadcast traffic, and heterogeneous QoS environment. Our analytical model is a generalization of several well-known models extensively used in the field. The obtained results are verified to demonstrate perfect agreement with ns-2 simulations.

  This paper is a collaborative work of the author and his supervisor with Luís Filipe Dias de Sousa from Network and Communication Group (Germany). The latter conducted his M.Sc. studies under the supervision of the author when carrying out this work.

- **[P5]** S. Andreev, K. Dubkov, and A. Turlikov, "IEEE 802.11 and 802.16 cooperation within multi-radio stations," *Wireless Personal Communications Journal (WIRE)*, vol. 58, no. 3, pp. 525–543, 2011.

  Description

  In [P5], we consider a multi-radio wireless network client that is capable of simultaneous operation in IEEE 802.16 and IEEE 802.11 communication networks. In order to enable the cooperative functioning of both networks we introduce the medium access control coordination concept. A set of coordination algorithms is then presented together with a simple approach to their performance analysis. Our performance evaluation shows that the saturation goodput of the proposed coordination algorithm is at least 50% higher than that of the existing coordination algorithms. Moreover, it allows for the considerable reduction in the data packet delay.

  This paper is a collaborative work of the author with Konstantin Dubkov from (formerly) Intel Corporation (Russia), as well as with Prof. Andrey Turlikov from St. Petersburg State University of Aerospace Instrumentation (Russia).

- **[P6]** S. Andreev, O. Galinina, and A. Vinel, "Performance evaluation of a three node client relay system," *International Journal of Wireless Networks and Broadband Technologies (IJWNBT)*, vol. 1, no. 1, pp. 73–84, 2011.

  Description

In [P6], we examine a client relay system comprising three wireless nodes. Closed-form expressions for the mean packet delay, as well as for the throughput, energy expenditure, and energy efficiency of the source nodes are obtained. The precision of the established parameters is verified via simulations.

This paper is a collaborative work of the author with Olga Galinina and Alexey Vinel from the same research group at Tampere University of Technology (Finland).

- **[P7]** S. Andreev, E. Pustovalov, and A. Turlikov, “A practical tree algorithm with successive interference cancellation for delay reduction in IEEE 802.16 networks,” in *Proc. of the 18th International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA)*, pp. 301–315, 2011.

  Description

  In [P7], we thoroughly study a modification of a tree algorithm with successive interference cancellation. In particular, we focus on the algorithm throughput and account for a single signal memory location, as well as cancellation errors of three types. The resulting scheme is robust to imperfect interference cancellation and is tailored to the UL bandwidth request collision resolution in an IEEE 802.16 cellular network. The mean packet delay is shown to be considerably reduced when using the proposed approach.

  This paper is a collaborative work of the author with Eugeny Pustovalov and Prof. Andrey Turlikov from St. Petersburg State University of Aerospace Instrumentation (Russia).

- **[P8]** S. Andreev, O. Galinina, and Y. Koucheryavy, “Energy-efficient client relay scheme for machine-to-machine communication,” in *Proc. of the 54th IEEE Global Communications Conference (GLOBECOM)*, 2011.

  In [P8], we consider a wireless cellular network capable of supporting Machine-to-Machine (M2M) applications. According to the recent IEEE 802.16p proposals, a wireless M2M device may act as an aggregation point and communicate data packets on behalf of other M2M devices, which may lack a cellular interface or have a weak communication link to the network. We propose a client relay scheme to improve the link reliability and energy efficiency of the devices with weak links. Performance of the proposed scheme is evaluated through analysis and simulation across several metrics covering client throughput, latency, and energy consumption. Our analytical approach is a novel queuing model that captures realistic traffic arrival patterns borrowed from the evaluation methodology. It is shown that the obtained analytical results demonstrate excellent agreement with simulation. We also conclude that the proposed client relay scheme may save power for the devices with weak communication links.

  This paper is a collaborative work of the author and his supervisor with Olga Galinina from the same research group at Tampere University of Technology (Finland).

7.2 AUTHOR'S CONTRIBUTION

The research work summarized in this thesis has been carried out in the Department of Communications Engineering, Tampere University of Technology, Finland. The author of this thesis is the main contributor to [P1]-[P8] and has originally proposed the research topic. Consequently, the reported research has been done mainly by the author, naturally supervised and guided by his supervisor Prof. Yevgeni Koucheryavy and by his co-supervisor Prof. Andrey Turlikov. As such, the manuscripts [P1]-[P8] have been written primarily by the author. Needless to say that numerous discussions with the supervisors helped the author shape the ideas presented in this thesis, as well as improve the quality and the style of his writing. Further, many particular features published in [P1]-[P8] have been developed in tight collaboration between the author, his international colleagues, and the members of the research group in Tampere. 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 of the considered energy efficient techniques, as well as for developing the overall evaluation methodology. In [P2], the author has formulated the general problem, introduced the system model, and generally conducted both analysis and simulation. In [P3], the author has also contributed the system model, as well as the extended simulation software that was originally used by [P2]. In [P4], the author has formulated the research hypothesis, detailed the system model, and then supervised the overall study which was carried out by his M.Sc. student. In [P5], the author has generally developed the analytical framework, as well as partly implemented the simulation software concerning the cellular network and the multi-radio device. In [P6], the author has proposed the research target and the system model, as well as suggested the performance benchmarking strategy. In [P7], the author has developed the system model and analyzed the original collision resolution algorithm. He was also responsible for the simulation part. In [P8], the author has proposed the original idea and the protocol, as well as defined the evaluation methodology.

Bibliography

[1] A. Anisimov, S. Andreev, A. Lokhanova, and A. Turlikov, “Energy efficient operation of 3GPP LTE-Advanced and IEEE 802.16m downlink channel,” in *Proc. of the 3rd International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT)*, 2011.

[2] S. Andreev, Y. Koucheryavy, N. Himayat, P. Gonchukov, and A. Turlikov, “Active-mode power optimization in OFDMA-based wireless networks,” in *Proc. of the 6th IEEE Broadband Wireless Access (BWA) Workshop co-located with IEEE GLOBECOM*, 2010.

[3] Z. Saffer, S. Andreev, and Y. Koucheryavy, “Performance evaluation of uplink delay-tolerant packet service in IEEE 802.16-based networks,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2011, pp. 1–13, March 2011.

[4] S. Andreev, O. Galinina, A. Lokhanova, and Y. Koucheryavy, “Analysis of client relay network with opportunistic cooperation,” in *Proc. of the 9th International Conference on Wired/Wireless Internet Communications (WWIC)*, pp. 247–258, 2011.

[5] T. Demoor, S. Andreev, K. de Turck, H. Bruneel, and D. Fiems, “On the effect of combining cooperative communication with sleep mode,” in *Proc. of the 9th Annual Conference on Wireless On-demand Network Systems and Services (WONS)*, 2012.

[6] A. Pyattaev, S. Andreev, O. Galinina, and Y. Koucheryavy, “System-level evaluation of opportunistic client cooperation in wireless cellular networks,” in *Proc. of the 20th IEEE International Conference on Computer Communications and Networks (ICCCN)*, 2011.

[7] N. Himayat, S. Talwar, K. Johnsson, S. Mohanty, X. Wang, G. Wei, E. Schooler, G. Goodman, S. Andreev, O. Galinina, and A. Turlikov, *Informative text on Smart Grid applications for inclusion in IEEE 802.16p Systems Requirements Document (SRD).* `IEEE 802.16p-10/0007r1`, November 2010.

[8] N. Himayat, S. Talwar, K. Johnsson, S. Andreev, O. Galinina, and A. Turlikov, *Proposed IEEE 802.16p performance requirements for network entry by large number of devices.* `IEEE 802.16p-10/0006r1`, November 2010.

[9] *IEEE 802.16m-2011, Amendment to IEEE Standard for Local and metropolitan area networks. Advanced Air Interface*, May 2011.

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

[11] G. Song, *Cross-Layer Resource Allocation and Scheduling in Wireless Multicarrier Networks.* PhD thesis, Georgia Institute of Technology, 2005.

[12] G. Miao, *Cross-Layer Optimization for Spectral and Energy Efficiency.* PhD thesis, Georgia Institute of Technology, 2008.

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

[14] K. Lahiri, A. Raghunathan, S. Dey, and D. Panigrahi, "Battery-driven system design: a new frontier in low power design," in *Proc. of the 15th International Conference on VLSI Design*, pp. 261–267, 2002.

[15] K. Pentikousis, "In search of energy-efficient mobile networking," *IEEE Communications Magazine*, vol. 48, pp. 95–103, January 2010.

[16] A. Ephremides and B. Hajek, "Information theory and communication networks: an unconsummated union," *IEEE Transactions on Information Theory*, vol. 44, pp. 2416–2434, October 1998.

[17] A. Goldsmith, *Wireless Communications.* Cambridge University Press, 2005.

[18] T. Rappaport, *Wireless Communications: Principles and Practice (2nd Edition).* Prentice Hall, 2002.

[19] S. Shakkottai, T. Rappaport, and P. Karlsson, "Cross-layer design for wireless networks," *IEEE Communications Magazine*, vol. 41, pp. 74–80, October 2003.

[20] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, pp. 1277–1294, June 2002.

[21] R. Knopp and P. Humblet, "Information capacity and power controlling single-cell multiuser communications," in *Proc. of the IEEE International Conference on Communications (ICC)*, pp. 331–335, 1995.

[22] Atheros Communications, *White paper: Power consumption & energy efficiency*, 2003.

[23] G. Miao, N. Himayat, and G. Li, "Energy-efficient link adaptation in frequency-selective channels," *IEEE Transactions on Communications*, vol. 58, pp. 545–554, February 2010.

[24] J. Chuang and N. Sollenberger, "Beyond 3G: wideband wireless data access based on OFDM and dynamic packet assignment," *IEEE Communications Magazine*, vol. 38, pp. 78–87, July 2000.

[25] G. Song and G. Li, "Cross-layer optimization for OFDM wireless networks – part I: theoretical framework," *IEEE Transactions on Wireless Communications*, vol. 4, pp. 614–624, March 2005.

[26] C. Wong, R. Cheng, K. Letaief, and R. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 1747–1758, October 1999.

[27] G. Song and G. Li, "Cross-layer optimization for OFDM wireless networks – part II: algorithm development," *IEEE Transactions on Wireless Communications*, vol. 4, pp. 625–634, March 2005.

[28] G. Miao, N. Himayat, G. Li, and A. Swami, "Cross-layer optimization for energy-efficient wireless communications: a survey," *Journal on Wireless Communications and Mobile Computing*, vol. 9, pp. 529–542, April 2009.

[29] P. Kolios, V. Friderikos, and K. Papadaki, "A practical approach to energy efficient communications in mobile wireless networks," *Mobile Networks and Applications*, vol. 17, pp. 267–280, April 2012.

[30] G. Song and G. Li, "Asymptotic throughput analysis for channel-aware scheduling," *IEEE Transactions on Communications*, vol. 54, pp. 1827–1834, October 2006.

[31] X. Lin, N. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, pp. 1452–1463, August 2006.

[32] L. Benini, A. Bogliolo, and G. de Micheli, "A survey of design techniques for system-level dynamic power management," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 8, pp. 299–316, June 2000.

[33] C. Schurgers, *Energy-Aware Wireless Communications.* PhD thesis, University of California Los Angeles, 2002.

[34] A. Pantelidou and A. Ephremides, "A cross-layer view of optimal scheduling," *IEEE Transactions on Information Theory*, vol. 56, pp. 5568–5580, November 2010.

[35] Y. Chen, S. Zhang, S. Xu, and G. Li, "Fundamental trade-offs on green wireless networks," *IEEE Communications Magazine*, vol. 49, pp. 30–37, June 2011.

[36] G. Miao, N. Himayat, G. Li, A. Koc, and S. Talwar, "Interference-aware energy-efficient power optimization," in *Proc. of the IEEE International Conference on Communications (ICC)*, 2009.

[37] V. Bhaghavan, A. Demers, S. Shenker, and L. Zhang, "MACAW: a media access protocol for wireless LAN's," in *Proc. of the Conference on Communications Architectures, Protocols and Applications (SIGCOMM)*, vol. 24, 1994.

[38] G. Ganesan, G. Song, and G. Li, "Asymptotic throughput analysis of distributed multichannel random access schemes," in *Proc. of the IEEE International Conference on Communications (ICC)*, vol. 5, pp. 3637–3641, 2005.

[39] V. Naware, G. Mergen, and L. Tong, "Stability and delay of finite-user slotted Aloha with multipacket reception," *IEEE Transactions on Information Theory*, vol. 51, pp. 2636–2656, July 2005.

[40] K. Bai and J. Zhang, "Opportunistic multichannel Aloha: distributed multiaccess control scheme for OFDMA wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 55, pp. 848–855, May 2006.

[41] M. Ngo, S. Adireddy, and L. Tong, "Optimal channel-aware Aloha protocol for random access in WLANs with multipacket reception and decentralized channel state information," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2575–2588, June 2008.

[42] I. Akyildiz, S. Weilian, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, pp. 102–114, August 2002.

[43] A. Ephremides, "Ad hoc networks: not an ad hoc field anymore," *Wireless Communications and Mobile Computing*, vol. 2, no. 5, pp. 441–448, 2002.

[44] I. Chlamtac, M. Conti, and J. Liu, "Mobile ad hoc networking: imperatives and challenges," *Ad Hoc Networks*, vol. 1, pp. 13–64, July 2003.

[45] I. Akyildiz and X. Wang, "A survey on wireless mesh networks," *IEEE Communications Magazine*, vol. 43, pp. S23–S30, September 2005.

[46] X. Wang, E. Knightly, M. Conti, and A. Ephremides, "A special issue on "wireless mesh networks"," *Ad Hoc Networks*, vol. 5, no. 6, pp. 649–651, 2007.

[47] G. Li, Z. Xu, C. Xiong, C. Yang, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient wireless communications: tutorial, survey, and open issues," *IEEE Wireless Communications*, vol. 18, pp. 28–35, December 2011.

[48] A. Marques, L. Lopez-Ramos, G. Giannakis, J. Ramos, and A. Caamano, "Optimal cross-layer resource allocation in cellular networks using channel and queue-state information," *IEEE Transactions on Vehicular Technology*, vol. 61, pp. 2789–2807, July 2012.

[49] S. Nanda, K. Balachandran, and S. Kumar, "Adaptation techniques in wireless packet data services," *IEEE Communications Magazine*, vol. 38, pp. 54–64, January 2000.

[50] K. Tenhonen, J. Hamalainen, R. Wichman, and K. Horneman, "On the effect of channel-aware scheduling to CDMA uplink capacity," in *Proc. of the 17th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2006.

[51] X. Liu, E. Chong, and N. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, pp. 2053–2064, October 2001.

[52] G. Cao and M. Singhal, "An adaptive distributed channel allocation strategy for mobile cellular networks," *Journal of Parallel and Distributed Computing*, vol. 60, pp. 451–473, April 2000.

[53] G. Stuber, *Principles of Mobile Communication (3rd Edition)*. Springer, 2011.

[54] M. Necker, "Coordinated fractional frequency reuse," in *Proc. of the 10th ACM Symposium on Modeling, Analysis, and Simulation of Wireless and Mobile Systems (MSWiM)*, pp. 296–305, 2007.

[55] J. Andrews, "Interference cancellation for cellular systems: a contemporary overview," *IEEE Wireless Communications*, vol. 12, pp. 19–29, April 2005.

[56] H. Zhang and H. Dai, "Cochannel interference mitigation and cooperative processing in downlink multicell multiuser MIMO networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2004, pp. 222–235, December 2004.

[57] W. Choi and J. Andrews, "Base station cooperatively scheduled transmission in a cellular MIMO TDMA system," in *Proc. of the 40th Annual Conference on Information Sciences and Systems*, pp. 105–110, 2006.

[58] P. Omiyi and H. Haas, "Improving time-slot allocation in 4th generation OFDM/TDMA TDD radio access networks with innovative channel-sensing," in *Proc. of the IEEE International Conference on Communications (ICC)*, vol. 6, pp. 3133–3137, 2004.

[59] V. Rodoplu and T. Meng, "Bits-per-Joule capacity of energy-limited wireless networks," *IEEE Transactions on Wireless Communications*, vol. 6, pp. 857–865, March 2007.

[60] G. Miao, N. Himayat, G. Li, and S. Talwar, "Low-complexity energy-efficient OFDMA," in *Proc. of the IEEE International Conference on Communications (ICC)*, 2009.

[61] R. Gallager, "Power limited channels: coding, multiaccess, and spread spectrum," in *Proc. of the Conference on Information Sciences and Systems*, vol. 1, 1988.

[62] V. Sethuraman and B. Hajek, "Capacity per unit energy of fading channels with a peak constraint," *IEEE Transactions on Information Theory*, vol. 51, pp. 3102–3120, September 2005.

[63] S. Verdu, "Spectral efficiency in the wideband regime," *IEEE Transactions on Information Theory*, vol. 48, pp. 1319–1343, June 2002.

[64] F. Meshkati, H. Poor, S. Schwartz, and N. Mandayam, "An energy-efficient approach to power control and receiver design in wireless networks," *IEEE Transactions on Communications*, vol. 53, pp. 1885–1894, November 2005.

[65] G. Li and G. Stuber, *OFDM for Wireless Communications.* Springer, 2006.

[66] A. Wang, S. Cho, C. Sodini, and 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, 2001.

[67] S. Cui, A. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Transactions on Wireless Communications*, vol. 4, pp. 2349–2360, September 2005.

[68] P. Grover, K. Woyach, and A. Sahai, "Towards a communication-theoretic understanding of system-level power consumption," *IEEE Journal on Selected Areas in Communications*, vol. 29, pp. 1744–1755, September 2011.

[69] C. Schurgers and M. Srivastava, "Energy optimal scheduling under average throughput constraint communications," in *Proc. of the IEEE International Conference on Communications (ICC)*, vol. 3, pp. 1648–1652, 2003.

[70] 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.

[71] R. Mangharam, R. Rajkumar, S. Pollin, F. Catthoor, B. Bougard, L. van der Perre, and I. Moeman, "Optimal fixed and scalable energy management for wireless networks," in *Proc. of the 24th IEEE Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, vol. 1, pp. 114–125, 2005.

[72] R. Hu, S. Talwar, and P. Zong, *Cooperative, Green and Mobile Heterogeneous Wireless Networks.* IEEE 802.16x, Future Wireless Networks tutorial, March 2011.

[73] J. Zhuang, L. Jalloul, R. Novak, and J. Park, *IEEE 802.16m Evaluation Methodology Document (EMD).* `IEEE 802.16m-08/004r5`, January 2009.

[74] I. Akyildiz, D. Gutierrez-Estevez, and E. Reyes, "The evolution to 4G cellular systems: LTE-Advanced," *Physical Communication*, vol. 3, no. 4, pp. 217–244, 2010.

[75] Z. Hasan, H. Boostanimehr, and V. Bhargava, "Green cellular networks: a survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 13, pp. 524–540, 4th quarter 2011.

[76] ETSI MCC, RWS-120052, *Draft Report of 3GPP RAN Workshop on Release 12 and onwards*, June 2012.

[77] G. Perrucci, *Energy Saving Strategies on Mobile Devices.* PhD thesis, Aalborg University, 2009.

[78] F. Meshkati, H. Poor, S. Schwartz, and R. Balan, "Energy efficiency and delay quality-of-service in wireless networks," in *Proc. of the Inaugural Workshop of the Center for Information Theory and its Applications*, 2006.

[79] N. Himayat, M. Venkatachalam, A. Koc, S. Talwar, H. Yin, S. Ahmadi, M. Ho, G. Miao, G. Li, H. Kim, P.-K. Liao, Y.-S. Chen, P. Cheng, Z. Yan-Xiu, M. Chen, and R. Li, *Improving client energy consumption in 802.16m.* IEEE C80216m-09_0107, January 2009.

[80] G. Miao, N. Himayat, G. Li, and D. Bormann, "Energy efficient design in wireless OFDMA," in *Proc. of the IEEE International Conference on Communications (ICC)*, pp. 3307–3312, 2008.

[81] C. Wijting, K. Doppler, K. Kalliojarvi, T. Svensson, M. Sternad, G. Auer, N. Johansson, J. Nystrom, M. Olsson, A. Osseiran, M. Dottling, J. Luo, T. Lestable, and S. Pfletschinger, "Key technologies for IMT-Advanced mobile communication systems," *IEEE Wireless Communications*, vol. 16, pp. 76–85, June 2009.

[82] R. Wang, J. Tsai, C. Maciocco, T.-Y. Tai, and J. Wu, "Reducing power consumption for mobile platforms via adaptive traffic coalescing," *IEEE Journal on Selected Areas in Communications*, vol. 29, pp. 1618–1629, September 2011.

[83] N. Himayat, M. Venkatachalam, A. Koc, S. Talwar, H. Yin, S. Ahmadi, M. Ho, R. Yang, S. Andreev, P. Gonchukov, A. Turlikov, G. Miao, G. Li, H. Kim, M. Kone, M.-H. Tao, and Y.-C. Hsiao, *Amendment Text Proposal for Section 10.5.3 on Power Management for Connected Mode.* IEEE C802.16m-09/0553r2, March 2009.

[84] D. Raychaudhuri and N. Mandayam, "Frontiers of wireless and mobile communications," *Proceedings of the IEEE*, vol. 100, pp. 824–840, April 2012.

[85] Z. Jiang, H. Mason, B. Kim, N. Shankaranarayanan, and P. Henry, "A subjective survey of user experience for data applications for future cellular wireless networks," in *Proc. of the Symposium on Applications and the Internet*, pp. 167–175, 2001.

[86] Z. Jiang, Y. Ge, and G. Li, "Max-utility wireless resource management for best effort traffic," *IEEE Transactions on Wireless Communications*, vol. 4, pp. 100–111, January 2005.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

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

[92] 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.

[93] 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.

[94] S. Patil and G. de Veciana, "Managing resources and quality of service in heterogeneous wireless systems exploiting opportunism," *IEEE/ACM Transactions on Networking*, vol. 15, pp. 1046–1058, October 2007.

[95] 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.

[96] 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.

[97] 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.

[98] G. de Veciana, T.-J. Lee, and T. Konstantopoulos, "Stability and performance analysis of networks supporting elastic services," *IEEE/ACM Transactions on Networking*, vol. 9, pp. 2–14, February 2001.

[99] G. Song and G. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless networks," *IEEE Communications Magazine*, vol. 43, pp. 127–134, December 2005.

[100] *IEEE 802.16-2009, Part 16: Air Interface for Broadband Wireless Access Systems*, May 2009.

[101] *IEEE 802.16-2004, Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, October 2004.

[102] Y. Sekercioglu, M. Ivanovich, and A. Yegin, "A survey of MAC based QoS implementations for WiMAX networks," *Computer Networks*, vol. 53, pp. 2517–2536, September 2009.

[103] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund, "Quality of service support in IEEE 802.16 networks," *IEEE Network*, vol. 20, pp. 50–55, March/April 2006.

[104] *WiMAX Forum, home page.* `http://www.wimaxforum.org/`.

[105] G. Paschos, I. Papapanagiotou, C. Argyropoulos, and S. Kotsopoulos, "A heuristic strategy for IEEE 802.16 WiMAX scheduler for quality of service," in *Proc. of the 45th Congress of the Federation of Telecommunications Engineers of the European Community (FITCE)*, 2006.

[106] L. de Moraes and P. Maciel, "A variable priorities MAC protocol for broadband wireless access with improved channel utilization among stations," in *Proc. of the International Telecommunications Symposium*, pp. 398–403, 2006.

[107] Y.-J. Chang, F.-T. Chien, and C.-C. Kuo, "Delay analysis and comparison of OFDM-TDMA and OFDMA under IEEE 802.16 QoS framework," in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM)*, 2006.

[108] D.-H. Cho, J.-H. Song, M.-S. Kim, and K.-J. Han, "Performance analysis of the IEEE 802.16 wireless metropolitan network," in *Proc. of the 1st International Conference on Distributed Frameworks for Multimedia Applications (DFMA)*, pp. 130–136, 2005.

[109] A. Vinel, Y. Zhang, Q. Ni, and A. Lyakhov, "Efficient request mechanism usage in IEEE 802.16," in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM)*, 2006.

[110] M. Wood, "An analysis of the design and implementation of QoS over IEEE 802.16," tech. rep., Washington University in St. Louis, 2006.

[111] T. Bohnert, D. Staehle, G.-S. Kuo, Y. Koucheryavy, and E. Monteiro, "Speech quality aware admission control for fixed IEEE 802.16 wireless MAN," in *Proc. of the IEEE International Conference on Communications (ICC)*, pp. 2690–2695, 2008.

[112] T. Efimushkina, N. Vassileva, D. Moltchanov, and Y. Koucheryavy, "Analytical performance evaluation of a WiMAX cell with VoIP/elastic data traffic," in *Proc. of the IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 509–514, 2011.

[113] I. Rubin, "Access-control disciplines for multi-access communication channels: reservation and TDMA schemes," *IEEE Transactions on Information Theory*, vol. 25, pp. 516–536, September 1979.

[114] S. Andreev, Z. Saffer, A. Turlikov, and A. Vinel, “Overall delay in IEEE 802.16 with contention-based random access,” in *Proc. of the 16th International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA)*, pp. 89–102, 2009.

[115] R. Iyengar, P. Iyer, and B. Sikdar, “Delay analysis of 802.16 based last mile wireless networks,” in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM)*, vol. 5, pp. 3123–3127, 2005.

[116] L. Lin, W. Jia, and W. Lu, “Performance analysis of IEEE 802.16 multicast and broadcast polling based bandwidth request,” in *Proc. of the IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1854–1859, 2007.

[117] H. Sadjadpour, R. Ulman, A. Swami, and A. Ephremides, “Wireless mobile ad hoc networks,” in *EURASIP Journal on Wireless Communications and Networking*, Hindawi Publishing Corporation, 2007.

[118] *IEEE 802.11-2012, Part 11: Local and metropolitan area networks*, March 2012.

[119] G. Bianchi, “Performance analysis of the IEEE 802.11 distributed coordination function,” *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 535–547, March 2000.

[120] D. Cavalcanti, D. Agrawal, C. Cordeiro, B. Xie, and A. Kumar, “Issues in integrating cellular networks WLANs, and MANETs: a futuristic heterogeneous wireless network,” *IEEE Wireless Communications*, vol. 12, pp. 30–41, June 2005.

[121] B. Walke, S. Mangold, and L. Berlemann, *IEEE 802 Wireless Systems: Protocols, Multi-Hop Mesh/Relaying, Performance and Spectrum Coexistence*. Wiley, 2007.

[122] J. Zhu, A. Waltho, X. Yang, and X. Guo, “Multi-radio coexistence: challenges and opportunities,” in *Proc. of the 16th International Conference on Computer Communications and Networks (ICCCN)*, pp. 358–364, 2007.

[123] T. Zetterman, A. Piipponen, K. Raiskila, and S. Slotte, “Multi-radio coexistence and collaboration on an SDR platform,” *Analog Integrated Circuits and Signal Processing*, vol. 69, pp. 329–339, December 2011.

[124] A. Kamerman, “Coexistence between Bluetooth and IEEE 802.11 CCK solutions to avoid mutual interference,” tech. rep., Lucent Technologies Bell Labs, `IEEE 802.11-00/162`, 1999/2000.

[125] I. Howitt, “WLAN and WPAN coexistence in UL band,” *IEEE Transactions on Vehicular Technology*, vol. 50, pp. 1114–1124, July 2001.

[126] F. Wang, A. Nallanathan, and H. Garg, “Introducing packet segmentation for the IEEE 802.11b throughput enhancement in the presence of Bluetooth,”

in *Proc. of the 59th IEEE Vehicular Technology Conference (VTC)*, vol. 4, pp. 2252–2256, 2004.

[127] P. Djukic and S. Valaee, "802.16 MCF for 802.11a based mesh networks: a case for standards re-use," in *Proc. of the 23rd Biennial Symposium on Communications*, pp. 186–189, 2006.

[128] S. Mangold, *Analysis of IEEE 802.11e and Application of Game Models for Support of Quality-of-Service in Coexisting Wireless Networks.* PhD thesis, RWTH Aachen University, 2003.

[129] L. Berlemann, C. Hoymann, G. Hiertz, and S. Mangold, "Coexistence and interworking of IEEE 802.16 and IEEE 802.11(e)," in *Proc. of the IEEE 63rd Vehicular Technology Conference (VTC)*, vol. 1, pp. 27–31, 2006.

[130] IEEE 802 Plenary Tutorial, *WPAN/WLAN/WWAN Multi-Radio Coexistence*, November 2007.

[131] S. Cui, A. Goldsmith, and A. Bahai, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 1089–1098, August 2004.

[132] S. Jayaweera, "An energy-efficient virtual MIMO architecture based on V-BLAST processing for distributed wireless sensor networks," in *Proc. of the 1st Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks (SECON)*, pp. 299–308, 2004.

[133] J. Rabaey, J. Ammer, J. da Silva Jr., and D. Patel, "PicoRadio: ad-hoc wireless networking of ubiquitous low-energy sensor/monitor nodes," in *Proc. of the IEEE Computer Society Workshop on VLSI*, pp. 9–12, 2000.

[134] M. Haenggi and D. Puccinelli, "Routing in ad hoc networks: a case for long hops," *IEEE Communications Magazine*, vol. 43, pp. 93–101, October 2005.

[135] S. Valentin, H. Lichte, H. Karl, G. Vivier, S. Simoens, J. Vidal, and A. Agustin, "Cooperative wireless networking beyond store-and-forward," *Wireless Personal Communications*, vol. 48, pp. 49–68, January 2009.

[136] X. Tao, X. Xu, and Q. Cui, "An overview of cooperative communications," *IEEE Communications Magazine*, vol. 50, pp. 65–71, June 2012.

[137] E. van der Meulen, "Three-terminal communication channels," *Advances in Applied Probability*, vol. 3, no. 1, pp. 120–154, 1971.

[138] T. Cover and A. el Gamal, "Capacity theorems for the relay channel," *IEEE Transactions on Information Theory*, vol. 25, pp. 572–584, September 1979.

[139] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity. Part I and Part II," *IEEE Transactions on Communications*, vol. 51, pp. 1927–1948, November 2003.

[140] A. Nosratinia, T. Hunter, and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Communications Magazine*, vol. 42, pp. 74–80, October 2004.

[141] A. Nosratinia and T. Hunter, "Grouping and partner selection in cooperative wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, pp. 369–378, February 2007.

[142] T. Ng and W. Yu, "Joint optimization of relay strategies and resource allocations in cooperative cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, pp. 328–339, February 2007.

[143] R. Tannious and A. Nosratinia, "Spectrally efficient relay selection with limited feedback," *IEEE Journal on Selected Areas in Communications*, vol. 26, pp. 1419–1428, October 2008.

[144] J. Sydir and R. Taori, "An evolved cellular system architecture incorporating relay stations," *IEEE Communications Magazine*, vol. 47, pp. 115–121, June 2009.

[145] R. Balakrishnan, X. Yang, M. Venkatachalam, and I. Akyildiz, "Mobile relay and group mobility for 4G WiMAX networks," in *Proc. of the IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1224–1229, 2011.

[146] K. Loa, C.-C. Wu, S.-T. Sheu, Y. Yuan, M. Chion, D. Huo, and L. Xu, "IMT-advanced relay standards," *IEEE Communications Magazine*, vol. 48, pp. 40–48, August 2010.

[147] K. Doppler, *In-band Relays for Next Generation Communication Systems.* PhD thesis, Aalto University, 2010.

[148] *IEEE 802.16j-2009, Amendment to IEEE Standard for Local and metropolitan area networks. Multihop Relay Specification*, June 2009.

[149] N. Abuzainab and A. Ephremides, "Energy efficiency of cooperative relaying over a wireless link," *IEEE Transactions on Wireless Communications*, vol. 11, pp. 2076–2083, June 2012.

[150] J. Luo and A. Ephremides, "On the throughput, capacity, and stability regions of random multiple access," *IEEE Transactions on Information Theory*, vol. 52, pp. 2593–2607, June 2006.

[151] B. Rong and A. Ephremides, "On opportunistic cooperation for improving the stability region with multipacket reception," in *Proc. of the 3rd Euro-NF Conference on Network Control and Optimization (NET-COOP)*, pp. 45–59, 2009.

[152] I. Krikidis, B. Rong, and A. Ephremides, "Network-level cooperation for a multiple-access channel via dynamic decode-and-forward," *IEEE Transactions on Information Theory*, vol. 57, pp. 7759–7770, December 2011.

[153] Y. Park and G. Hwang, “Performance modelling and analysis of the sleep-mode in IEEE 802.16e WMAN,” in *Proc. of the IEEE 65th Vehicular Technology Conference (VTC)*, pp. 2801–2806, 2007.

[154] Y. Park and G. Hwang, “An efficient power saving mechanism for delay-guaranteed services in IEEE 802.16e,” *IEICE Transactions on Communications*, vol. E92-B, pp. 277–287, January 2009.

[155] S. Alouf, E. Altman, and A. Azad, “Analysis of an M/G/1 queue with repeated inhomogeneous vacations with application to IEEE 802.16e power saving mechanism,” in *Proc. of the 5th International Conference on Quantitative Evaluation of Systems (QEST)*, pp. 27–36, 2008.

[156] K. de Turck, S. de Vuyst, D. Fiems, and S. Wittevrongel, “Performance analysis of the IEEE 802.16e sleep mode for correlated downlink traffic,” *Telecommunication Systems*, vol. 39, no. 2, pp. 145–156, 2008.

[157] K. de Turck, S. Andreev, S. de Vuyst, D. Fiems, S. Wittevrongel, and H. Bruneel, “Performance of the IEEE 802.16e sleep mode mechanism in the presence of bidirectional traffic,” in *Proc. of the International Workshop on Green Communications (GreenComm) co-located with IEEE ICC*, 2009.

[158] S. de Vuyst, K. de Turck, D. Fiems, S. Wittevrongel, and H. Bruneel, “Delay versus energy consumption of the IEEE 802.16e sleep-mode mechanism,” *IEEE Transactions on Wireless Communications*, vol. 8, pp. 5383–5387, November 2009.

[159] Z. Saffer and M. Telek, “Analysis of BMAP vacation queue and its application to IEEE 802.16e sleep mode,” *Journal of Industrial and Management Optimization*, vol. 6, no. 3, pp. 661–690, 2010.

[160] A. Anisimov, S. Andreev, O. Galinina, and A. Turlikov, “Comparative analysis of sleep mode control algorithms for contemporary metropolitan area wireless networks,” in *Proc. of the International Conference on Next Generation Wired/ Wireless Advanced Networking (NEW2AN)*, pp. 184–195, 2010.

[161] S. Jin, M. Choi, and S. Choi, “Performance analysis of IEEE 802.16m sleep mode for heterogeneous traffic,” *IEEE Communications Letters*, vol. 14, pp. 405–407, May 2010.

[162] R. Kalle, M. Gupta, A. Bergman, E. Levy, S. Mohanty, M. Venkatachalam, and D. Das, “Advanced mechanisms for sleep mode optimization of VoIP traffic over IEEE 802.16m,” in *Proc. of the IEEE Global Telecommunications Conference (GLOBECOM)*, 2010.

[163] S. Baek and B. Choi, “Performance analysis of sleep mode operation in IEEE 802.16m with both uplink and downlink packet arrivals,” in *Proc. of the 16th IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pp. 112–116, 2011.

[164] C. Bontu and E. Illidge, "DRX mechanism for power saving in LTE," *IEEE Communications Magazine*, vol. 47, pp. 48–55, June 2009.

[165] L. Zhou, H. Xu, H. Tian, Y. Gao, L. Du, and L. Chen, "Performance analysis of power saving mechanism with adjustable DRX cycles in 3GPP LTE," in *Proc. of the IEEE 68th Vehicular Technology Conference (VTC)*, 2008.

[166] S. Baek and B. Choi, "Analysis of discontinuous reception (DRX) with both downlink and uplink transmissions in 3GPP LTE," in *Proc. of the 6th International Conference on Queueing Theory and Network Applications (QTNA)*, 2011.

[167] S. Cui, *Cross-Layer Optimization in Energy Constrained Networks.* PhD thesis, Stanford University, 2005.

[168] J. Turkka and M. Renfors, "Path loss measurements for a non-line-of sight mobile-to-mobile environment," in *Proc. of the 8th International Conference on ITS Telecommunications (ITST)*, pp. 274–278, 2008.

[169] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Electronics Letters*, vol. 27, pp. 2145–2146, November 1991.

[170] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011-2016*, February 2012.

[171] C. Sankaran, "Data offloading techniques in 3GPP Rel-10 networks: a tutorial," *IEEE Communications Magazine*, vol. 50, pp. 46–53, June 2012.

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

[173] M. Dohler, D.-E. Meddour, S.-M. Senouci, and H. Moustafa, *Cooperative Communications for Improved Wireless Network Transmission: Framework for Virtual Antenna Array Applications*, ch. Cooperative Communication System Architectures for Cellular Networks, pp. 522–547. IGI Global, 2010.

[174] 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.

[175] 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.

[176] 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.

[177] 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.

[178] 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.

[179] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, “FlashLinQ: a synchronous distributed scheduler for peer-to-peer ad hoc networks,” in *Proc. of the 48th Annual Allerton Conference on Communication, Control, and Computing*, pp. 514–521, 2010.

[180] 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.

[181] TSG SA WG1, `SP-110638`, *WID on Proposal for a study on Proximity-based Services*, September 2011.

[182] H. Cho and J. Puthenkulam, *Machine to Machine (M2M) Communication Study Report.* `IEEE 802.16ppc-10/0002r6`, May 2010.

[183] Harbor Research, Inc., *Machine-To-Machine (M2M) & Smart Systems Forecast 2010-2014*, 2009.

[184] Finnish Strategic Centre for Science, Technology, and Innovation, *Internet of Things Strategic Research Agenda*, September 2011.

[185] K. David, V. Vinodrai, and J. Yao, *WWRF Introduction and Vision*, November 2010.

[186] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. Johnson, “M2M: from mobile to embedded Internet,” *IEEE Communications Magazine*, vol. 49, pp. 36–43, April 2011.

[187] *IEEE 802.16p/D6, [Draft] Enhancements to Support Machine-to-Machine Applications*, July 2012.

[188] *Study on RAN Improvements for Machine-Type Communications.* 3GPP Technical Report (TR) 37.868, September 2011.

[189] 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.

[190] S.-Y. Lien, T.-H. Liau, C.-Y. Kao, and K.-C. Chen, “Cooperative access class barring for machine-to-machine communications,” *IEEE Transactions on Wireless Communications*, vol. 11, pp. 27–32, January 2012.

[191] D. Boswarthick, O. Elloumi, and O. Hersent, eds., *M2M Communications: A Systems Approach.* Wiley-Blackwell, 2012.

[192] Z. Fadlullah, M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, and Y. Nozaki, "Toward intelligent machine-to-machine communications in smart grid," *IEEE Communications Magazine*, vol. 49, pp. 60–65, April 2011.

[193] D. Niyato, L. Xiao, and P. Wang, "Machine-to-machine communications for home energy management system in smart grid," *IEEE Communications Magazine*, vol. 49, pp. 53–59, April 2011.

[194] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Communications Magazine*, vol. 49, pp. 66–74, April 2011.

[195] Y. Yu and G. B. Giannakis, "SICTA: a 0.693 contention tree algorithm using successive interference cancellation," in *Proc. of the 24th IEEE Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, vol. 3, pp. 1908–1916, 2005.

[196] Y. Yu and G. Giannakis, "High-throughput random access using successive interference cancellation in a tree algorithm," *IEEE Transactions on Information Theory*, vol. 53, pp. 4628–4639, December 2007.

[197] B. Tsybakov and V. Mikhailov, "Free synchronous packet access in a broadcast channel with feedback," *Problems of Information Transmission*, vol. 14, pp. 32–59, October 1978.

[198] J. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Transactions on Information Theory*, vol. 25, pp. 505–515, September 1979.

[199] Qualcomm Incorporated, `R2-105619`, *Simulation Assumptions for MTC and RACH Load Simulation Results for UMTS*, October 2010.

[200] K. Zheng, F. Hu, W. Wang, W. Xiang, and M. Dohler, "Radio resource allocation in LTE-Advanced cellular networks with M2M communications," *IEEE Communications Magazine*, vol. 50, pp. 184–192, July 2012.

Publications

Publication 1

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, no. 14, pp. 1662–1671, 2012.

Energy Efficient Communications for Future Broadband Cellular Networks

Sergey Andreev[†], Pavel Gonchukov, Nageen Himayat, Yevgeni Koucheryavy, and Andrey Turlikov

***Abstract*— Energy efficiency is increasingly important for wireless cellular systems due to the limited battery resources of mobile clients. While modern cellular standards emphasize low client battery consumption, existing techniques do not explicitly focus on reducing power that is consumed when a client is actively communicating with the network. In this paper, we evaluate the performance of the recently introduced power-bandwidth optimization techniques using realistic cellular system simulation model, which is compliant with the methodology proposed for the IEEE 802.16m standard.**

The paper addresses several practical trade-offs associated with the implementation of energy efficient schemes. Our simulation results indicate that energy efficient techniques continue to provide considerable power savings, even when accounting for realistic system parameters and channel environments.

***Index Terms*— energy efficiency, power-bandwidth optimization, wireless cellular networks, IEEE 802.16m.**

I. Introduction

Adoption of wireless technology has become increasingly widespread as new high data rate broadband wireless standards emerge, allowing for improved access to services and applications previously only supported through fixed broadband systems. The Institute of Electrical and Electronics Engineers (IEEE) 802.16 work group and the 3rd Generation Partnership Project (3GPP) are introducing *fourth generation* (4G) metropolitan wireless standards [1] and [2] respectively. However, future success of wireless communication systems significantly depends on the solution to overcome the mismatch between the requested *quality of service* (QoS) and limited *network resources*.

Spectrum is a natural resource that cannot be replenished. As such, the need for its effective use introduces the problem of *spectral efficiency*. Similarly, *energy efficiency* is also becoming increasingly important primarily for small form factor mobile devices due to the growing gap between the available and the required battery capacity, which is demanded by rich and ubiquitous use of multimedia applications [3]. Client energy efficiency, therefore, is an important consideration in wireless system design.

Current standards [4] support reductions in client power consumption through maximizing ”sleep” and ”idle” periods at the clients. However, they do not explicitly focus on active mode power consumption. Given the battery-limited power budget of mobile devices and the high data rate demands of multimedia applications, active mode power consumption is also expected to become an important consideration for wireless system design and standards development [5].

Recent work on active mode energy efficiency has focused on minimizing the transmit power consumption at the client [6] as it comprises a significant portion of the active power budget. This work proposes energy-aware cross layer power-bandwidth optimization approaches to improve energy consumption at the client [7], [8]. Preliminary performance evaluation in the context of uplink *orthogonal frequency-division multiple access* (OFDMA) systems suggests that significant gains in client power efficiency are possible with such techniques [7], [8].

In this paper, we carry out an in-depth performance evaluation of energy efficient techniques using a realistic cellular system simulation model, which is compliant with the most advanced evaluation methodology proposed by the IEEE 802.16m standards group [9]. In particular, we adapt several existing approaches and associated modifications for practical use. Further, we evaluate these to show that significant active mode power savings are possible for wireless clients even under realistic system environments.

The contribution of the paper is therefore the detailed investigation of important practical trade-offs such as dependence of the proposed schemes on circuit/idle power consumption, as well as amplifier efficiency and fairness aspects. We also consider the important relationship between inter-cell interference and power reduction and compare performance of energy efficient schemes with power-control based interference management schemes. The results reported in this paper illustrate the pros and cons associated with applying power-bandwidth optimization approaches for improving client energy efficiency and develop insights for future research in this area.

The organization of the paper is as follows. In Section II, we briefly survey past work and cover important considerations associated with energy efficiency enhancement in cellular systems. Section III outlines the basic cellular system model considered and analytically describes the energy efficient schemes evaluated in this paper. Section IV describes the advanced system level evaluation methodology. Section VI concludes on the performance of the energy efficient schemes using extensive simulation results from Section V and highlights areas for future research.

S. Andreev and Y. Koucheryavy are with Tampere University of Technology, Finland; P. Gonchukov and A. Turlikov are with St. Petersburg State University of Aerospace Instrumentation, Russia; N. Himayat is with Wireless Commun. Lab./Commun. Tech. Lab., Santa Clara, CA, USA

[†] S. Andreev is the contact author: P.O.Box 553, FI-33101, Tampere, Finland; phone: +358 44 329 4200; e-mail: serge.andreev@gmail.com

Manuscript received Jun 29, 2011, revised Jan 16, 2012, and accepted Apr 26, 2012

II. Background and Related Work

A. Cross-Layer Approaches

Currently, layered architecture dominates in networking design and each layer is operated independently to maintain transparency. Among these layers, the *physical* (PHY) layer is responsible for the raw-bit transmission, whereas the *medium access control* (MAC) layer arbitrates access to the shared wireless resources. However, the traditional layer-wise architecture turns out to be inflexible and results in the inefficient wireless resource utilization. An integrated and adaptive design across different layers is thus required to overcome this limitation. As a consequence, *cross-layer* optimization across PHY and the MAC layers is desired for wireless resource allocation and packet scheduling [10].

In cross-layer optimization, *channel-aware* approaches are introduced and developed to explicitly take into account wireless *channel state information* (CSI). Taking advantage of the channel variation across clients, channel-aware approaches are shown to substantially improve the network performance through multi-user diversity, whose gain increases with the number of clients [11].

Since wireless channels are shared and highly dynamic, resource management is believed to be the most challenging element in the design of channel-aware systems [12], [13]. Future schedulers should account for at least three primary performance metrics: overall system capacity (or spectral efficiency), energy consumption (or energy efficiency) of wireless clients, and their quality of service perception [14]. It is also desirable to have a high degree of control over the trade-offs associated with these metrics.

Clearly, spectral and energy efficiency are affected by all components of system design, ranging from radio frequency (RF) circuits to applications. Cross-layer approaches may significantly improve system performance as well as adaptability to service, traffic, and environment dynamics [15], [16], [17]. Cross-layer optimization for throughput improvement has been a popular research direction [18]. However, as wireless clients become increasingly mobile, the focus of recent efforts tends to shift toward energy consumption at all layers of communication systems, from architectures [19] to algorithms [20].

B. Transmit Power Optimization and Metrics

A key consideration in designing energy efficient systems are the metrics used for measuring energy efficiency. Typical metrics used thus far are ”bits-per-Joule” [21] or ”throughput-per-Joule” [22]. Conversely, ”Joules-per-bit” metric may also be used. Whereas link level energy efficient criteria have been discussed in more detail [7], little attention has been paid to the aggregate system level metrics.

In multi-user system environment, several novel metrics that measure system-wide energy efficiency performance across clients may also be defined and we seek to propose some of these below. Fairness is also an important aspect when choosing a metric for performance optimization across several clients. These various choices for energy efficient metrics will be discussed in detail in later sections.

Transmit power consumption can dominate the client's active power consumption budget in cellular systems due to the need to overcome significant path loss for reliable signal reception, as well as the poor efficiency of typical power amplifiers at the client. Therefore, uplink transmit power optimization for the client is a critical aspect of energy efficient design in cellular systems.

Here Shannon's law for a point-to-point link can be used to provide intuition on possible approaches for transmit power reduction:

$$c = f \log\left(1 + \frac{g \cdot p}{\sigma^2}\right), \tag{1}$$

where f is the allocated bandwidth, g is the channel gain, p is the transmit power of a client, σ^2 is the noise power, and c is the achievable capacity for a client.

The channel capacity is known to be the maximum rate at which reliable communication is possible in the system. Given that it is linearly related to bandwidth but exponentially related to power, client transmit energy consumption may be reduced by the following.

- For a fixed rate of transmission, if transmission bandwidth is increased, power can be exponentially decreased in the system.
- For a fixed rate, client experiencing good channel conditions can be scheduled.
- If delay can be tolerated and transmission rate is reduced, power can be exponentially reduced.

Therefore, network can allocate power and bandwidth, and control delay across clients [23] to conserve transmit energy.

Several approaches exist to optimize transmit energy efficiency, which include water-filling power allocation schemes [24], [25], and adaptation of both overall transmit power and its allocation, according to the CSI [26], [27]. However, the client energy efficiency is affected not only by the layers of the point-to-point communication link, but also by the interaction between the individual links. Therefore, as indicated earlier, a cross-layer approach, including both wireless *link adaptation* and multi-user *resource allocation*, is required.

In what follows, we discuss recently introduced cross-layer power-bandwidth optimization techniques and evaluate their performance using realistic system level simulation model.

C. Link Adaptation and Resource Allocation

As the quality of wireless channel varies with time, frequency, and client, link adaptation can be used to improve transmission performance. With it, modulation order, coding rate, and transmit power can be selected according to CSI. Earlier research on link adaptation focuses on power allocation to improve individual channel capacity [28], whereas state-of-the-art approaches emphasize the need for joint link adaptation and resource allocation [22].

More specifically, since channel frequency responses vary for different frequencies and clients, data rate adaptation over each sub-carrier, dynamic sub-carrier assignment, and adaptive power allocation can significantly improve the performance of *orthogonal frequency-division multiplexing* (OFDM) networks. Through data rate adaptation, the transmitter can use higher transmission rates and reduced power consumption over the sub-carriers with better conditions so as to improve throughput [29].

Due to limited wireless resources, intricate performance trade-offs arise between an individual client and the entire network. While extensive efforts have been undertaken to improve energy efficient resource management in *time domain* [30], little effort has been devoted to *frequency domain*. Here, while increasing transmission bandwidth improves energy efficiency, the entire system bandwidth can not be allocated exclusively to one client in a multi-user system since this may hurt the energy efficiency of other clients, as well as that of the overall network [22]. Hence, frequency-domain resource management is critical in determining overall network energy efficiency. Frequency selectivity of broadband wireless channels further accentuates this necessity.

D. Interference Mitigation

Modern wireless networks, especially those with cellular topology [1], [2], are becoming increasingly interference-limited as more clients share the same spectrum to receive high-rate multimedia service. In modern cellular systems, *co-channel interference* (CCI) is expected to become one of the major performance-limiting factors, especially as these systems shift toward aggressive frequency reuse scenarios [31]. A popular CCI mitigation technique is to assign different sets of channels to neighboring cells [32] and a good summary of channel assignment can be found in [33].

While the overall spectral efficiency may indeed improve with aggressive frequency reuse, the performance of cell-edge clients degrades dramatically. Power control is an important method of reducing interference in cellular networks [34], therefore, we expect that techniques designed to control interference will also help reduce power consumption.

Interactions between power-bandwidth optimization techniques designed for throughput versus energy optimization is, therefore, an important consideration in our investigation.

E. Standardization Efforts

As discussed, wireless client energy efficiency has been an important consideration in defining cellular system standards, where protocols for maximizing ”sleep” and ”idle” durations have been included to save client power [4]. The IEEE 802.16m Systems Requirements Document (SRD) [35] requires that the standard supports ”enhanced power savings mode”, as well as reporting mechanisms for communicating power related information. Therefore, the IEEE 802.16m protocol supports several enhancements for sleep and idle mode optimization. The standard also provides hooks for clients to communicate their battery status to the base station so that it can initiate energy savings mechanisms for them. Therefore, cross-layer optimization techniques may easily be utilized in future cellular systems to improve client energy efficiency.

Network energy efficiency to lower power consumed in the network is also becoming important due to environmental concerns as well as due to the operator’s desire to reduce operational costs (see 3GPP-LTE work items on green RAN). However, our focus in this paper would be on client energy efficiency.

III. Energy Efficient Scheme for OFDMA System

A. System Description

Power-bandwidth optimization (see Fig. 1) plays a pivotal role in both interference management and energy utilization. An implicit discussion can be found in [36] and [37], which summarizes existing approaches in the context of power control for *code division multiple access* (CDMA) networks. However, given that the dominant 4G standards are based on OFDMA technology, we focus on techniques applicable for OFDMA systems. Works investigating energy efficient optimization for OFDM and OFDMA communications are [22], [26], [27].

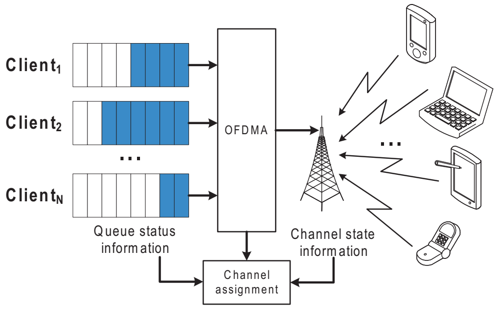


Fig. 1. General power-bandwidth optimization scheme.

For the sake of simplicity here we consider a single cell of a wireless cellular network. There are N subscriber stations (clients) and one base station (BS) in such a system. The BS arbitrates all activity within the cell and may communicate with its clients in the downlink (DL) sub-frames. In the uplink (UL) sub-frames, the clients transmit scheduled data (see Fig. 2).

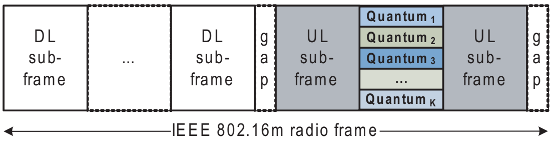


Fig. 2. Basic IEEE 802.16m frame structure.

In what follows, we focus on the uplink channel only, as data transmission consumes much more client power than reception [7], [8]. Channel time is broken into (sub-)frames. Each frame is composed of K frequency sub-channels or *quanta* of resources (see Fig. 2). Exactly one client may transmit its data at one quantum per one frame. However, a client may utilize more than one and up to K different quanta for its data transmission in a single frame. Each client n transmits data on each quantum k with the corresponding attenuation factor g_{nk}.

It is assumed that attenuation factors are known at the BS. They should be taken into account to minimize the subsequent power consumption. Information about quanta assignment may be sent to clients in the DL and clients transmit their packets according to that assignment. Currently, we also assume that the packet buffer of a client is always full. Whereas this assumption invokes the static steady-state cellular system operation, it nevertheless may bring important conclusions about the performance of the energy efficient schemes [38].

B. Energy Efficient Metrics

In order to detail a power-optimal joint resource allocation and link adaptation scheme, energy efficient metrics are discussed in [7]. A measure of average energy efficiency at the client n in the time frame t is the total data size sent by this client by the time t ($D_n[t]$) divided by the total consumed energy ($E_n[t]$):

$$u_n[t] = \frac{D_n[t]}{E_n[t]}. \tag{2}$$

We note that for a fixed throughput, an energy efficient metric should consume the lowest power. Alternately, for a fixed power budget an energy efficient scheme will deliver the maximum throughput. The choice of using average or aggregate system metrics is driven by the desired *low complexity* solutions to the energy efficient power-bandwidth optimization problem [7], [8]. Our evaluation will verify that the low complexity solutions derived from average performance metrics perform close to more complex near optimal *iterative* techniques derived for the instantaneous performance criteria.

Due to the fact that the frames have equal size, equation (2) could be rewritten as:

$$u_n[t] = \frac{T_n[t]}{P_n[t]}, \tag{3}$$

where $T_n[t]$ is the throughput of the client n, $P_n[t]$ is the total consumed power. The $T_n[t]$ and $P_n[t]$ may be calculated recursively by:

$$T_n[t] = T_n[t-1] + r_n[t], \quad \text{and} \tag{4}$$

$$P_n[t] = P_n[t-1] + p_n[t], \tag{5}$$

where $r_n[t]$ is the data rate of the client n at the frame t, $p_n[t]$ is the consumed power by the client n at the frame t.

Thus, energy efficiency shows how many data bits are sent by a client per a Joule of consumed energy (bpJ). We also propose novel important energy efficiency-related metrics which are summarized in Table I, where T is the total system operation time.

The primary task of any energy efficient scheme is to schedule and control client transmissions to maximize a particular energy efficient criterion. In this paper, we consider the following two energy efficient criteria:

1) An arithmetic-mean criterion:

$$U_{AM}[t] = \sum_{n=1}^{N} u_n[t]. \tag{6}$$

2) A geometric-mean criterion:

$$U_{GM}[t] = \sum_{n=1}^{N} \log\left(u_n[t]\right). \tag{7}$$

TABLE I
ENERGY EFFICIENT METRICS

Metric name	Expression
Average system energy efficiency	$\frac{\sum\limits_{t=1}^{T}\sum\limits_{n=1}^{N} r_n[t]}{\sum\limits_{t=1}^{T}\sum\limits_{n=1}^{N} p_n[t]}$
Average energy efficiency per client	$\frac{1}{N}\sum\limits_{n=1}^{N}\frac{\sum\limits_{t=1}^{T} r_n[t]}{\sum\limits_{t=1}^{T} p_n[t]}$
Average energy efficiency per time frame	$\frac{1}{T}\sum\limits_{t=1}^{T}\frac{\sum\limits_{n=1}^{N} r_n[t]}{\sum\limits_{n=1}^{N} p_n[t]}$
Average instantaneous energy efficiency	$\frac{1}{NT}\sum\limits_{t=1}^{T}\sum\limits_{n=1}^{N}\frac{r_n[t]}{p_n[t]}$

The purpose of the optimization problem posed by the above equations is to assign frame quanta and power to clients with pending data packets accounting for their energy efficiency. This problem may be solved through iterative utility based optimization. For further details see [25] and [39]. However, by using the above time-averaged throughput per Joule metric, the iterative solutions summarized in [40], can be replaced by closed-form metrics that can be computed on a per quantum basis [7], greatly simplifying the resource allocation solution. The approach from [7] is summarized in the following subsections.

C. Solution for Energy Efficient Link Adaptation

Consider function $S(r_{nk}[t])$, which represents *signal-to-noise ratio* (SNR) value for the client n at the frame t on the quantum k. In [7], it was shown that the optimal data rate of the client n on the quantum k is established as:

$$r_k^{opt}[t] = \max(S^{'-1}(\frac{g_k[t]}{u[t-1]\cdot\sigma^2}), 0), \tag{8}$$

where $S^{'}$ is the first-order derivative of the function S and S^{-1} is its inverse. The corresponding optimal power allocation on the quantum k is given by:

$$p_k^{opt}[t] = \frac{S(r_k^{opt}[t])\cdot\sigma^2}{g_k[t]}. \tag{9}$$

If Shannon's law (1) is used to approximate data rate on each quantum, $S(r) = 2^{\frac{r}{f}} - 1$, the above simplifies to:

$$p_k^{opt}[t] = \max(\frac{f}{u[t-1]\cdot\log 2} - \frac{\sigma^2}{g_k[t]}, 0). \tag{10}$$

D. Solution for Energy Efficient Resource Allocation

Define C_n^* to be the set of quanta assigned to the client n. Total quanta allocation is therefore:

$$C = \bigcup_{i=1}^{N} C_i^*. \tag{11}$$

An energy efficient scheduler creates such an allocation C that energy efficient criterion is maximized:

$$C : \lim_{t\to\infty} U(C, r, p, t) \to \max. \tag{12}$$

In [7], it is shown that energy efficiency tends to its maximum if quanta allocation is defined as:

$$C_n^* = \{k \,|J(n,k) > J(m,k), \forall m \neq n\}, \forall n, \tag{13}$$

where $J(n, k)$ depends on the selected criterion. $J(n, k)$ is then calculated for the

- arithmetic-mean criterion as:

$$J_{AM}(n,k) = \frac{r_{nk}[t]}{T_n[t-1]} - u_n[t-1]\frac{p_{nk}[t]}{P_n[t-1]}, \tag{14}$$

- geometric-mean criterion as:

$$J_{GM}(n,k) = \frac{r_{nk}[t]}{T_n[t-1]} - \frac{p_{nk}[t]}{P_n[t-1]}, \tag{15}$$

where $r_{nk}[t]$ is the data rate of the client n at the frame t on the quantum k, $p_{nk}[t]$ is the power consumed by the client n at the frame t for the data transmission on the quantum k.

E. Summary

Summarizing, the proposed energy efficient scheme consists of the resource allocation part (scheduling algorithm) and the link adaptation part (power control algorithm). Its operation (see Fig. 3) could be described as follows.

1) The BS calculates $J(n, k)$ metric for all the clients at all quanta accounting for their data rate and the power consumed.
2) For each quantum the BS determines a client with maximum $J(n, k)$ value and assigns the quantum to this client.
3) Information about quanta assignment is sent to the clients.
4) The clients transmit data in the assigned quanta.

The described low complexity energy efficient scheme has the property of increasing the selected energy efficient criterion up to some suboptimal value with time. However, there exist high complexity iterative energy efficient approaches that try to find a near-optimal solution [41]. The performance gap between the low complexity and the iterative algorithms is evaluated as part of this paper.

As can be seen, the low complexity solution to energy efficiency optimization is fairly easy to implement and requires simple modifications to the standard metric computation used for throughput optimization. However, this scheme is purely distributed in the sense that no inter-cell coordination is used. While simple to implement, a distributed approach may ignore the impact of interference from the neighboring cells in its optimization and the performance of the cell-edge clients may degrade as a result. This aspect is evaluated in greater detail in the later sections.

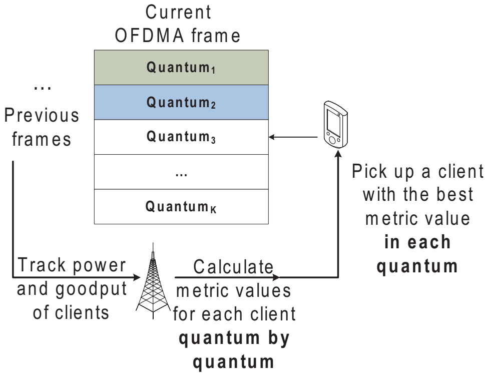

Current
OFDMA frame
Quantum₁
Quantum₂
Quantum₃
...
Quantum_K
...
Previous frames
Pick up a client with the best metric value **in each quantum**
Track power and goodput of clients
Calculate metric values for each client **quantum by quantum**

Fig. 3. Simplified energy efficient scheme operation.

IV. System Level Evaluation Methodology

A. Advanced System Model

In this section, we evaluate the system performance of low complexity energy efficient scheme described in the previous section within a more realistic system model. According to the IEEE 802.16m methodology described in [9], the cellular system is modeled as a network of 19 cells with central target cell surrounded by interfering cells (see Fig. 4). Each cell is hexagonal, with cell radius R determined by the link budget. The cell radius R is defined as half the distance between the centers of the two closest cells in the network.

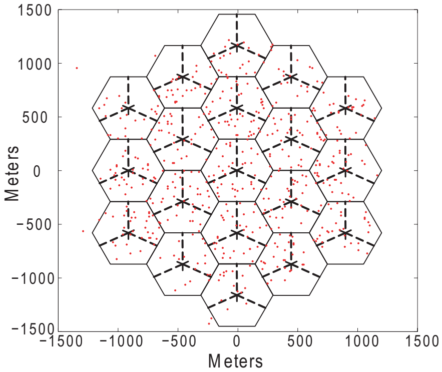


Fig. 4. Example layout of clients.

The System Level Simulator (SLS) creates N_c cells, and every cell may have N_s sectors each with a boresight direction (ϕ_{BS}) and a frequency (F_{BS}). The network is planned with N_a frequency allocations, yielding a frequency reuse pattern of $N_s \times N_a$. The parameters in Table II describe the generic network configuration.

TABLE II
GENERIC NETWORK CONFIGURATION

Parameter	Description	Value
N_c	Number of cells	19
N_s	Number of sectors/cell	3
$N_s \times N_c$	Total number of sectors	57
R	Cell radius	0.5 km
ϕ_{BS}	Orientation (boresight angle) of each sector	$\phi_{BS} = 30, 150, 270$
N_a	Number of frequency allocations in the network (frequency reuse)	1

The SLS assumptions are compliant with IEEE 802.16m evaluation methodology [9]. In particular, Fig. 4 shows an example of client layout within the considered system model. The clients (represented by the dots) are placed randomly inside the simulated cellular system and are then associated with the BSs. Additionally to the *network model* (cells, sectors, and reuse scheme) and *deployment model* (macro/micro cell, fixed/mobile client) described above, the SLS also includes client and BS *equipment model* (power, height, and antenna pattern), various *channel models* (slow fading, fast fading, and spatial model), *air interface model* (OFDMA permutation, frame, and resource allocation), as well as *interference model* (channel of interferer, loading).

We restrict our further explorations to the most interference-limited frequency reuse 3×1 pattern, where the same frequency allocation is deployed in all sectors throughout the network. If aggressive 3×1 reuse is deployed, a client can suffer significant interference from neighboring sectors/cells.

B. Scheduling Algorithms

An OFDMA scheduler at the BS (see Fig. 1) partitions the UL frame resources (quanta) between active clients. The scheduler makes decisions for the entire frame at frame boundaries using available CSI. Its objective is to allocate resources in a fashion that maximizes link utilization while managing QoS requirements [42] and controlling signaling overhead. A single client is assigned a quantum to transmit its data in it. Naturally, the size of the quantum determines the number of clients that can be accommodated within the frame as well as the control signaling overhead. We assume the quantum size is fixed. We also assume the sub-channelization scheme covered by *partial usage of sub-channels* (PUSC), which effectively randomizes the sub-carriers used in a given quantum across different sectors in the system. Such randomization averages the interference across the allocation quanta and renders the effective channel across all quanta to be roughly similar. Hence with the PUSC sub-channelization scheme the channel across the quanta is flat fading and a single channel quality may be applicable for the entire OFDMA band.

A summary of the scheduling algorithms implemented in the SLS is given below.

Round robin (RR): This simplest scheduler is channel-agnostic. It cycles through the active clients, scheduling one quantum to each active client in each cycle.

Maximum SINR (MS): The metric used by this scheduler results in the selection of the client with the highest instantaneous effective *signal-to-interference-plus-noise ratio* (SINR). Although this channel-aware scheduler maximizes aggregate cell throughput, it is inherently unfair. Scheduling delays could also be arbitrarily large for clients with relatively low SINR. Such a metric is, therefore, only suited for applications that can tolerate higher delay.

Proportional fair (PF): At any scheduling instant, the metric $H(n)$ used by the PF scheduler is given by:

$$H(n) = \frac{r_n[t]}{T_n[t]}, \tag{16}$$

where $r_n[t]$ is the rate that can be supported at time t. It is a function of the instantaneous SINR, and consequently of the modulation and coding parameters that can meet the QoS requirement. $T_n[t]$ is the time averaged throughput at the scheduling instant t.

Energy efficient fair (EEF): This energy efficient scheduler adopts the proposed geometric mean criterion given by (7). In particular, from (15) and PUSC properties it follows that the EEF metric is calculated as:

$$J(n) = \frac{r_n[t]}{T_n[t-1]} - \frac{p_n[t]}{P_n[t-1]}. \tag{17}$$

The use of the EEF scheduler results in some fairness when the clients are selected, because if at least one client has never been scheduled the geometric mean takes the value of zero.

Energy efficient (EE): Finally, we also consider the arithmetic mean criterion given by (6) and the respective metric from (14). However, the EE scheduler is also inherently unfair as it selects a client with the highest instantaneous energy efficiency. It is accounted for only to establish the maximum energy efficiency of the system.

C. Power Control Algorithms

We compare the performance of the implemented energy efficient scheme with existing power control methods in place for 4G systems.

Full power (FP): FP is the simplest power control method, where the scheduled clients always transmit with the maximum power allowed.

SINR target (ST): ST is a fixed SINR target method, where power is adjusted at the client to ensure a fixed SINR at the receiver for all clients.

Energy efficient (EE): EE follows the proposed link adaptation algorithm using equation (10).

Simplified maximum sector throughput (SMST): SMST is a variable SINR target [43], where each client has an SINR target depending on its location. Cell-center clients can have a higher SINR target, whereas cell-edge clients have a lower SINR target. The target SINR per client is based on the possible interference caused to other cells by the transmissions of the client. This approach implicitly provides inter-cell coordination to reduce interference. By contrast, other power control schemes operate in distributed manner and do not provide inter-cell coordination. Therefore, it is expected that performance of schemes providing inter-cell coordination may be superior to purely distributed approaches.

D. Energy Efficient Schemes

Fig. 5 illustrates the operation of energy efficient schemes within the SLS. A scheme defines both power control algorithm and scheduling algorithm. In each frame t the proposed energy efficient scheme uses the necessary statistics from the previous frames and iterates through the available quanta. The EE(F) scheduler calculates the $J(n)$ metric values for all the clients in the sector. Notice that now the metric does not depend on the quantum number as for PUSC scheme all the quanta have near-equal quality. The BS then selects a client with the best metric value to transmit over the current quantum.

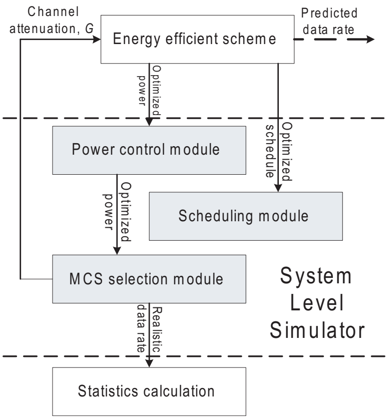


Fig. 5. Energy efficient scheme integration details.

The EE + EEF energy efficient scheme (EE power control and EEF scheduler) is implemented in a way that different clients within a frame may be scheduled. Hence, up to K clients could transmit simultaneously. Using Shannon's law, the scheme "predicts" power and goodput in the already scheduled quanta of the current frame, while processing the remaining quanta. In the course of its operation, the energy efficient scheme interacts with the primary SLS modules: power control, scheduling, and *modulation and coding scheme* (MCS) selection. The SLS defines 11 MCSs given by Table III.

TABLE III
MCS PARAMETERS

MCS number	Convolution code rate	Bits per QAM or QPSK	Number of repetitions	Bits per quantum
1	1/2	2	6	280
2	1/2	2	4	420
3	1/2	2	2	840
4	1/2	2	1	1680
5	3/4	2	1	2520
6	1/2	4	1	3360
7	3/4	4	1	5040
8	1/2	6	1	5040
9	2/3	6	1	6720
10	3/4	6	1	7560
11	5/6	6	1	8400

As the result, for each BS and every client link, the SLS computes the channel and interference power on the loaded data sub-carriers of each quantum. The received signal power level at the sub-carrier k for the target client n is calculated as:

$$p_{rx}(n,k,t) = \frac{p_{tx}(n,t) \cdot d(n) \cdot g(n,k,t)}{q_s}, \qquad (18)$$

where $p_{tx}(n,t)$ is the total transmit power from BS (per sector) or client n, $d(n)$ is the path loss including shadowing and antenna gains, $g(n,k,t)$ is attenuation factor, and q_s is the total number of loaded sub-carriers per quantum.

The CCI power level (from the remaining BSs in DL and clients in UL) at the sub-carrier k of the target client n is calculated as:

$$p^*(n,k,t) = \sum_{l=2}^{q_i} p^*(l,n,k,t) = \qquad (19)$$

$$= \sum_{l=2}^{q_i} \frac{g(l,n,k,t) \cdot d(l,n) \cdot p_{tx}(l,t)}{q_s},$$

where $p^*(l,n,k,t)$ is CCI power level from the interferer l to target client n, and q_i is the number of co-channel interferers. Finally, the SINR of the target client n at the sub-carrier k is:

$$S(n,k,t) = \frac{p_{rx}(n,k,t)}{\sigma_s^2 + p^*(n,k,t)} = \frac{p_{rx}(n,k,t)}{\sigma_s^2 + \sum_{l=2}^{q_i} p^*(l,n,k,t)}, \quad (20)$$

where σ_s^2 is the per sub-carrier noise power level.

The obtained SINR dependence on the number of bits per quantum from Table III is given in Fig. 6. To use equation (10) of the proposed EE link adaptation it is important to coordinate the empirical curve from the SLS with Shannon's law approximation of the data rate. We thus vary the allocated bandwidth parameter f from equation (1) to artificially "reduce" the available bandwidth and thus match the SINR mapping between the SLS and the energy efficient scheme. The resulting curve is also presented in Fig. 6.

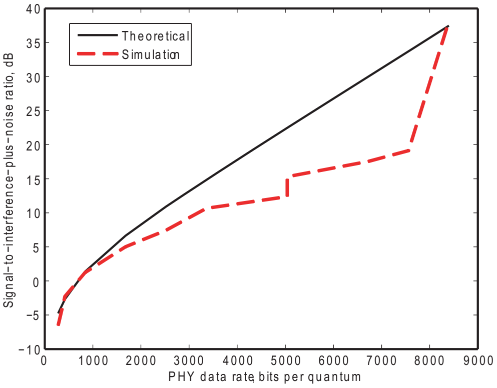


Fig. 6. Proposed SINR mapping function.

V. Practical Features and Results

A. Client Power Profile

The conventional information-theoretic results derived in [44] and [6] focus only on transmit power when considering energy consumption during transmission. Typically, a device will incur additional *circuit power*, which is relatively independent of the transmission rate [29], [45]. Thus, the circuit power consumption should be accounted for explicitly in maximizing energy efficiency.

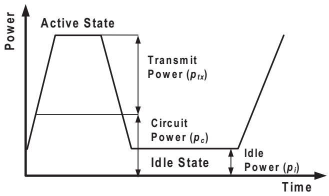


Fig. 7. Typical client power profile.

Consequently, the known method to transmit with the longest duration is no longer the best [46], [47], [48] since circuit energy consumption increases with transmission duration. Considering the impact of circuit power, the focus shifts toward using optimization theory for determining energy-optimal link settings [26]. Accounting for the circuit power, the energy efficiency is reformulated as:

$$u(r) = \frac{r}{p(r)} = \frac{r}{p_c + p_{tx}(r)}, \qquad (21)$$

where r is the data rate and p_c is the constant circuit power of the client respectively.

Summarizing, total wireless power consumption p at the client varies as a function of its state (see Fig. 7), whether idle or active. When a typical client is actively transmitting to the network, it not only consumes RF power in the power amplifier to communicate its signal reliably over the air, but also additional power in the electronic circuitry, which is greater than its idle power consumption p_i. We ensured that within the considered simulation methodology the overall energy consumption of a client is not only affected by the useful power needed for reliable communication, but also the overhead energy consumed due to power consumed in circuit electronics.

Another important power-related issue of wireless cellular system methodology is the consideration of the PHY amplifier operation. In practice, the amplifier works with considerably low efficiency. Therefore, energy efficient scheme should explicitly account for its inefficiency. We conducted the amplifier inefficiency analysis with the simplified approach assuming linear amplifier characteristics. As such, client total power consumption p may be given by:

$$p = \alpha(\beta p_{tx} + p_c) + (1 - \alpha)p_i, \qquad (22)$$

where α is the activity factor, which is equal to 1 each time the client is scheduled in a frame and is equal to 0 otherwise, β is the amplifier coefficient value.

Our SLS simulations show that the increase in the amplifier coefficient by only 25 % causes significant drop of the client total power. Therefore, accounting for the amplifier coefficient in the energy efficient scheme could improve the performance of the proposed mechanism. Simulation results also demonstrate that the explicit consideration of the amplifier coefficient at link adaptation and resource allocation stages of the energy efficient scheme operation considerably reduces consumed client power and, consequently, improves energy efficient performance of the proposed energy efficient scheme.

B. Performance Results

In our simulations, we use goodput, power and energy efficiency performance metrics to compare the EE + EEF energy efficient scheme using the mean geometric criterion (7) against the other 4G approaches. Additionally, we set the system parameters shown in Table IV [34]. In order to obtain performance data with acceptable confidence, each SLS scheme has been simulated for 200 frames (1000 ms) per a Monte-Carlo trial and then the results were averaged across at least 50 independent trials.

TABLE IV

Simulated network configuration

System Parameter	Value
Cell geometry	19 cell system, 3 sectors, reuse 1
Carrier frequency	2.5 GHz
System bandwidth	10 MHz (1024 FFT size)
Power control	Energy efficient (EE), Simplified maximum sector throughput (SMST), SINR target (ST), Full power (FP)
Number of clients per sector	10
Scheduling	Energy efficient (EE), Energy efficient fair (EEF), Proportional fair (PF)
MIMO configuration	1×2
Circuit power, p_c	100 mW
Idle power, p_i	10 mW
Maximum transmit power	23 dBm
Channel model	ITU-Ped B, 3 kmph
Sub-channel permutation	PUSC

The resulting EE + EEF *cumulative distribution functions* (CDFs) are generally compared with those of the other reference schemes described before (see Fig. 8). In particular, we look at FP + PF, ST + PF, and SMST + PF schemes. We notice that currently the EE + EEF scheme favors cell-center clients for the cost of the cell-edge clients. This is different from the results presented in [7]. It was established that such a performance difference is due to the huge variation of channel attenuation factors in the simulated wireless environment.

We also investigate the cell-edge behavior of the energy efficient schemes further, using the spread diagrams shown in Fig. 9. The diagrams illustrate distributions of individual client metrics against the SNR values determined by the client locations. Comparing different schemes, we conclude that EE + EEF scheme performs well for the high quality channel, whereas loses some performance to the SMST + PF scheme when SNR drops. We also notice that EE + EEF scheme is very conservative in assigning resources to SNR-limited clients

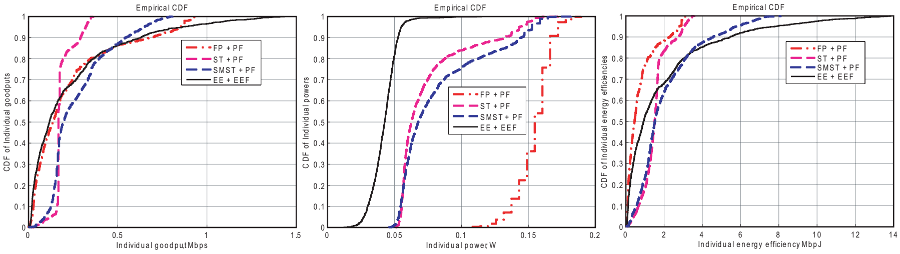


Fig. 8. Empirical goodput, power, and energy efficiency results.

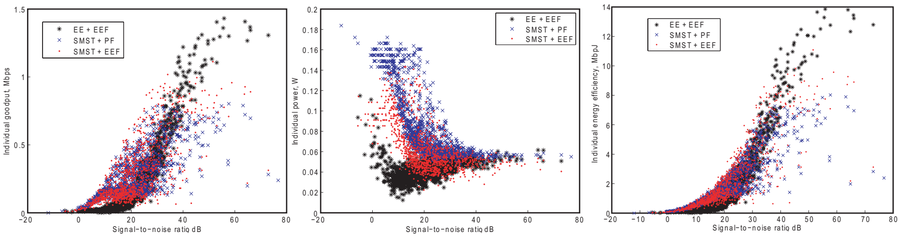


Fig. 9. Empirical goodput, power, and energy efficiency spread diagrams for some schemes.

that may need to transmit with higher power. This property explains the poor cell-edge performance.

C. Fairness Improvement

The shape of the goodput CDF (see Fig. 8, left) for EE + EEF scheme shows that there is a strong variation in goodput across the clients. For the IEEE 802.16m standard it is sometimes beneficial to have a more ”vertical” CDF shape to guarantee that the clients QoS requirements are equally satisfied. Therefore, the discordance between the individual client goodputs should be decreased. Currently, the mean geometric criterion (7) does not explicitly minimize this discordance. Therefore, the optimization criterion might be reconsidered. There is a variety of candidate criteria of which the fairness index by R. Jain [49] appears to be the most suitable one:

$$\gamma = \frac{\left(\sum\limits_{i=1}^{N} X_i\right)^2}{N\sum\limits_{i=1}^{N} X_i^2}, \tag{23}$$

where X is the parameter of interest.

We propose a modification of the energy efficient criterion capable of trading system performance for client fairness. Remember that when the existing EEF scheduler is run, the $J(n,k)$ metric (15) is calculated for each client. We propose to account for an additional metric:

$$F(n,k) = \frac{D_n[t-1]}{D_s[t-1]}, \tag{24}$$

where $D_n[t-1]$ is the total data size sent by client n up to the time t and $D_s[t-1]$ is the total data size sent by the entire system up to the time t.

As such, we only schedule clients that have maximum $J(n,k)$ value and at the same time $F(n,k)$ of which does not exceed some threshold value. When the threshold is small enough, the goodput CDF is more fair, which is, however, achieved for the cost of some overall system performance degradation. Therefore, Table V summarizes EE + EEF results with high threshold.

It can be seen from Table V, that our overall simulation results report the average goodput gain of 22 %, average power gain of up to 43 %, and average energy efficiency gain of up to 54 % for the proposed EE + EEF scheme. Here, system-wide and per-client energy efficiencies are calculated using *system* and *instantaneous* metric (see Table I) respectively. Additionally, we establish the fairness of the considered approaches using the index (23) and include the inherently unfair EE + EE scheme only to assess that the EE + EEF scheme does not lose much performance in comparison.

We have also shown through extensive simulations that the EE + EEF low complexity scheme performs close to near optimal iterative solution from [41] even with realistic system and channel assumptions. Summarizing, the scheme from [41] adapts both overall transmit power and its allocation according to the states of all sub-channels and circuit power consumption to maximize energy efficiency in frequency-selective channels. It has been implemented into the SLS in a similar way as the EE + EEF low complexity scheme, but the respective simulation time is much longer due to iterative performance. For the same configuration scenario, the average goodput degradation was within 5 %, the average power increased by 23 % and the average energy efficiency was reduced by 27 %.

TABLE V

OVERALL SIMULATION RESULTS

Scheme	Goodput, Mbps			Power, W			Energy efficiency, MbpJ		
	System-wide	Per-client	Fairness	System-wide	Per-client	Fairness	System-wide	Per-client	Fairness
FP + PF	129.51	0.22	46 %	88.14	0.15	99 %	1.46	0.75	46 %
ST + PF	106.70	0.18	91 %	43.03	0.07	87 %	2.47	1.53	82 %
EE + EEF	**130.29**	**0.22**	**40 %**	**24.12**	**0.04**	**95 %**	**5.40**	**1.96**	**36 %**
SMST + PF	153.57	0.26	71 %	48.26	0.08	86 %	3.18	2.04	61 %
SMST + EEF	**165.15**	**0.28**	**63 %**	**35.97**	**0.06**	**87 %**	**4.59**	**2.22**	**54 %**
EE + EE	194.84	0.34	17 %	17.32	0.03	86 %	11.24	3.15	16 %

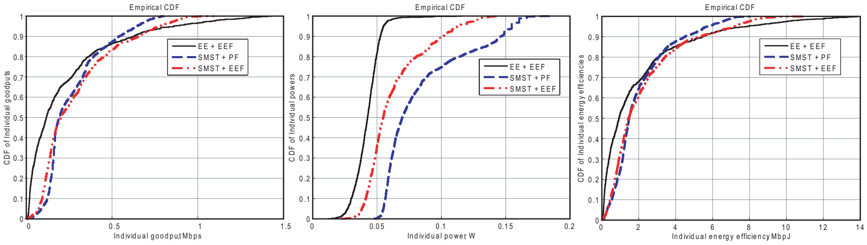


Fig. 10. Empirical goodput, power, and energy efficiency results for interference-aware and EE + EEF schemes.

D. *Combining with Interference-Aware Power Control*

The performance of wireless cellular systems is always affected by the other-cell interference. The existing EE + EEF scheme is based on completely distributed processing which self-regulates the interference caused to adjacent cells. As such, the performance of the cell-edge clients is impacted.

As can be seen, one of the strongest performing SLS power control algorithms is SMST [43]. SMST + PF scheme shows good cell-edge performance (see Fig. 10) and, additionally, provides a flexible trade-off control between overall system throughput and cell-edge performance. This performance enhancement is due, in part, to the implicit inter-cell coordination provided by the SMST. Therefore, we extend the EE + EEF scheme to account for inter-cell information by considering a combination of SMST power control with EEF scheduler to further improve performance.

Our results confirm that there is some cell-edge goodput gain for the combined SMST + EEF scheme. This gain is explained by the cross-sector optimization of the SMST power control. Summarizing, the simulations report the average goodput gain of 54 %, average power gain of up to 16 %, and average energy efficiency gain of up to 46 % for the SMST + EEF scheme. Therefore, utilizing inter-cell coordination schemes is a prominent research direction toward the enhancement of the existing energy efficient solutions. Going back to the spread diagrams shown earlier, we can see from Fig. 9, that the combined SMST + EEF scheme demonstrates excellent performance by maintaining the balance between the cell-edge and the cell-center clients.

VI. CONCLUSION

In this paper, we studied the system level performance of several energy efficient power and resource optimization solutions for OFDMA-based wireless cellular networks. The IEEE 802.16m evaluation methodology was used to investigate a reference 4G system. Also the realistic system parameters, channel environment and implementation considerations were addressed. In particular, the low complexity energy efficient scheme proposed in [7] and [8] was evaluated and shown to perform similar to near-optimal, but significantly more complex, iterative approach. Performance comparison with existing state of the art throughput efficient power optimization schemes was also considered.

Through extensive simulations we showed that the energy efficient schemes demonstrate significant power savings across the cell (greater than 70 % comparing to the fixed-power approaches), and are more energy efficient in terms of bits per Joule metric for cell-center clients. The performance for cell-edge clients requires further improvement, which may be provided through use of ”fairer” metrics or by combining the energy efficient scheduler with the interference-aware power control. Our simulations also show that energy efficient schemes perform well with practical amplifier efficiencies in the range of 10-20 %.

The system level performance characterization of energy efficient wireless transmission techniques studied in this paper appears to be the first of its kind and indicates significant promise for this research area. Future extensions of this work need to focus on more advanced system models and algorithms. In particular, the full-buffer assumption, used in this paper needs to be relaxed and traffic models for multimedia services need to be included [50]. Queuing models, arrival flows, and traffic-aware energy efficient scheduling must also be included.

It should also be noted that IEEE 802.16m standard now provides specific hooks for mobile devices to initiate active mode power savings. The results reported herein were important toward enabling this feature in the standards [34] and may be investigated for enhancing client energy efficiency in future cellular system implementations.

ACKNOWLEDGMENTS

This work is supported by Tampere Graduate School in Information Science and Engineering, Nokia Foundation, HPY Research Foundation, and Intel Corporation. The authors would also like to acknowledge helpful discussions with Guowang Miao and Prof. Geoffrey Li of Georgia Tech.

REFERENCES

[1] IEEE Std 802.16m-2011, Air Interface for Broadband Wireless Access Systems – Advanced Air Interface, May 2011.

[2] LTE Release 10 & beyond (LTE-Advanced), see `http://www.3gpp.org/lte-advanced`.

[3] K. Lahiri, A. Raghunathan, S. Dey, and D. Panigrahi, "Battery-driven system design: A new frontier in low power design," Proc. Intl. Conf. on VLSI Design, pp. 261-267, Jan. 2002.

[4] IEEE Std 802.16-2009. Part 16: Air Interface for Broadband Wireless Access Systems, May 2009.

[5] N. Himayat et al., Improving Client Energy Consumption in 802.16m, C802.16m-09/107, January 2009.

[6] F. Meshkati, H. Poor, S. Schwartz, and N. Mandayam, "An energy-efficient approach to power control and receiver design in wireless networks," IEEE Trans. Commun., vol. 5, pp. 3306-3315, Nov. 2006.

[7] G. Miao, N. Himayat, Y. G. Li, and S. Talwar, "Low-complexity energy-efficient OFDMA," Proc. of the IEEE Int. Conf. on Commun., 2009.

[8] G. Miao, N. Himayat, Y. Li, S. Talwar, "Low-complexity energy-efficient scheduling for uplink OFDMA," IEEE Trans. Commun., vol. 60, pp. 112-120, 2012.

[9] J. Zhuang, L. Jalloul, R. Novak, and J. Park, IEEE 802.16m Evaluation Methodology Document (EMD). IEEE 802.16 Contribution 802.16m-08/004r5, January 2009.

[10] S. Shakkottai, T. Rappaport, and P. Karlsson, "Cross-layer design for wireless networks," IEEE Commun. Magazine, vol. 41, pp. 74-80, Oct. 2003.

[11] R. Knopp and P. Humblet, "Information capacity and power controlling single-cell multiuser communications," Proc. IEEE Int. Conf. on Commun., pp. 331-335, Jun. 1995.

[12] Z. Jiang, Y. Ge, and Y. Li, "Max-utility wireless resource management for best effort traffic," IEEE Trans. on Wireless Commun., vol. 4, no. 1, pp. 100-111, Jan. 2005.

[13] H. Kim and G. de Veciana, "Leveraging dynamic spare capacity in wireless systems to conserve mobile terminals energy," IEEE/ACM Trans. on Network., vol. 1, pp. 1–14, 2008.

[14] F. Meshkati, H. Poor, S. Schwartz, and R. Balan, "Energy-Efficient Resource Allocation in Wireless Networks with Quality-of-Service Constraints," IEEE Trans. Commun., vol. 57, no. 11, pp. 3406-3414, Nov. 2009.

[15] G. Song, Cross-Layer Optimization for Spectral and Energy Efficiency. PhD thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology, 2005.

[16] G. Miao, Cross-Layer Optimization for Spectral and Energy Efficiency. PhD thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology, 2008.

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

[18] G. Song and Y. Li, "Asymptotic throughput analysis for channel-aware scheduling," IEEE Trans. Commun., vol. 54, no. 10, pp. 1827-1834, Oct. 2006.

[19] L. Benini, A. Bogliolo, and G. de Micheli, "A survey of design techniques for system-level dynamic power management," IEEE Trans. VLSI Syst., vol. 8, pp. 299-316, Jun. 2000.

[20] C. Schurgers, Energy-Aware Wireless Communications. PhD thesis, University of California Los Angeles, 2002.

[21] V. Rodoplu and T. Meng, "Bits-per-Joule Capacity of Energy-limited Wireless Networks," IEEE Trans. Wireless Commun., vol. 6, no. 3, pp. 857-865, Mar. 2007.

[22] G. Miao, N. Himayat, Y. Li, and A. Swami, "Cross-layer optimization for energy-efficient wireless communications: A survey," Wiley J. Wireless Commun. and Mob. Comp., vol. 9, no. 4, pp. 529-542, Apr. 2009.

[23] F. Meshkati, H. Poor, S. Schwartz, and R. Balan, "Energy Efficiency and Delay Quality-of-Service in Wireless Networks," Proc. Inaugural Workshop of the Center for Information Theory and Its Applications, 2006.

[24] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks – part I: theoretical framework," IEEE Trans. Wireless Commun., vol. 4, no. 2, pp. 614-624, 2005.

[25] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks – part II: algorithm development," IEEE Trans. Wireless Commun., vol. 4, no. 2, pp. 625-634, 2005.

[26] G. Miao, N. Himayat, and Y. Li, "Energy-efficient link adaptation in frequency-selective channels," IEEE Trans. Commun., vol. 58, no. 2, pp. 545-554, Feb. 2010.

[27] G. Miao, N. Himayat, Y. Li, A. Koc, and S. Talwar, "Interference-aware energy-efficient power optimization," Proc. IEEE Int. Conf. on Commun., Jun. 2009.

[28] Y. Li and G. Stuber, OFDM for Wireless Communications. Springer, 2006.

[29] A. Wang, S. Cho, C. Sodini, and A. Chandrakasan, "Energy efficient modulation and MAC for asymmetric RF microsensor system," Proc. Int. Symp. Low Power Electronics and Design, pp. 106-111, 2001.

[30] R. Mangharam, R. Rajkumar, S. Pollin, F. Catthoor, B. Bougard, L. van der Perre, and I. Moeman, "Optimal fixed and scalable energy management for wireless networks," Proc. IEEE INFOCOM 2005, vol. 1, pp. 114-125, Mar. 2005.

[31] M. Necker, "Coordinated fractional frequency reuse," Proc. ACM Symp. on Mod., Anal., and Sim. of wireless and Mob. Syst., pp. 296-305, 2007.

[32] G. Cao and M. Singhal, "An adaptive distributed channel allocation strategy for mobile cellular networks," J. Parallel and Dist. Comput., vol. 60, pp. 451-473, 2000.

[33] G. L. Stuber, Principles of Mobile Communication. Norwell, MA: Kluwer Academic Publishers, 2001.

[34] N. Himayat et al., Amendment Text Proposal for Section 10.5.3 on Power Management for Connected Mode, IEEE C802.16m-09/0553r2, March 2009.

[35] IEEE 802.16m-07/002r7, IEEE 802.16m Systems Requirements Document (SRD), December 2008.

[36] F. Meshkati, H. Poor, and S. Schwartz, "Energy-efficient resource allocation in wireless networks," IEEE Commun. Magazine, pp. 58-68, May 2007.

[37] F. Meshkati, H. Poor, and S. Schwartz, "Energy-Efficient Resource Allocation in Wireless Networks: An Overview of Game-Theoretic Approaches," IEEE Sign. Proc. Magazine, May 2007.

[38] S. Andreev, Y. Koucheryavy, N. Himayat, P. Gonchukov, and A. Turlikov, "Active-Mode Power Optimization in OFDMA-Based Wireless Networks," Proc. IEEE BWA Workshop Globecom, Dec. 2010.

[39] G. Miao, N. Himayat, Y. G. Li, and D. Bormann, "Energy efficient design in wireless OFDMA," Proc. IEEE Int. Conf. on Commun., vol. 1, pp. 3307-3312, 2008.

[40] G. Miao and N. Himayat, "Low complexity utility based resource allocation for 802.16 OFDMA systems," Proc. of the IEEE Wireless Commun. and Network. Conf., pp. 1465-1470, 2008.

[41] G. Miao, N. Himayat, and Y. G. Li, "Energy-efficient transmission in frequency-selective channels," Proc. of the IEEE GLOBECOM, 2008.

[42] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless networks," IEEE Commun. Magazine, vol. 43, no. 12, pp. 127-135, Dec. 2005.

[43] R. Yang et al., Uplink Open Loop Power Control Recommendations for IEEE 802.16m Amendment, IEEE C802.16m-0546, March 2009.

[44] S. Verdu, "Spectral efficiency in the wideband regime," IEEE Trans. Inf. Theory, vol. 48, pp. 1319-1343, June 2002.

[45] S. Cui, A. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," IEEE Trans. Wireless Commun., vol. 4, pp. 2349-2360, Sep. 2005.

[46] H. Kim, C.-B. Chae, G. de Veciana, and J. R. W. Heath, "Energy-efficient adaptive MIMO systems leveraging dynamic spare capacity," Proc. of the Conf. on Inform. Sciences and Syst., vol. 1, pp. 68–73, 2008.

[47] H. Kim, C.-B. Chae, G. de Veciana, and R. Heath Jr., "A Cross-layer approach to energy efficiency for adaptive MIMO systems exploiting Spare Capacity," IEEE Trans. Wireless Commun., vol. 8, pp. 4264-4275, Aug. 2009.

[48] H. Kim and G. de Veciana, "Leveraging Dynamic Spare Capacity in Wireless Systems to Conserve Mobile Terminals' Energy," IEEE Trans. on Networking, Jul. 2010.

[49] R. Jain, D. M. Chiu, and W. Hawe, "A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Systems," DEC Research Report TR-301, 1984.

[50] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed alpha-Optimal User Association and Cell Load Balancing in Wireless Networks," IEEE/ACM Trans. on Networking, vol. 20, pp. 177-190, 2012.

Publication 2

S. Andreev, Z. Saffer, A. Turlikov, and A. Vinel, “Upper bound on overall delay in wireless broadband networks with non real-time traffic,” in *Proc. of the 17th International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA)*, pp. 262–276, 2010.

Upper Bound on Overall Delay in Wireless Broadband Networks with Non Real-Time Traffic

Sergey Andreev[1], Zsolt Saffer[2],
Andrey Turlikov[3], and Alexey Vinel[4]

Tampere University of Technology[1] (TUT), FINLAND
sergey.andreev@tut.fi
Budapest University of Technology and Economics[2] (BUTE), HUNGARY
safferzs@hit.bme.hu
State University of Aerospace Instrumentation[3] (SUAI), RUSSIA
turlikov@vu.spb.ru
Saint-Petersburg Institute for Informatics and Automation[4]
Russian Academy of Sciences (SPIIRAS), RUSSIA
vinel@ieee.org

Abstract. In this paper we consider the non real-time traffic in IEEE 802.16-based wireless broadband networks with contention-based bandwidth reservation mechanism. We introduce a new system model and establish an upper bound on the overall data packet delay. The model enables symmetric Poisson arrival flows and accounts for both the reservation and the scheduling delay components. The analytical result is verified by simulation.

Keywords: IEEE 802.16, queueing system, Markov chain, overall delay, contention-based request mechanism

1 Introduction and Background

IEEE 802.16 telecommunication protocol [1] defined by the respective networking standard specifies a high data rate wireless broadband network with inherent support for various multimedia applications. Media access control (MAC) layer of IEEE 802.16 provides unified service for the set of physical (PHY) layer profiles, each of which corresponds to a specific operation environment. Currently we observe the proliferation of IEEE 802.16-based networks due to their relatively low cost, wide coverage and MAC mechanisms supporting a variety of quality of service (QoS) requirements.

Performance evaluation of IEEE 802.16 QoS mechanisms is addressed by numerous research papers. In particular, the so-called bandwidth reservation stage is often considered, at which a network user can reserve a portion of the channel resources. A general description of the different reservation techniques can be found in [2]. IEEE 802.16 protocol allows the usage of random multiple access (RMA) for bandwidth requesting and it specifies the truncated binary exponential backoff (BEB) algorithm as means of collision resolution between the requests.

Asymptotic behavior of the BEB algorithm has been thoroughly investigated in the scientific literature. In [3] the BEB algorithm was shown to be unstable in the infinitely-many user model. By contrast, in [4] the BEB algorithm was demonstrated to be stable for sufficiently small arrival rates and finitely-many user model, even for the high number of users. Infinitely-many user model is known to highlight the limiting performance metrics of the algorithm, whereas finitely-many user model provides insight to the practical applicability of the algorithm. Finally, the operation of the BEB algorithm in the saturation conditions, where every network user always has pending data packets, was investigated by means of Markov models in [5] and [6].

Together with the separate analysis of the BEB collision resolution algorithm itself, its proper usage in the framework of IEEE 802.16 system is of interest. According to IEEE 802.16 protocol the BEB algorithm works with broadcast and multicast polling mechanisms (see [7] for details). The performance evaluation of broadcast polling was studied in [8]. Several important BEB application scenarios for the delay-sensitive traffic were discussed in [9].

The overall packet delay is strongly influenced by the choice of an appropriate bandwidth reservation mechanism. In [10] an efficient RMA algorithm is proposed, which may serve as an alternative to the standardized BEB algorithm at the reservation stage. IEEE 802.16 imposes no limitations on the methods for processing the bandwidth requests from the network users. Consequently, numerous scheduling algorithms were proposed.

For instance, in [11] a prioritized scheme for the request processing is developed together with the dynamic on-demand channel resource allocation. The performance of the proposed scheduling is also analyzed. A novel reservation algorithm is considered by [12], for which the corresponding analytical model is detailed. The model allows the evaluation of the reservation delay, but the scheduling delay is not addressed. Finally, in [13] an approach to estimate the overall packet delay is demonstrated. However, the scheduler-independent results there are approximations and thus they give only a rough delay estimate.

Therefore, we conclude that there is a lack of adequate models for the overall packet delay evaluation, including both reservation and scheduling delay. In our previous work [14] we gave an approximation for the overall packet delay. As a continuation of it in this paper we propose an analytical model that provides an upper bound on the overall data packet delay in IEEE 802.16 network.

2 IEEE 802.16 Short Summary

IEEE 802.16 standard specifies both PHY and MAC layers and provides dynamic resource allocation via bandwidth requesting and scheduling. Two operation modes are supported, where the point-to-multipoint mode is mandatory and the mesh mode is optional. MAC structure is composed of the three hierarchical sub-layers.

At the convergence sub-layer IP, ATM and Ethernet traffic is processed uniformly. At the common part sub-layer five different QoS profiles are defined. Various traffic flows with respective QoS requirements are mapped onto these profiles. According to the MAC specification the data packets may vary in size, subject to the proper aggregation/fragmentation. At the privacy sublayer the data encryption service is provided, as well as some additional cryptographic mechanisms.

The baseline PHY technology of IEEE 802.16 is orthogonal frequency division multiplexing (OFDM). Two OFDM-based schemes are defined: plain and OFD multiple access (OFDMA). Both schemes support adaptive modulation and coding to ensure reliable transmission under multipath propagation and over long distances. The growing number of IEEE 802.16 implementations are OFDMA-based, as OFDMA results in the higher spectral efficiency. However, consideration of OFDMA scheme is complicated due to the higher number of parameters and therefore we restrict our further explorations to the plain OFDM scheme.

The core IEEE 802.16 architecture comprises a base station (BS) and a set of subscriber stations (SSs) in its vicinity (see Figure 1). BS performs the polling of the SSs and manages the scheduling of SSs transmissions ensuring that the QoS guarantees of each data flow at each SS are satisfied. The BS and the SSs exchange packets via disjoint communication channels. In the downlink channel the BS broadcasts data to the SSs, whereas in the uplink channel the transmissions from the SSs are multiplexed.

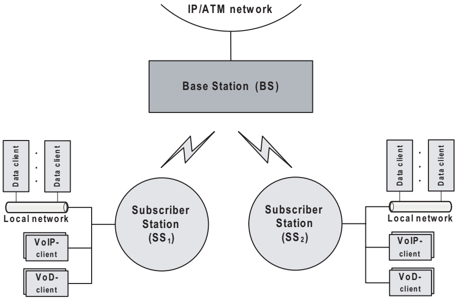


Fig. 1. Core IEEE 802.16 architecture

IEEE 802.16 provides two duplexing modes for the aforementioned downlink and uplink channels. In the time division duplex (TDD) mode a time frame is divided into downlink and uplink sub-frames, respectively. The simplified TDD frame structure is shown in Figure 2. In the frequency division duplex (FDD) mode the channel frequency range is divided into non-overlapping sub-ranges to avoid cross-interference.

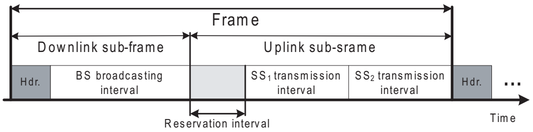


Fig. 2. Simplified TDD frame structure

As mentioned above, the BS broadcasts information to the wirelessly connected SSs. Together with the data packets, BS also sends relevant scheduling information for both downlink and uplink channels. The uplink sub-frame schedule is incorporated into the UL-MAP (uplink map) management packet of the downlink sub-frame and is used by the SSs to determine the start time of their transmission in the uplink sub-frame. In order to enable the SSs to indicate their bandwidth needs to the BS, the so-called reservation interval, a portion of channel resources, is provided. The SSs are allowed to send their bandwidth requests during this interval. These requests are processed in the course of the scheduling.

There is a set of bandwidth requesting mechanisms at the reservation stage. Unicast polling is a contention-free mechanism, according to which the BS provides each SS with one transmission opportunity in a number of frames. Once provided, the transmission opportunity is used by the SS to send its bandwidth request. By contrast, broadcast and multicast polling are contention-based mechanisms. When broadcast polling is enabled the BS provides a number of transmission opportunities and each SS chooses one of them randomly. In case of multicast polling the SSs are grouped and broadcast polling is applied individually to each group. Simultaneous request transmissions may arise, if two or more SSs choose the same transmission opportunity to send their requests. Such request collisions are subject to the subsequent resolution by the BEB algorithm. Piggybacking feature allows an SS to append its bandwidth request to the transmitted data packet, when a connection to the BS is established.

As discussed previously, IEEE 802.16 successfully manages various multimedia connections. It is equally suitable for both high data rate (VoIP, audio and video) and low data rate (web) applications. The protocol supports bursty data flows and delay-sensitive traffic. In order to ensure the satisfaction of the QoS requirements for all these applications IEEE 802.16 standard introduces five QoS profiles. In particular, each profile specifies the type of a bandwidth requesting mechanism (contention-free and/or contention-based) to be used. Summarizing, a data flow with a dedicated identifier (ID) is mapped onto one of the following QoS profiles:

1. Unsolicited grant service (UGS). Used for real-time data sources with constant bit-rate (VoIP traffic without silence suppression). Uplink channel resource is granted periodically without explicit reservation.

2. Real-time polling service (rtPS). Used for real-time data sources with variable bit-rate (MPEG traffic). Uplink channel reservation is organized via unicast polling.
3. Extended real-time polling service (ertPS). Used for real-time data sources with variable bit-rate, which require more strict delay and throughput guarantees (VoIP traffic with silence suppression). This profile is introduced in one of the latest versions, IEEE 802.16e-2005 [15]. Uplink channel reservation is performed via unicast, multicast or broadcast polling.
4. Non real-time polling service (nrtPS). Used for non real-time data sources with variable packet length (FTP traffic). The allowed uplink channel reservation mechanisms are unicast, multicast or broadcast polling.
5. Best effort (BE). Used for non real-time data sources, which do not require delay and throughput guarantees (HTTP traffic). This profile utilizes the remaining bandwidth after scheduling all the above profiles. Multicast or broadcast polling can be used for uplink channel reservation.

Remember that all uplink transmissions are controlled by the BS scheduler. After a new data flow is mapped onto a particular QoS profile (UGS, rtPS, ertPS, nrtPS or BE) the SS proceeds with the uplink channel reservation by sending the corresponding bandwidth request. The BS sends back an UL-MAP management packet in the downlink sub-frame, which indicates the portion of the uplink sub-channel reserved for sending data packets.

The above summary implies that contention-based polling is the most widespread reservation mechanism in IEEE 802.16. Moreover, it is more difficult to analyze it due to its randomized nature in comparison to the analysis of the contention-free mechanism [16], [17]. Below we formulate a set of assumptions and detail the joint model to account for both the reservation and the scheduling stages.

3 System Model

In this section we describe the detailed model of IEEE 802.16-based network, which is used to evaluate the delay at both the reservation and the scheduling stages.

We consider the system that comprises a BS and M SSs, in which we focus only on the uplink transmissions. The BS is in the transmission range of all its SSs and all the SSs are in the reception range of the BS. In order to make the further analysis tractable we impose the following restrictions on the system operation according to IEEE 802.16 protocol description:

Restriction 1. The system operates in the point-to-multipoint mode.

Restriction 2. The time division duplex mode and the plain OFDM PHY scheme are used.

Restriction 3. The delay analysis is conducted for nrtPS QoS profile only, but both nrtPS and BE QoS profiles are considered.

Restriction 4. Only contention-based polling schemes are considered. We concentrate on the broadcast polling.

The system operation time is divided into frames and T_{frame} denotes the frame duration. The consecutive frames are indexed by integer nonnegative numbers, $t = 0, 1, \ldots$. The duration of the packet transmission is τ. The packets arriving to SS i are also referred to as i-packets. At each SS the packet arrival process is Poisson. For simplicity we consider only symmetric arrival flows. Hence at each SS the arrival rate is the same, λ. Thus the overall arrival rate is $\Lambda = \lambda M$. The duration of each contention-based transmission opportunity is α. Moreover the reservation interval of each frame comprises exactly K contention-based transmission opportunities. A bandwidth request is issued by the i-th SS whenever at least one new data packet arrives, of which the BS should be notified. The request contains the information about all the newly arrived packets since the last request sending. If a packet arrives to an empty outgoing buffer of SS i during the reservation interval the SS must wait with sending the bandwidth request for this packet until the next reservation interval.

Additionally, below we introduce a set of assumptions to shape the system model. As such, we use the modified classical multiple access model, which is known from the substantial literature on multiple access techniques and applications, e.g. [18] and [19]. It is often addressed to compare various multiple-access protocols uniformly and has proved its usefulness over passing years.

1. The system
   - The number of contention-based transmission opportunities, K, is constant throughout the system operation.
   - The piggybacking is not used.
2. The BS
   - The BS maintains an individual grant buffer for each SS.
   - The individual BS buffers of the SSs have infinite capacity.
3. The SSs
   - Each SS is supplied with a infinite buffer to store data packets.
   - Each SS maintains exactly one active nrtPS connection.
   - Each SS can transmit exactly one packet in each uplink sub-frame.
4. The channel
   - The channel propagation time is negligible.
   - The uplink channel is noise-free. Consequently, if an SS transmits the BS always receives successfully. The downlink channel is also noise-free. Thus, all the SSs successfully receive the schedule of their transmissions.
   - In each contention-based transmission opportunity only one of the following events may arise at the same time: a single SS transmits its bandwidth request (SUCCESS), none of the SSs transmit (EMPTY), two or more SSs transmit their request simultaneously (COLLISION).
5. The feedback
   - The feedback for each SS about the success/failure of its own bandwidth request transmission (SUCCESS or NON-SUCCESS) is available. This feedback is necessary for the BEB algorithm operation.
   - The notification about the success of the bandwidth request transmissions is provided by the BS at the beginning of the following frame, that is, once in K transmission opportunities.

The BS uses the individual buffer of SS i to store the information about the number and the order of the i-packets (see Figure 3). At the end of each contention-based transmission opportunity the BS process a successfully received request, if any. The information about the newly arrived i-packets is extracted and placed into the corresponding BS buffer. Instead of each i-packet consideration it is equivalent to consider a grant assigned to it. These grants are placed into the individual BS grant buffer of SS i in the order of their extraction from the bandwidth request. This guarantees the first-come-first-served service.

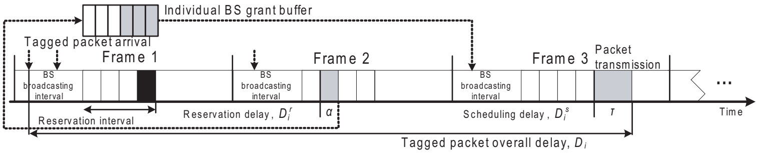


Fig. 3. An example of the request processing procedure

The BS processes the grants from the i-th buffer one by one until the buffer empties. During a frame only one grant can be processed from each grant buffer. When a grant is processed the BS forwards the scheduling information to the corresponding SS in the next frame for the uplink transmission of the corresponding packet. In case the i-th BS grant buffer was empty upon the reception of the new bandwidth request from SS i, the BS starts the service of the grants placed in the buffer immediately. Thus, the i-packet corresponding to the first i-grant will be included into the uplink schedule in the next frame.

When one or more SSs has empty BS grant buffer the BS utilizes the unused uplink transmission capacity to schedule BE packets. Such a behavior allows for avoiding the channel resource waste if individual BS buffer gets empty and therefore it results in a more efficient capacity utilization.

4 Overall Delay Analysis

In this section we conduct the evaluation of the overall data packet delay in the considered wireless broadband network. This delay includes both the reservation and the scheduling parts. We denote the durations of the downlink (DL) and the uplink (UL) sub-frames by T_{DL} and T_{UL}, respectively. Then for these durations it holds:

$$T_{UL} = T_{RI} + T_{UD}, \tag{1}$$

where T_{RI} is the duration of the reservation interval (RI) and T_{UD} is the maximum allowable duration of the UL sub-frame for sending the uplink data (UD).

Remember that according to the system model each frame comprises K contention-based transmission opportunities, which yields $T_{RI} = K\alpha$, where

α is the bandwidth request duration. Then we can rewrite the expression for T_{UD} as:

$$T_{UD} = T_{UL} - K\alpha. \tag{2}$$

Otherwise, accounting for the fact that each SS transmits at most one data packet per uplink sub-frame, we establish:

$$T_{UD} = M\tau. \tag{3}$$

Combining (1), (2) and (3) as well as assuming that the channel propagation time is negligible we obtain the following expression for the frame duration:

$$T_{frame} = T_{DL} + K\alpha + M\tau. \tag{4}$$

Let ρ denote the load at SS i. As an SS transmits at most one packet per frame, we obtain:

$$\rho = \lambda T_{frame} = \frac{\Lambda T_{frame}}{M}. \tag{5}$$

Clearly, the considered system is stable when $\rho < 1$ or $\Lambda < \frac{M}{T_{frame}}$, that is the number of arriving packets does not on average exceed the number of departing packets.

Consider the overall packet delay D_i for the i-th SS, which is a continuous random variable. This delay arises due to both queueing in the outgoing SS buffer during the reservation delay and queueing in the BS buffer during the scheduling delay. The overall packet delay is thus defined as the time interval from the moment the packet arrives into the system to the moment when its successful uplink transmission ends. Figure 3 illustrates the following components of the overall tagged packet delay:

$$D_i = D_i^r + \alpha + D_i^s + \tau, \tag{6}$$

where the components are defined as follows.

- D_i^r – reservation delay from the moment the packet arrives into the outgoing buffer of SS i to the start of the successful transmission of the corresponding bandwidth request in the reservation interval.
- α – time of the successful bandwidth request transmission, which equals the duration of the transmission opportunity.
- D_i^s – scheduling delay from the end of the successful bandwidth request transmission of the i-th SS to the start of the corresponding data packet transmission in the uplink sub-frame.
- τ – data packet transmission time.

The main assumption of the analysis is that the probability of the successful bandwidth request transmission in a reservation interval is constant. Let p_r denote this probability as it is independent of the SS index. Accounting for the fact that each SS has an individual BS buffer and an own, separate data packet transmission period in the uplink sub-frame, we conclude that the statistical behavior of an SS is independent of that one for the other SSs. As such, to establish the overall packet delay of the tagged SS, it is enough to model its behavior separately from the rest of the system.

According to this we consider the system shown in Figure 3 from the point of view of the tagged SS i. For the sake of simplicity in the following description we omit the index i. We construct an embedded Markov chain [20] at the sequence of begin times of the consecutive reservation intervals. The state of the chain consists of the number of packets in the SS and BS buffers. More precisely, we assume that there are three buffers for the data packets (see Figure 4). The first buffer is the one at the tagged SS where the packet is queued during the reservation delay. After that the packet is immediately transferred to the virtual buffer at the beginning of the corresponding reservation interval. The virtual buffer accounts for the fact that a packet cannot be transmitted in the current frame, that is, experiences the delay of at least one frame. After this additional delay the packet enters the individual BS buffer of the tagged SS. There the packet is queued until the end of the scheduling delay. Finally, the packet is transmitted. Note that in this equivalent queueing system we implicitly assume that the transitions between the buffers and leaving the last buffer happen at the embedded epochs, i.e. somewhat earlier comparing to e.g. the BS processing at the end of the contention-based transmission opportunities.

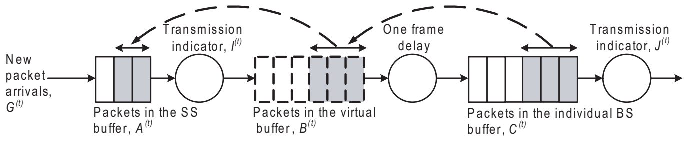

Transmission indicator, $I^{(t)}$
One frame delay
Transmission indicator, $J^{(t)}$
New packet arrivals, $G^{(t)}$
Packets in the SS buffer, $A^{(t)}$
Packets in the virtual buffer, $B^{(t)}$
Packets in the individual BS buffer, $C^{(t)}$

Fig. 4. Equivalent queueing system description

Let $\{A^{(t)}\}$, $\{B^{(t)}\}$ and $\{C^{(t)}\}$ denote the number of packets in the first buffer, in the virtual buffer and in the individual BS buffer at the embedded epoch in the t-th frame, respectively. The dynamics of the number of packets in the first SS buffer at the consecutive embedded time epochs in frame t and $t+1$ can be expressed by the following expression:

$$A^{(t+1)} = (A^{(t)} + G^{(t)})(1 - I^{(t)}), \tag{7}$$

where $G^{(t)}$ is the number of newly arriving packets, which enter the SS buffer during the interval between the t-th and $(t+1)$-th embedded epochs and $I^{(t)}$ is the discrete indicator function showing if the corresponding bandwidth request is transmitted successfully in the reservation interval of the t-th frame:

$$I^{(t)} = \begin{cases} 1 & \text{with probability } p_r, \\ 0 & \text{with probability } 1 - p_r. \end{cases} \tag{8}$$

The dynamics of the number of packets in the virtual buffer $\{B^{(t)}\}$ could be described as follows:

$$B^{(t+1)} = (A^{(t)} + G^{(t)})I^{(t)}. \tag{9}$$

Finally, the evolution of the number of packets in the individual BS buffer $\{C^{(t)}\}$ at the embedded time moments could be written as:

$$C^{(t+1)} = C^{(t)} - J^{(t)} + B^{(t)}, \tag{10}$$

where $J^{(t)}$ is the discrete indicator function showing if the packet is transmitted successfully in the uplink sub-frame of the t-th frame:

$$J^{(t)} = \begin{cases} 1, & \text{if } C^{(t)} > 0, \\ 0, & \text{if } C^{(t)} = 0. \end{cases} \tag{11}$$

We are interested in obtaining the steady-state expression for the mean number of packets in all the considered buffers. Let $E[A^j]$, $E[B^j]$ and $E[C^j]$ stand for the limiting j-th moments of the $\{A^{(t)}\}$, $\{B^{(t)}\}$ and $\{C^{(t)}\}$ random variables, for $j = 1, 2, \ldots$, i.e. $E[A^j] = \lim\limits_{t\to\infty} E[(A^{(t)})^j]$, $E[B^j] = \lim\limits_{t\to\infty} E[(B^{(t)})^j]$ and $E[C^j] = \lim\limits_{t\to\infty} E[(C^{(t)})^j]$. The ergodicity of the considered Markov chain ensures the existence of the limiting distributions of $\{A^{(t)}\}$, $\{B^{(t)}\}$ and $\{C^{(t)}\}$.

To determine the steady-state mean number of packets in all three buffers we derive relations for the above first two moments from the equations (7), (9) and (10). We average both parts of these expressions. Also we raise them to the second power and then take the mathematical expectation of both parts. Utilizing the mutual independence of $A^{(t)}$, $G^{(t)}$ and $I^{(t)}$ as well as the mutual independence of $B^{(t)}$ and $C^{(t)}$ the required mean quantities can be established as:

$$\begin{aligned} E[A] &= \rho\frac{1-p_r}{p_r}, \\ E[B] &= \rho, \\ E[C] &= \frac{2\rho - \rho^2\left(3 - \frac{2}{p_r}\right) - 2E[AJ]}{2(1-\rho)}. \end{aligned} \tag{12}$$

We note that the expression of $E[C]$ explicitly incorporates $E[AJ]$, since the $A^{(t)}$ and $J^{(t)}$ are dependent random variables. The exact calculation of this term is not easy and therefore here we replace it with zero. As such, we derive the upper bound on $E[C]$. The total number of packets at the embedded epoch is $E[A] + E[B] + E[C]$. Note that in the three buffers model every transition between the buffers occurs at an embedded epoch and only packet arrivals happen between these epochs. If the state of the Markov chain at an embedded epoch is given, the stochastic evolution of the the number of packets in the system repeats itself in the intervals between the consecutive embedded epochs. Thus, to get the number of packets in the system at an arbitrary epoch (Q^{rs}) it is enough to consider it at an arbitrary epoch in the intervals between two consecutive embedded epochs with the length of T_{frame}. Hence $E[Q^{rs}]$ plus the number of arriving packets during the forward recurrence time of such an interval $(\lambda\frac{T_{frame}}{2})$ is exactly the total number of packets at the embedded epochs. This yields:

$$E[Q^{rs}] = E[A] + E[B] + E[C] - \frac{\rho}{2}. \tag{13}$$

Applying (13) in the Little's formula [21] we can obtain the upper bound on the sum of the reservation and the scheduling packet delays. However the virtual buffer accounts only for a part of the delay from scheduling a grant of a packet to the start of the uplink transmission of that packet. The rest of this delay for an i-packet is given as $\alpha K + (i-1)\tau$. Averaging over every possible $i = 1, \ldots, M$ yields $\alpha K + \tau\frac{M-1}{2}$. Furthermore as the embedded epoch happens at least by α time earlier as the end of the contention-based transmission opportunities the above delay part is upper bounded by $\alpha(K-1) + (i-1)\tau$. Taking this term into account in the scheduling packet delay, expressing the sum of the reservation and the scheduling packet delays by applying Little's formula and using (6) results in the upper bound on overall packet delay in the considered wireless broadband network as:

$$E[D] \leq \left(\frac{3}{2} + \frac{1-p_r}{p_r}\right)T_{frame} + \frac{\rho T_{frame}\,(2-p_r)}{2p_r(1-\rho)} + \alpha K + \tau\frac{M+1}{2}. \tag{14}$$

The probability of the successful bandwidth request transmission in a reservation interval p_r can be determined by means of a second Markov chain model, which uses the quantity p_t, which is defined as the probability of a transmission attempt of an SS.

Firstly, we briefly summarize the determination of the probability p_t, which is presented in [22]. At the reservation stage IEEE 802.16 users follow the rules of the BEB algorithm used for the collision resolution. The BEB algorithm operation is thoroughly investigated in [22]. According to [5] and [6] the consideration of the entire system could be reduced to the consideration of the tagged SS only. For a contention-based transmission opportunity the conditional collision probability, conditioning on the fact that the SS attempts the transmission (p_c) is introduced as:

$$p_c = 1 - (1-p_t)^{M-1}. \tag{15}$$

This probability may be established by:

$$p_t = \frac{2(1-2p_c)}{(1-2p_c)(W_0+K) + p_c W_0(1-(2p_c)^m)}, \tag{16}$$

where W_0 and m are the parameters of the BEB algorithm and they are termed as initial contention window and maximum backoff stage, respectively. Hence, the probabilities p_t and p_c can be determined by solving the system of two non-linear equations (15) and (16).

As stated above having the probability p_t a second Markov chain model can be set up for determination of p_r. This can be described analogously to its description in our previous work [14]. From the point of view of the bandwidth requesting each SS may reside in an active or an inactive state. Active SS participates in the contention process, i.e. it has at least one pending data packet, for which a successful bandwidth request has not yet been issued. Inactive SS does not initiate the reservation process as it has no packets, of which the BS has not yet been successfully informed. We introduce a Markov chain embedded at the sequence of the ends of the contention-based transmission opportunities. The state of this Markov chain $\{N^{(u)}\}$, for $u = 1, \ldots$, composes of the number of active SSs. In each frame the first packet arrives to an inactive SS with the probability $y = 1 - e^{-\lambda T_{frame}}$. After the first packet arrival the SS enters the active state, issues a new bandwidth request and starts the contention process, for which all the subsequent arrivals are irrelevant. According to these the transition probabilities among the $M + 1$ states of the chain can be written as:

$$p_{i,j} = \Pr\{N^{(t+1)} = j | N^{(t)} = i\} = \tag{17}$$
$$= \begin{cases} 0, & \text{if } j \leq i-2, \\ ip_t(1-p_t)^{i-1}(1-y)^{M-i+1}, & \text{if } j = i-1, \\ ip_t(1-p_t)^{i-1}(M-i+1)y(1-y)^{M-i}+ & \\ +(1-ip_t(1-p_t)^{i-1})(1-y)^{M-i}, & \text{if } j = i, \\ ip_t(1-p_t)^{i-1}\begin{pmatrix} M-i+1 \\ j-i+1 \end{pmatrix} y^{j-i+1}(1-y)^{M-j}+ & \\ +(1-ip_t(1-p_t)^{i-1})\begin{pmatrix} M-i \\ j-i \end{pmatrix} y^{j-i}(1-y)^{M-j}, & \text{if } j \geq i+1. \end{cases}$$

It may be shown that the considered Markov chain is finite and irreducible for $p_t, y > 0$ [23]. Therefore, its stationary probability distribution exists, which may be obtained by solving a linear system of $M + 1$ equations:

$$\begin{cases} P_j = \sum\limits_{i=0}^{M} P_i p_{i,j} \quad \text{for } j = 0, 1, \ldots, M, \\ \sum\limits_{i=0}^{M} P_i = 1. \end{cases} \tag{18}$$

We determine the joint probability at the end of a contention-based transmission opportunity that the number of active SSs is n $(n = 1, \ldots, M)$ and the tagged SS is among them and the tagged SS has successful bandwidth request transmission. This probability is denoted by $s(n)$. Due to the symmetry of the model the probability that the tagged SS is among the i active SSs is given by $\frac{\binom{M-1}{n-1}}{\binom{M}{n}} = \frac{n}{M}$. Thus the $s(n)$ can be expressed as:

$$s(n) = \frac{n}{M} p_t (1-p_t)^{n-1}. \tag{19}$$

Let p_s denote of the successful bandwidth request transmission of the tagged SS at the end of a contention-based transmission opportunity. p_s can be calculated with the help of the stationary distribution $\{P_n\}_{n=\overline{0,M}}$ of the Markov chain as:

$$p_s = \sum_{n=0}^{M} s(n) P_n. \tag{20}$$

A bandwidth request transmission in a reservation interval can be successful in any of the K provided contention-based transmission opportunity. As these events exclude each other, p_r can be given by:

$$p_r = K p_s. \tag{21}$$

5 Numerical Results and Conclusion

In order to verify the adequacy of the model assumptions made during the performance analysis we developed a simplified IEEE 802.16 MAC simulator. It accounts for the restrictions of the system model and was previously used in [7], [16], [14] and [17]. According to [24] we set the typical simulation parameters and summarize them in Table 1.

Table 1. Typical simulation parameters

IEEE 802.16 parameter	Value
DL:UL proportion	60:40
PHY type	OFDM
Frame duration (T_{frame})	5 ms
Channel bandwidth	7 MHz
Contention-based transmission opportunity duration (α)	170 μs
Data packet length	4096 bits

The result of the verification for this typical parameter set is demonstrated in Figure 5, where curves show analytical results and symbols are obtained with simulation. The accuracy of the model depends on the overall arrival rate and some system parameters, such as p_r. Although we do not include results for different values and system parameters in this paper, we have shown through extensive simulations that the derived model is reasonably accurate for the realistic protocol settings. Therefore, it is a useful tool for the evaluation of the overall packet delay, as well as for fine-tuning the wireless system to control it.

In this paper we proposed an analytical model to estimate the overall data packet delay in IEEE 802.16 network. The model accounts for the delay at both the reservation and the scheduling stages. Several assumptions of the presented model can be relaxed and hence the analysis can be generalized in these

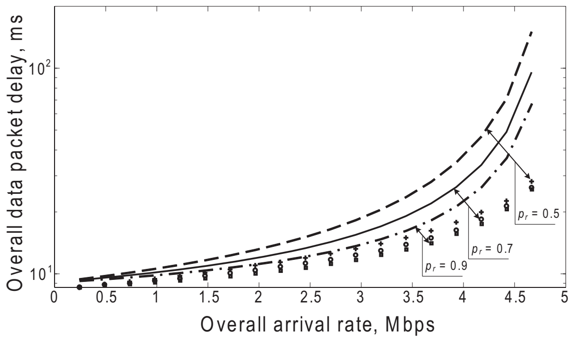


Fig. 5. Overall data packet delay in logarithmic scale for $M = 6$ and $K = 1$

directions. According to IEEE 802.16 protocol each SS may potentially establish **multiple connections** with the BS. The developed system model may be generalized for this case by considering connections instead of SSs. The assumption about the noise-free uplink and downlink channels is also non-realistic. In practice the transmissions are always corrupted by the adverse wireless channel effects. The analytical model enables the extension for the case of the **noisy channel**. Finally, the proposed analytical model could be modified to account also for the **unicast polling** of the SSs incorporating the models of [16] and [17].

Acknowledgments

This work is supported by the Russian Foundation for Basic Research (projects # 10-08-01071-a and # 08-08-00403-a), as well as by the Branch of Nano- and Information Technologies of Russian Academy of Sciences (project 2.3)

References

1. *IEEE 802.16-2009. IEEE Standard for Local and metropolitan area networks*, May 2009.
2. I. Rubin, “Access-control disciplines for multi-access communication channels: reservation and TDMA schemes,” *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 516–536, 1979.
3. D. Aldous, “Ultimate instability of exponential back-off protocol for acknowledgment based transmission control of random access communication channels,” *IEEE Transactions on Information Theory*, vol. 33, no. 2, pp. 219–223, 1987.
4. J. Goodman, A. Greenberg, N. Madras, and P. March, “Stability of binary exponential backoff,” *Journal of the ACM*, vol. 35, no. 3, pp. 579–602, 1988.
5. G. Bianchi, “Performance analysis of the IEEE 802.11 distributed coordination function,” *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.

6. N. Song, B. Kwak, and L. Miller, “On the stability of exponential backoff,” *Journal of Research of the NIST*, vol. 108, no. 4, pp. 289–297, 2003.
7. S. Andreev, A. Turlikov, and A. Vinel, “Contention-based polling efficiency in broadband wireless networks,” in *Proc. of the 15th International Conference on Analytical and Stochastic Modeling Techniques and Applications*, pp. 295–309, 2008.
8. L. Lin, W. Jia, and W. Lu, “Performance analysis of IEEE 802.16 multicast and broadcast polling based bandwidth request,” in *Proc. of the IEEE Wireless Communications and Networking Conference*, pp. 1854–1859, 2007.
9. O. Alanen, “Multicast polling and efficient VoIP connections in IEEE 802.16 networks,” in *Proc. of the 10th ACM Symposium on Modeling, Analysis, and Simulation of Wireless and Mobile Systems*, pp. 289–295, 2007.
10. A. Kobliakov, A. Turlikov, and A. Vinel, “Distributed queue random multiple access algorithm for centralized data networks,” in *Proc. of the 10th IEEE International Symposium on Consumer Electronics*, pp. 290–295, 2006.
11. D. Cho, J. Song, M. Kim, and K. Han, “Performance analysis of the IEEE 802.16 wireless metropolitan network,” in *Proc. of the 1st International Conference on Distributed Frameworks for Multimedia Applications*, pp. 130–136, 2005.
12. L. Moraes and P. Maciel, “Analysis and evaluation of a new MAC protocol for broadband wireless access,” in *Proc. of the International Conference on Wireless Networks, Communications and Mobile Computing*, vol. 1, pp. 107–112, 2005.
13. R. Iyengar, P. Iyer, and B. Sikdar, “Delay analysis of 802.16 based last mile wireless networks,” in *Proc. of the 48th IEEE Global Telecommunications Conference*, vol. 5, pp. 3123–3127, 2005.
14. S. Andreev, Z. Saffer, A. Turlikov, and A. Vinel, “Overall delay in IEEE 802.16 with contention-based random access,” in *Proc. of the 16th International Conference on Analytical and Stochastic Modeling Techniques and Applications*, p. 89102, 2009.
15. *IEEE 802.16e-2005. Amendment to IEEE Standard for Local and Metropolitan Area Networks*, February 2006.
16. Z. Saffer and S. Andreev, “Delay analysis of IEEE 802.16 wireless metropolitan area network,” in *Proc. of the 15th International Conference on Telecommunications*, pp. 1–5, 2008.
17. S. Andreev, Z. Saffer, and A. Anisimov, “Overall delay analysis of IEEE 802.16 network,” in *Proc. of the IEEE International Conference on Communications*, pp. 1–6, 2009.
18. D. Bertsekas and R. Gallager, *Data Networks.* Prentice Hall, 1992.
19. R. Rom and M. Sidi, *Multiple Access Protocols: Performance and Analysis.* Springer-Verlag, 1990.
20. L. Kleinrock, *Queueing Systems: Volume II - Computer Applications.* New York, 1976.
21. L. Kleinrock, *Queueing Systems: Volume I – Theory.* New York, 1975.
22. S. Andreev and A. Turlikov, “Binary exponential backoff algorithm analysis in the lossy system with frames,” in *Proc. of the 12th International Symposium on Problems of Redundancy in Information and Control Systems*, p. 201210, 2009.
23. L. Kleinrock and S. Lam, “Packet-switching in a multi-access broadcast channel: performance evaluation,” *IEEE Transactions on Communications*, vol. 23, no. 4, pp. 410–423, 1975.
24. D. Sivchenko, N. Bayer, B. Xu, V. Rakocevic, and J. Habermann, “Internet traffic performance in IEEE 802.16 networks,” in *Proc. of the 12th European Wireless Conference*, pp. 1–5, 2006.

Publication 3

S. Andreev, Z. Saffer, and A. Turlikov, “Delay analysis of wireless broadband networks with non real-time traffic,” in *Proc. of the 4th International Workshop on Multiple Access Communications (MACOM)*, pp. 206–217, 2011.

Delay Analysis of Wireless Broadband Networks with Non Real-Time Traffic

Sergey Andreev[1], Zsolt Saffer[2], and
Andrey Turlikov[3]

Tampere University of Technology[1] (TUT), FINLAND
sergey.andreev@tut.fi
Budapest University of Technology and Economics[2] (BUTE), HUNGARY
safferzs@hit.bme.hu
State University of Aerospace Instrumentation[3] (SUAI), RUSSIA
turlikov@vu.spb.ru

Abstract. In this paper, we present the analysis of the mean overall packet delay of non real-time traffic in IEEE 802.16-based wireless broadband networks. We consider the case of contention-based bandwidth reservation. The system model accounts for both bandwidth reservation and packet transmission delay components of the overall delay. The queueing analysis is based on the description of the joint content of the outgoing subscriber station buffer and the base station grant buffer. This is achieved by means of a properly chosen bivariate embedded Markov chain. The mean overall packet delay is computed from its equilibrium solution. The analytical approach is verified by means of simulation. The corresponding analytical and simulation results show excellent agreement with each other.

Keywords: IEEE 802.16, queueing system, Markov chain, contention-based bandwidth reservation.

1 Introduction and Background

In wireless broadband networks, the users are distributed across a large geographic area and communicate via a base station (BS). As such, the BS is the coordinator of the network activity, which controls user communication in its vicinity. Recently, the proliferation of IEEE 802.16-based [1] broadband networks is observed. This is due to their relatively low cost, wide coverage and MAC mechanisms supporting a variety of quality of service (QoS) requirements.

The performance evaluation of IEEE 802.16 QoS features with bandwidth reservation is addressed by numerous research papers (see e.g. [2], [3], and [4]). The overall operation of the considered wireless broadband network is shown in Figure 1, in which both downlink (DL) and uplink (UL) transmissions are demonstrated. In the DL and the UL data packets are sent from the BS to its subscriber stations (SSs) and in the opposite direction, respectively. Initially, a SS issues a bandwidth request (BW-Req) in the UL (1UL), which is received by the BS (2UL). After processing these requests, the BS forms a transmission schedule and then forwards it to the SSs in the DL (1DL). Each SS receives the schedule (2DL) and transmits own data packets accordingly in the dedicated time-frequency slots (3UL, 4UL). If necessary, the BS may also transmit data packets to SSs in the DL (3DL, 4DL). As such, the overall system operation consists of bandwidth reservation and packet transmission functionalities, therefore the overall system model should account for both bandwidth reservation and packet transmission stages.

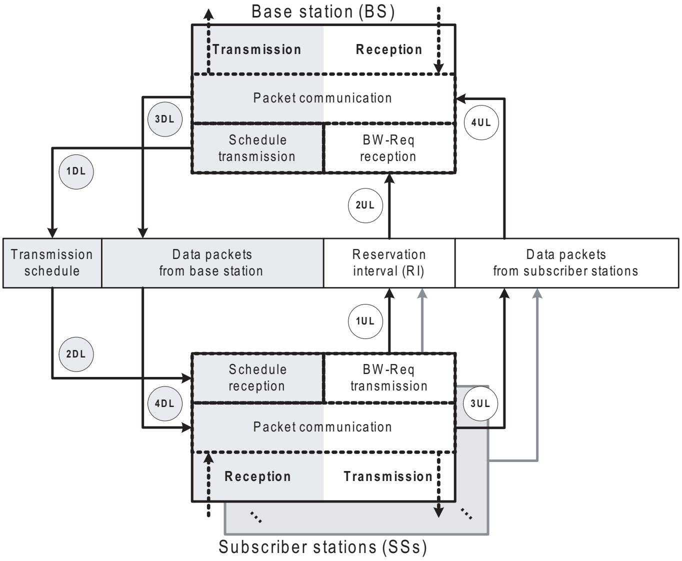


Fig. 1. System overall description

There has been little effort taken to address both aforementioned stages of the wireless broadband network functionality. This is due to the complexity of the overall system operation. The performance of IEEE 802.16 network was studied either by simulation [5], [6] or particular special cases were addressed analytically [7]. Moreover, in the majority of existing research papers the reservation and the transmission are considered separately. However, the operation of the real-world network (see Figure 1) includes both functionalities, which should be taken into account. In our previous work [8], we have studied the overall delay by addressing its both components and constructed a simple analytical upper bound on its mean value. In this paper, we continue our work and calculate the exact value of the mean overall delay.

2 System Model

In this section, we briefly outline the model of IEEE 802.16-based network, which is used to evaluate the delay at both reservation and scheduling stages. For more details, see our previous paper [8].

We consider the system that comprises a BS and N SSs, in which we focus only on the uplink transmissions. The delay analysis is conducted for non real-time (nrtPS) QoS profile. Only contention-based polling schemes are considered. Particularly, we concentrate on the broadcast polling.

The system operation time is divided into frames and T_f denotes the frame duration. The duration of a packet transmission is τ. At each SS, the packet arrival process is Poisson. The arrival rate is λ_i at SS i. Thus the overall arrival rate is $\lambda = \lambda_i N$. The duration of each contention-based transmission opportunity is ν. Moreover, the reservation interval (RI) of each frame comprises exactly K contention-based transmission opportunities.

A BW-Req is issued by the i-th SS whenever at least one new data packet arrives, of which the BS should be notified. The request contains the information about all the newly arrived packets since the last request sending. If a packet arrives to an empty outgoing buffer of SS i during the RI the SS must wait with sending the BW-Req for this packet until the next RI. We define $p_i^{(b)}$ as the probability of the successful BW-Req transmission at SS i, given that this SS takes part in the contention process (i.e. there is at least one new i-packet belonging to SS i in its outgoing buffer).

Additionally, below we introduce a set of assumptions to shape the system model.

Assumption 1. The probability of the successful BW-Req transmission at each SS in the RI of a frame, $p_i^{(b)}$, is assumed to be constant (see [8]).

Assumption 2. The BS maintains an individual grant buffer of infinite capacity for each SS.

Assumption 3. The grants in the individual BS grant buffer of each SS are processed in first-come-first-served order.

Assumption 4. Each SS can transmit exactly one packet in each UL sub-frame.

Assumption 5. For each SS, the feedback on the success/failure of its own BW-Req transmission, which is necessary for the collision resolution algorithm operation, is available.

The system is stable when for every SS on average the number of arriving packets does not exceed the number of departing packets, i.e.

$$\lambda_i < \frac{1}{T_f}, \quad i = 1, \ldots, N. \tag{1}$$

In general, our approach enables asymmetric traffic arrival patterns and different $p_i^{(b)}$ for the individual SSs. However, determining this probability for asymmetric system is more complicated. For the sake of simplicity, we consider only the symmetric model, i.e. at each SS the arrival rate of the uplink traffic is the same, λ_i. Similarly, $p_i^{(b)}$ is also the same for every SS.

3 Queueing Analysis

In this section, we detail a queueing model for the reservation and scheduling parts of the system. The statistical behavior of a SS is independent of that one for the other SSs, since each SS has an individual BS buffer and a separate data packet transmission period in the uplink sub-frame. Consequently, it is enough to model the behavior of the tagged SS separately from the rest of the system. Accordingly, we consider the system from the point of view of the tagged SS i.

We study the behavior of the tagged SS i at the embedded epochs, which are the end epochs of the first contention-based transmission opportunities in the RIs of the frames. In the queueing model, we assume that the BW-Req transmissions happen at the embedded epochs, i.e. in several cases somewhat earlier than in the real system, in which they happen at the end of the contention-based transmission opportunities. From the point of view of the queueing model, the "service start" of an i-packet is associated with an embedded epoch, in the frame preceding the one, in which that i-packet is transmitted. Such a "service start" of an i-packet is modeled by the event that the corresponding i-grant leaves the i-grant buffer of the BS, i.e. that i-grant is scheduled. The interval between the end epochs of the first contention-based transmission opportunities in the RIs in two consecutive frames is called a *cycle*.

By definition, the time instant of the successful BW-Req transmission from SS i is the *i-reservation event*. Similarly, by definition the instants of scheduling the BS grants in the BS grant buffer of SS i are the *i-scheduling events*. The positioning of the embedded epochs relatively to the corresponding *i-reservation* and *i-scheduling events* are shown in Figure 2.

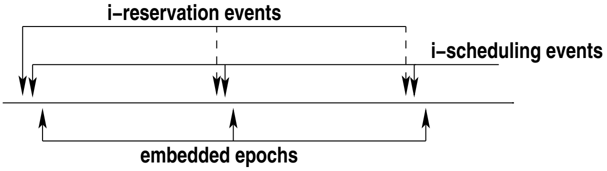


Fig. 2. Positions of the embedded epochs

The main assumption of the queueing analysis (see Assumption 1) is that the probability of the successful BW-Req transmission at SS i, given that this SS takes part in the contention process, $p_i^{(b)}$, is constant.

3.1 The joint content of the outgoing and BS grant buffers of SS i at the embedded epochs

Let $q_i^{(r)}(\ell)$ be the number of i-packets in the outgoing buffer of SS i at the end of the first contention-based transmission opportunity in the RI in the ℓ-th frame for $\ell > 0$. Similarly, let $q_i^{(s)}(\ell)$ be the number of i-grants in the BS grant buffer of SS i at the end of the first contention-based transmission opportunity in the RI in the ℓ-th frame for $\ell > 0$. The sequence $\{(q_i^{(r)}(\ell), q_i^{(s)}(\ell)), \ell > 0\}$ is a bivariate homogeneous embedded Markov chain on the state space $(\{0, 1, \ldots\}, \{0, 1, \ldots\})$. We say that the chain is in state (j, k) when $q_i^{(r)}(\ell) = j$ and $q_i^{(s)}(\ell) = k$. Let $p_i(j, k, n, m)$ denote the probability of transition from state j, k to state n, m in this Markov chain, i.e.

$$\begin{aligned} p_i(j,k,n,m) = {} & P\{q_i^{(r)}(\ell+1) = n, q_i^{(s)}(\ell+1) = m \mid q_i^{(r)}(\ell) = j, q_i^{(s)}(\ell) = k\}, \\ & \ell \geq 1, \quad j, k, n, m \geq 0. \end{aligned} \tag{2}$$

Let us consider the transitions from state (j, k) to state (n, m) in the above defined Markov chain. The transition from state $(0, 0)$ to state $(0, 0)$ happens either if there are no i-packet arrivals during the actual cycle or there is exactly one i-packet arrival during a cycle and the BW-Req transmission at the end of that cycle is successful, i.e. the newly generated i-grant is immediately scheduled to be sent. Thus the probability of this transition is given as

$$p_i(0,0,0,0) = e^{-\lambda_i T_f} + p_i^{(b)} \lambda_i T_f e^{-\lambda_i T_f}. \tag{3}$$

The transition from state $(0, 0)$ to $(n, 0)$ for $n \geq 1$ happens if there are n i-packet arrivals during a cycle and the BW-Req transmission at the end of that cycle is not successful. This leads to

$$p_i(0,0,n,0) = (1 - p_i^{(b)}) \frac{(\lambda_i T_f)^n}{n!} e^{-\lambda_i T_f}, \quad n \geq 1. \tag{4}$$

The transition from state $(0, k)$ to $(0, k-1)$ for $k \geq 1$ happens if there are no i-packet arrivals during the actual cycle. Thus we have

$$p_i(0,k,0,k-1) = e^{-\lambda_i T_f}, \quad k \geq 1. \tag{5}$$

The transition from state $(0, k)$ to $(0, k)$ for $k \geq 1$ happens if there is exactly one i-packet arrival during a cycle and the BW-Req transmission at the end of that cycle is successful, i.e. the newly generated i-grant is immediately scheduled to be sent. This results in

$$p_i(0,k,0,k) = p_i^{(b)} \lambda_i T_f e^{-\lambda_i T_f}, \quad k \geq 1. \tag{6}$$

The transition from state $(0, k)$ to $(0, m)$ for $k \geq 0$ and $m > k$ happens if there are exactly $m - k + 1$ i-packet arrivals during a cycle and the BW-Req transmission at the end of that cycle is successful. The corresponding transition probability is given as

$$p_i(0, k, 0, m) = p_i^{(b)} \frac{(\lambda_i T_f)^{m-k+1}}{(m-k+1)!} e^{-\lambda_i T_f}, \quad k \geq 0, \ m > k. \tag{7}$$

The transition from state $(0, k)$ to $(n, k-1)$ for $n \geq 1$ and $k \geq 1$ happens if there are exactly n i-packet arrivals during a cycle and the BW-Req transmission at the end of that cycle is not successful. This leads to

$$p_i(0, k, n, k-1) = (1 - p_i^{(b)}) \frac{(\lambda_i T_f)^n}{n!} e^{-\lambda_i T_f}, \quad n, k \geq 1. \tag{8}$$

The transition from state $(1, 0)$ to $(0, 0)$ happens if there are no i-packet arrivals during the actual cycle and the BW-Req transmission at the end of that cycle is successful. Thus it results in

$$p_i(1, 0, 0, 0) = p_i^{(b)} e^{-\lambda_i T_f}. \tag{9}$$

The transition from state $(j, 0)$ to $(n, 0)$ for $j \geq 1$ and $n \geq j$ happens if there are exactly $n - j$ i-packet arrivals during a cycle and the BW-Req transmission at the end of that cycle is not successful. The corresponding transition probability is given as

$$p_i(j, 0, n, 0) = (1 - p_i^{(b)}) \frac{(\lambda_i T_f)^{n-j}}{(n-j)!} e^{-\lambda_i T_f}, \quad j \geq 1, \ n \geq j. \tag{10}$$

The transition from state $(1, k)$ to $(0, m)$ for $k = 0$, $m \geq 1$ or $k \geq 1$, $m \geq k$ happens if there are exactly $m - k$ i-packet arrivals during a cycle and the BW-Req transmission at the end of that cycle is successful. This yields

$$p_i(1, k, 0, m) = p_i^{(b)} \frac{(\lambda_i T_f)^{m-k}}{(m-k)!} e^{-\lambda_i T_f}, \ \ k = 0, \ m \geq 1, \ \text{or } k \geq 1, \ m \geq k. \tag{11}$$

The transition from state (j, k) to $(0, m)$ for $j \geq 2$, $k \geq 0$ and $m \geq k + j - 1$ happens if there are exactly $m - k - j + 1$ i-packet arrivals during a cycle and the BW-Req transmission at the end of that cycle is successful. This leads to

$$p_i(j, k, 0, m) = p_i^{(b)} \frac{(\lambda_i T_f)^{m-k-j+1}}{(m-k-j+1)!} e^{-\lambda_i T_f}, \ \ j \geq 2, \ k \geq 0, \ m \geq k + j - 1. \tag{12}$$

Finally the transition from state (j, k) to $(n, k-1)$ for $j \geq 1$, $k \geq 1$ and $n \geq j$ happens if there are exactly $n - j$ i-packet arrivals during a cycle and the BW-Req transmission at the end of that cycle is not successful. The corresponding transition probability is given as

$$p_i(j, k, n, k-1) = (1 - p_i^{(b)}) \frac{(\lambda_i T_f)^{n-j}}{(n-j)!} e^{-\lambda_i T_f}, \;\; j \geq 1, \;\; k \geq 1, \;\; n \geq j. \qquad (13)$$

Let $p_i^{(e)}(j, k)$ denote the equilibrium joint probability that the above Markov chain is in state j, k. To keep the computation of the joint probabilities tractable we apply an upper limit X_i both on the number of i-packets in the outgoing buffer of SS i and on the number of i-grants in the BS grant buffer of SS i, i.e. $j, k \leq X_i$. This results in finite number of equilibrium joint probabilities and transition probabilities and hence finite number of equilibrium equations. The proper value of X_i depends on the required precision and can be determined on iterative manner until the difference of consecutive values of probabilities $p_i^{(e)}(j, k)$, for every $j, k \leq X_i$, becomes less than the specified error. In the computation, the probabilities $p_i^{(e)}(j, k)$ for $j > X_i$ or $k > X_i$ are set 0, since they can be neglected.

Let $\mathbf{e}_j^{X_i+1} = (0, \ldots, 0, 1, 0, \ldots, 0)$ denote the $1 \times (X_i + 1)$ vector with 1 at the j-th position. Additionally, let $\otimes$ stand for the Kronecker product. We define the $1 \times (X_i + 1)^2$ vector $\boldsymbol{\theta}_i$ representing the equilibrium joint probabilities of the above Markov chain as

$$\boldsymbol{\theta}_i = \sum_{j=0}^{X_i} \sum_{k=0}^{X_i} \; p_i^{(e)}(j, k) \; \mathbf{e}_j^{X_i+1} \otimes \mathbf{e}_k^{X_i+1}. \qquad (14)$$

We also define the $(X_i+1)^2 \times (X_i+1)^2$ matrix $\boldsymbol{\Pi}_i$ representing the transition probabilities of the embedded Markov chain as

$$\boldsymbol{\Pi}_i = \sum_{j=0}^{X_i} \sum_{k=0}^{X_i} \; \sum_{n=0}^{X_i} \sum_{m=0}^{X_i} \; p_i(j, k, n, m) \; \left(\mathbf{e}_j^{X_i+1} \otimes \mathbf{e}_k^{X_i+1}\right)^T \; \left(\mathbf{e}_n^{X_i+1} \otimes \mathbf{e}_m^{X_i+1}\right) \qquad (15)$$

In the matrix $\boldsymbol{\Pi}_i$ the values of j, k and the values of n, m specify the row and the column indices of the corresponding transition probability $p_i(j, k, n, m)$.

The equilibrium joint probabilities of the embedded Markov chain can be uniquely determined from the following system of linear equations

$$\boldsymbol{\theta}_i \boldsymbol{\Pi}_i = \boldsymbol{\theta}_i, \quad \boldsymbol{\theta}_i \mathbf{e}^{(X_i+1)^2} = \sum_{j=0}^{X_i} \sum_{k=0}^{X_i} p_i^{(e)}(j, k) = 1, \qquad (16)$$

where $\mathbf{e}^{(X_i+1)^2}$ denotes the $(X_i + 1)^2 \times 1$ column vector having all elements equal to one.

The mean number of packets in the outgoing buffer of SS i at the end of the first contention-based transmission opportunity in the RI, $E[q_i^{(r)}]$, can be computed from the equilibrium joint distribution as

$$E[q_i^{(r)}] = \sum_{j=0}^{X_i} \sum_{k=0}^{X_i} j \; p_i^{(e)}(j,k). \tag{17}$$

Similarly, the mean number of i-grants in the BS grant buffer of SS i at the end of the first contention-based transmission opportunity in the RI, $E[q_i^{(s)}]$, can be computed also from the equilibrium joint distribution as

$$E[q_i^{(s)}] = \sum_{j=0}^{X_i} \sum_{k=0}^{X_i} k \; p_i^{(e)}(j,k). \tag{18}$$

3.2 The mean of the joint content of the outgoing and BS grant buffers of SS i at an arbitrary moment

Let q_i stand for the joint content of the outgoing and BS grant buffers of SS i at an arbitrary moment, i.e. the sum of the number of i-packets in the outgoing buffer of SS i and the number of i-grants in the BS grant buffer of SS i at an arbitrary moment.

The number of i-grants can change only just before the embedded observation epochs. This implies that the number of i-grants in the BS grant buffer of SS i at an arbitrary moment is the same as the one at the last embedded epoch.

The number of i-packets in the outgoing buffer of SS i at an arbitrary moment is the sum of the i-packets at the last embedded observation epoch and those, which arrive in the interval from the last embedded observation epoch to the arbitrary moment. This interval is the backward recurrence cycle time, whose mean length is $\frac{T_f}{2}$. Thus using (17) and (18), the mean number of i-packets in the outgoing buffer of SS i at an arbitrary moment can be expressed as

$$E[q_i] = \sum_{j=0}^{X_i} \sum_{k=0}^{X_i} j \; p_i^{(e)}(j,k) + \sum_{j=0}^{X_i} \sum_{k=0}^{X_i} k \; p_i^{(e)}(j,k) + \frac{\lambda_i T_f}{2}. \tag{19}$$

4 Overall Delay Analysis

4.1 The components of the overall delay

We define the *overall delay* (W_i) of the tagged i-packet as the time interval spent from its arrival into the outgoing buffer of SS i up to the end of its successful transmission in the UL. We define also the *grant time of the tagged i-packet* as the time of the i-scheduling event of its i-grant, i.e. the end epoch of the first contention-based transmission opportunity in the RI in the frame preceding the one, in which the tagged i-packet is transmitted. The *overall delay* is composed of several parts:

$$W_i = W_i^r + \nu + W_i^s + W_i^t + \tau. \tag{20}$$

where the individual parts are defined as follows.

- W_i^r – reservation delay from the moment the i-packet arrives into the outgoing buffer of SS i to the start of the successful transmission of the corresponding BW-Req in the RI.
- ν – time of the successful BW-Req transmission, which equals the duration of the transmission opportunity.
- W_i^s – scheduling delay from the end of the successful BW-Req transmission of the tagged i-packet to its grant time.
- W_i^t – transmission delay from the grant time of the tagged i-packet to the start of its successful transmission in the next UL sub-frame
- τ – data packet transmission time.

4.2 Reservation and scheduling delays

The definition of the reservation delay implies that the reservation delay of the tagged i-packet is exactly the sojourn time of the tagged i-packet in the outgoing buffer of SS i. Similarly, it follows from the definition of the scheduling delay that it equals to the sojourn time of the i-grant assigned to the tagged i-packet in the BS grant buffer of SS i.

Consequently, the mean of the sum of the reservation and scheduling delays can be determined by applying Little's law on the mean number of joint content of the outgoing and BS grant buffers of SS i at an arbitrary moment. Using (19), this leads to

$$E\left[W_i^r + W_i^s\right] = \frac{1}{\lambda_i}\left(\sum_{j=0}^{X_i}\sum_{k=0}^{X_i} j\ p_i^{(e)}(j,k) + \sum_{j=0}^{X_i}\sum_{k=0}^{X_i} k\ p_i^{(e)}(j,k)\right) + \frac{T_f}{2}. \tag{21}$$

4.3 Transmission delay

The transmission delay is a sum of the fixed time from the grant time of the tagged i-packet to the start of transmission of the i-packets in the next frame. Hence the mean transmission delay can be expressed as

$$E[W_i^t] = T_f + (K-1)\nu + (i-1)\tau. \tag{22}$$

4.4 Mean overall delay

Taking the mean of (20) and substituting the expressions (21) and (22), we obtain the expression of the mean overall delay as

$$E\left[W_i\right] = \frac{1}{\lambda_i}\left(\sum_{j=0}^{X_i}\sum_{k=0}^{X_i} j\ p_i^{(e)}(j,k) + \sum_{j=0}^{X_i}\sum_{k=0}^{X_i} k\ p_i^{(e)}(j,k)\right) + \frac{3T_f}{2} + i\tau + K\nu. \quad (23)$$

5 Numerical Results and Conclusion

In this section, we apply the derived analytical model to the performance evaluation of the uplink nrtPS packet service in the IEEE 802.16-2009 network [1] with contention-based reservation mechanism. We provide numerical examples to assess the performance of the IEEE 802.16 uplink nrtPS service flow evaluated with the considered analytical model. In order to generate performance data, a simulation program for IEEE 802.16-2009 MAC was developed. The program is an event-driven simulator that accounts for the discussed assumptions of the considered system model.

Table 1. Basic evaluation parameters

Parameter	Value
PHY layer	OFDMA
Frame duration (T_f)	5 ms
Sub-channelization mode	PUSC
DL/UL ratio	2 : 1
Channel bandwidth	10 MHz
MCS	16 QAM $^3/_4$
Packet length	160 bytes
Number of SSs (N)	15
Total capacity per frame for all SSs	15 packets

In our simulations, we set the default values recommended by WiMAX Forum [9] system evaluation methodology, which are also common values used in practice [10] (see Table 1). We assume a 10 MHz TDD system with 5 ms frame duration, PUSC sub-channelization mode and a DL : UL ratio of 2 : 1. In the numerical examples we use normalized arrival flow rate, in which the critical arrival rate that saturates the system is 1.

According to [11], the UL sub-frame comprises 175 slots. Assuming MCS of 16 QAM $^3/_4$, the IEEE 802.16-2009 system transmits 16 bytes per UL slot. We consider fixed packet length of 1600 bytes (10 slots) for all service flows, which results in capacity for sending 15 packets per UL sub-frame. The remaining 25 UL slots represent the necessary control overhead including exactly 1 contention-based transmission opportunity per RI to minimize the overhead (i.e. $K = 1$).

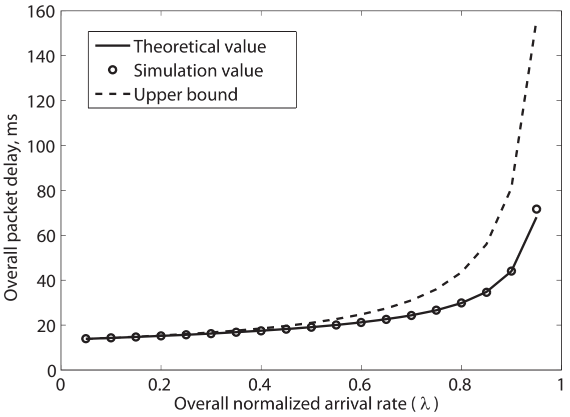


Fig. 3. Numerical results: overall delay vs. arrival rate

The determination of the probability of the successful BW-Req transmission $p_i^{(b)}$ for symmetric system can be found in our previous work [8].

In Figure 3 and Figure 4, we compare analytical and simulation results for the mean overall packet delay. In particular, in Figure 3 and in Figure 4 the dependency of the mean delay on the overall normalized arrival rate at fixed $p_i^{(b)} = 0.5$ and on $p_i^{(b)}$ at fixed normalized $\lambda = 0.5$ can be seen, respectively. The presented analytical approach to the overall mean packet delay computation demonstrates excellent agreement with simulation data. Moreover, we also plot the closed-form analytical upper bound from [8] to conclude that the currently proposed analysis is much more precise in predicting the mean delay values.

6 Final Remarks

It is our future work to use a public simulator to further verify the presented analytical approach also under real-world settings.

Furthermore, we notice that a potential perspective to continue this work is to investigate the determination of the probability of the successful BW-Req transmission $p_i^{(b)}$ also for the asymmetric system. It would enable the relaxation of the assumption on symmetric uplink traffic used in the numerical evaluation.

Acknowledgments

This work is supported by the Russian Foundation for Basic Research (projects #10-08-01071-a and #11-08-01016-a) as well as by the Branch of Nano- and Information Technologies of Russian Academy of Sciences (project 2.3).

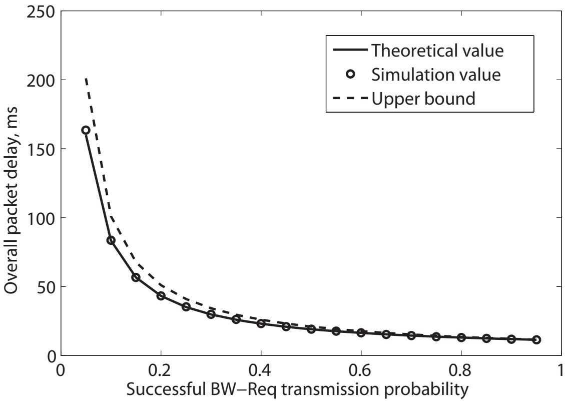


Fig. 4. Numerical results: overall delay vs. contention success probability

References

1. IEEE 802.16-2009, Part 16: Air Interface for Broadband Wireless Access Systems, Standard for Local and Metropolitan Area Networks, May 2009.
2. Vinel, A., Zhang, Y., Ni, Q., Lyakhov, A.: Efficient request mechanism usage in IEEE 802.16. In: IEEE Global Telecomm. Conf. (GLOBECOM). (2006)
3. Vinel, A., Ni, Q., Staehle, D., Turlikov, A.: Capacity analysis of reservation-based random access for broadband wireless access networks. IEEE Journal on Selected Areas in Communications **27**(2) (2009) 172–181
4. Saffer, Z., Andreev, S., Koucheryavy, Y.: Performance evaluation of uplink delay-tolerant packet service in IEEE 802.16-based networks. EURASIP Journal on Wireless Communications and Networking **1** (2011) 1–12
5. Redana, S., Lott, M.: Performance analysis of IEEE 802.16a in mesh operation mode. In: Proc. of the 13th IST SUMMIT, Lyon, France (June 2004)
6. Klein, A., Pries, R., Staehle, D.: Performance study of the WiMAX FDD mode. In: Proc. of the OPNETWORK 2006, Washington D.C. (August 2006)
7. Doha, A., Hassanein, H., Takahara, G.: Performance evaluation of reservation medium access control in IEEE 802.16 networks. In: IEEE International Conference on Computer Systems and Applications. (March 2006) 369–374
8. Andreev, S., Saffer, Z., Turlikov, A., Vinel, A.: Upper bound on overall delay in wireless broadband networks with non real-time traffic. In: Int. Conf. on Analyt. and Stochast. Modeling Techniques and Applications (ASMTA). (2010) 262–276
9. WiMAX Forum, home page: $http://www.wimaxforum.org/$
10. Sivchenko, D., Bayer, N., Xu, B., Rakocevic, V., Habermann, J.: Internet traffic performance in IEEE 802.16 networks. In: European Wireless Conf. (2006)
11. So-In, C., Jain, R., Tamimi, A.K.: Capacity evaluation for IEEE 802.16e mobile WiMAX. Journal of Computer Systems, Networks, and Communications **1** (2010) 1–12

Publication 4

S. Andreev, Y. Koucheryavy, and L. Sousa, "Calculation of transmission probability in heterogeneous ad hoc networks," in *Proc. of the Baltic Congress on Future Internet and Communications (BCFIC)*, pp. 75–82, 2011.

Calculation of Transmission Probability in Heterogeneous Ad Hoc Networks

Sergey Andreev, Yevgeni Koucheryavy
Tampere University of Technology
Tampere, FINLAND
Email: sergey.andreev@tut.fi, yk@cs.tut.fi

Luis Filipe Dias de Sousa
Network and Communication Group (NCG)
Amberg, GERMANY
Email: luis.sousa@advantech.de

***Abstract*—This paper addresses the problem of MAC performance evaluation of a contemporary IEEE 802.11 WLAN. The network is observed under saturation conditions and the packet transmission probability analysis is conducted with the novel regenerative approach. The proposed model accounts for collision resolution protocol parameters, packet retry limit, coexistence of unicast and broadcast traffic, and heterogeneous QoS environment. Our analytical model is a generalization of several well-known models extensively used in the field. The obtained results are verified to demonstrate perfect agreement with ns-2 simulations.**

Keywords: ad hoc network, saturation, collision resolution, transmission probability, throughput.

I. Introduction

The *Wireless Local Area Networks* (WLANs) based on IEEE 802.11 standard [1] is a rapidly growing technology worldwide. Whereas this type of networks is easy to deploy, it provides an effective low-cost solution to achieve wireless connectivity between various devices. There are important special cases of a WLAN, such as *Ad Hoc* networks and *Mobile Ad Hoc Networks* (MANETs). An Ad Hoc network is a self-configuring network of devices connected wirelessly. What differentiates MANETs from Ad Hoc networks is the mobility of connected devices. The emerging IEEE 802.11 technologies motivate the need to evaluate the network performance in order to support increasing user population and to exploit limited wireless resources more efficiently. The lack of infrastructure makes the *Medium Access Control* (MAC) analysis of an Ad Hoc network challenging.

Typically, Ad Hoc networks rely on two kinds of traffic: unicast traffic of point-to-point connections and broadcast traffic of point-to-multipoint connections. As these two types of traffic serve different purposes and generally coexist within a network, neither of them should be neglected at the performance evaluation stage. It is known that the user population of contemporary WLANs is increasing, whereas not all the network clients have equal necessities with respect to channel resources. For this reason, traffic differentiation is crucial in modern networks. As such, a group of privileged users (e.g. those generating real-time traffic) may have higher channel access probability than the others.

Enabling traffic differentiation, *Enhanced Distributed Channel Access* (EDCA) [2] replaced the legacy *Distributed Coordination Function* (DCF) of IEEE 802.11 [1] in order to provide improved *Quality of Service* (QoS) to wireless clients. The extended version of the protocol defines QoS levels for various types of traffic and allows for better utilization of limited network resources. However, in the research literature, the influence of broadcast traffic on the performance of diverse user groups (heterogeneity) has never been studied. Therefore, the focus of this work is to propose a novel tractable analytical model that captures the IEEE 802.11 MAC saturation performance. The model accounts for the mixture of traffic, heterogeneous groups of users with different QoS requirements, and limited number of retries after a failed transmission.

II. Research Background

A. Collision Resolution Protocol

As long as in 1975, Lam and Kleinrock [3] introduced a heuristic algorithm they called *Retransmission Control Procedure* (RCP) that is believed to be the first version of the notorious *Binary Exponential Back-off* (BEB) protocol [4]. Accounting for the channel conditions, the BEB protocol adapts the probability with which users access the medium by changing its *Contention Window* (CW) length parameter. The seeming simplicity of BEB motivated many researchers to address its performance by means of mathematical models. Research works in [5] and [6] proposed useful Markov chain models that capture the stationary equilibrium performance of a saturated network. Following these approaches, it is possible to obtain the saturation throughput by expressing the packet transmission probability as a function of the CW length.

Whereas some BEB models (e.g. [6]) indeed help to understand important BEB features, the actual protocol implementation [2] truncates the CW growth at certain stage to control medium access fairness. Studying it back in 2000, Bianchi [5] assumed that in the stationary equilibrium the *conditional collision probability* of each user is constant. This allowed him to construct a simple but nevertheless very accurate saturation model. Bianchi's key assumption has been believed to be controversial until Bordenave et al. [7] strictly proved that it is valid for a reasonably large number of network users.

Under the same set of assumptions as in [5], Medepalli and Tobagi [8] proposed an alternative model based on the *average cycle time* approach. The network throughput was also established and the new model was argued to be more accurate

978-1-4244-8513-0/11/$26.00 ©2011 IEEE

than that by Bianchi [5]. However, the actual difference between the models is only due to some timing simplifications adopted by Bianchi rather than a methodological flaw. Neither model, however, takes into account several practical features of the BEB operation within IEEE 802.11 [2], such as finite packet retry limit. Markov chains are still usable to account for the additional parameter, but few works present close-form transmission probability functions due to increasing analytical complexity.

Recently, we [9] introduced an alternative technique based on *regeneration cycle concept* to extend [5] and account for the packet retry limit analytically. The new approach is a powerful tool to extend existing models as it only requires operations with simple mathematical series. By contrast, the research work proposed by Wu et al. [10] adopts Markov chain technique, but the result is not reliable since it does not converge to Bianchi's model, when the retry limit tends to infinity. Kwak et al. [6] and later Oliveira et al. [11] truncate the respective Markov chain to also account for the effect of the retry limit, but their solutions are not complete.

B. Broadcast Traffic

The majority of research works on BEB performance only account for unicast traffic. However, in a real network unicast and broadcast traffic flows often coexist. There are only a few papers, where broadcast traffic is considered. In 2007, Ma and Chen [12] studied saturation throughput of broadcast traffic only. Later, Chen et al. [13] extended [12] to account for the mean packet delay and the packet delivery ratio. Finally, in 2008 this research group combined [12] and [13] to publish a revised paper [14], where the influence of the initial CW length on the network performance is addressed.

One more set of results for saturated broadcast traffic by Wang and Hassan [15] was based on the approach somewhat similar to that of [12]. They established that for a one-hop network, broadcast reliability does not depend on the packet size but rather depends on the initial CW length. What differentiates [12] from [15] is that the latter accounts for more realistic BEB counter freezing, when the channel gets busy.

C. Coexistence of Unicast and Broadcast Traffic

Another independent research group addressed the mixture of traffic in a network. The first work of 2006 by Oliveira et al. [11] is an extension of [5], where the percentage of generated broadcast traffic was accounted for. Later in 2007, Oliveira et al. [16] extended their model to account for the unsaturated traffic. The combined results of [11] and [16] were presented in [17], where several important conclusions were made.

In 2009, Wang et al. [18] proposed a model that extends [14] and [15] by considering both saturated and unsaturated traffic and the freezing process of BEB counter, when the channel is busy. The authors concluded that when compared to unicast traffic, broadcast achieves higher optimal throughput under low traffic conditions. However, as the load increases the throughput of broadcast deteriorates much faster than that of unicast. Another research work of Wang et al. [19] extended and improved [11] and [16] by considering the freezing process of BEB and unsaturated traffic. This work evaluated on a separate basis the unicast and the broadcast throughput.

D. Heterogeneity

All the research works presented so far only account for one group of homogeneous users accessing the channel. However, contemporary networks comprise diverse groups that may need service differentiation. In 2003, Li and Battiti [20] presented an extension of Bianchi's work [5], where heterogeneous groups of users were considered in saturation. Later in 2005, Bellalta et al. [21] formulated a model for unsaturated heterogeneous networks. This model is a useful tool to evaluate the aggregate network throughput and queue utilization. In 2007, Malone et al. [22] also proposed their model for unsaturated traffic. This model is more complete than that in [21] as it evaluates both aggregate throughput and per-node throughput, mean queueing delay, as well as the influence of the initial CW length on throughput and user fairness.

Summarizing, there are several independent models that capture saturation network behavior, when the unicast and the broadcast traffic flows coexist, e.g. [17] and [19]. However, it is important to construct a model backward compatible with previous well-known models like [5] and [6], which would precisely follow the retransmission mechanism of BEB. Moreover, there is no model addressing the properties of traffic mixture, when heterogeneous groups of users are considered. These are the issues that this research intends to cover in what follows.

III. IEEE 802.11 MAC Layer Protocols

A. Distribution Coordination Function

In 1999, the legacy IEEE 802.11 standard [1] was finalized to define the features of the Physical (PHY) layer and the MAC layer. At MAC, the DCF was a medium access protocol that allowed clients to share wireless channel resources through the use of *Carrier Sense Multiple Access with Collision Avoidance* (CSMA/CA) mechanism based on the BEB collision resolution protocol.

According to the standard, the time interval between frames is called the *Interframe Space* (IFS). IFS values allow various packet types accessing the medium differently. For instance, an *Acknowledgment* (ACK) frame always has priority over a regular data frame. Figure 1 difference between these xIFS times. For a more detailed information on timings the reader is referred to [1] and [2].

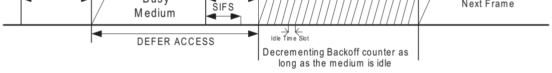

Lower Priority
Immediate access when the medium is free >= DIFS/AIFS
AIFS[j]
Higher Priority
AIFS[i]
DIFS/AIFS
DIFS
Busy Medium
SIFS
Next Frame
DEFER ACCESS
Idle Time Slot
Decrementing Backoff counter as long as the medium is idle

Fig. 1: Time intervals of DCF/EDCA.

The BEB protocol has a so-called *Back-off Counter* (BC) initiated with a random integer between zero and initial CW length minus one. For each idle slot, the BC is decremented by one. This means that the *Carrier Sense* (CS) mechanism has to indicate that the medium is idle. When the BC reaches zero the user is allowed to start the transmission. If a collision occurs the BEB doubles its previously used CW length and the contention process is repeated again. The CW is doubled until the maximum CW length is reached.

The DCF defines two acknowledgment-based transmission schemes for unicast and one acknowledgment-free transmission scheme for broadcast. All three schemes assume immediate transmission after the DIFS or EIFS, when a packet arrives into an empty queue. However, the BEB instance shall be started if a user has a pending packet but the CS indicates that the medium is busy or after an unsuccessful packet transmission.

A packet transmission may fail due to corruption by the channel noise or due to simultaneous transmissions in the same time slot. Thus, a user is ready to send a packet if the medium is idle at least for a period of time equal to DIFS, the *Network Allocation Vector* (NAV) timer shows zero, the PHY indicates that the medium is idle and the BC has reached zero.

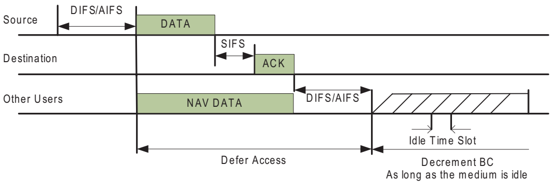


Fig. 2: Unicast basic access scheme mechanism.

There are two mechanisms to send a unicast data frame: the Basic Access and the *Request to Send* (RTS) and *Clear to Send* (CTS) mechanism depicted in Figures 2 and 3, respectively. The basic access is used whenever the data frame length is equal or below a given threshold (*dot11RTSThreshold* in IEEE 802.11). Figure 2 also shows why the basic access is not always a good option for multi-hop networks. The reason is that typically only the sender's neighbors update the NAV timer. Hence, if there are hidden users ready to transmit, the data frames are more likely to be corrupted because of the simultaneous transmissions.

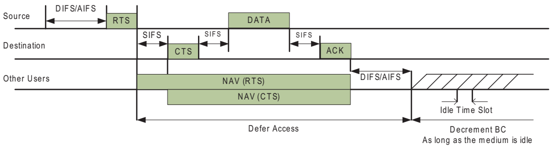


Fig. 3: Unicast RTS/CTS access scheme mechanism.

Figure 3 shows the RTS/CTS mechanism. After the medium is sensed idle for a DIFS, a reservation RTS frame is sent. This frame contains the information about the total time needed to finalize the transmission. If the recipient's CS indicates that the medium is idle, it answers the sender with a CTS. Further, the sender transmits the data frame. The recipient then returns the ACK frame to acknowledge the correct reception of the packet.

The broadcast transmission scheme is a special case of the basic access mechanism. In particular, there is no acknowledgment as there may be multiple candidates to send it. Thus, a packet may be lost and never retransmitted. This is the reason why broadcast traffic is unreliable. The retransmission mechanism is thus a feature of unicast traffic only.

There is a limit on the maximum number of retransmissions that a packet can suffer. If the packet is not successfully received after that many retransmissions, it is discarded. The 802.11 specification allows for two different retry limits, dot11ShortRetryLimit and dot11LongRetryLimit, for packets that are shorter than and longer than the dot11RTSThreshold respectively.

B. Enhanced Distributed Channel Access

The EDCA enhances the IEEE 802.11 DCF by enabling traffic differentiation. The QoS is handled via introducing four separate queues for different packet types. Each one of the queues, or *Access Categories* (ACs), has a separate BEB instance to control the medium access. The four different ACs are: *Voice*, *Video*, *Background*, and *Best Effort*. Typically, the multimedia ACs (Voice and Video) access the channel with higher priority. Figure 4 demonstrates the EDCA model and highlights its differences from the DCF. Each AC has to ensure that the medium is idle for a specific AIFS before a packet transmission. Hence, each AC has its own AIFS time.

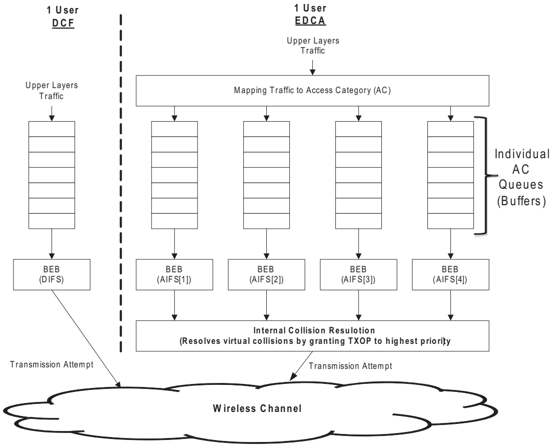


Fig. 4: EDCA reference model.

The unicast and broadcast transmission schemes of EDCA are basically the same as these of DCF. However, EDCA introduces the *Block Acknowledgment* (BA) mechanism. The idea of BA is similar to the fragmentation burst in legacy IEEE 802.11, but instead of sending several fragments of a frame the sender transmits a burst of data frames belonging to a specific AC.

IV. Regeneration Cycle Concept

Following our previous work [9], we define a regeneration cycle under saturation conditions as the time interval from the moment of the first BEB counter generation, to the end of successful packet transmission or packet discard. A simplified example of a regeneration cycle for the infinite number of BEB stages and unlimited number of retransmission attempts is shown in Figure 5. In what follows, W_0 is the initial CW length.

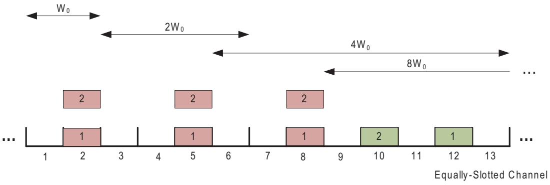


Fig. 5: Regeneration cycle example.

Figure 5 represents a 'classical' lossless system [6] in which a packet is sent until it is successfully received. The regeneration cycle starts in slot 1 for users 1 and 2 and it ends in slots 10 and 12 for users 2 and 1, respectively. Our concept accounts for the average number of packet transmissions during a regeneration cycle. Hence, the equation to calculate the packet transmission probability p_t is as follows:

$$p_t = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} B^{(i)}}{\sum_{i=1}^{n} D^{(i)}} = \frac{E[B]}{E[D]}, \tag{1}$$

where $B^{(i)}$ is the number of transmissions in a cycle and $D^{(i)}$ is the mean number of contention time slots for the i^{th} transmission attempt. Notice that $B^{(i)}$ and $D^{(i)}$ are independent and identically distributed with respect to i, but both are dependent on the conditional collision probability p_c. Here, $E[B]$ stands for the average number of transmissions in a cycle and $E[D]$ is the average number of slots that a user shall back-off until it starts its last transmission.

The regeneration approach for the maximum number of BEB stages m and the packet retry limit k is presented below.

V. Lossy System

A. Baseline Model

Here, the formulation described in [9] is discussed. We now consider a realistic lossy model in which after k unsuccessful transmission attempts the packet is discarded. Moreover, there is also a limiting number of m BEB stages. To compute the p_t within such a system, $E[B]$ and $E[D]$ are given below.

$$\begin{aligned} E[B] &= \sum_{i=1}^{k} i \Pr\{B=i\} = \\ &= \sum_{i=1}^{k} i p_c^{i-1}(1-p_c) + p_c^k k = \frac{1-p_c^k}{1-p_c}. \end{aligned} \tag{2}$$

Notice that there are two equations for $E[D]$: one for $k \le m+1$ and another one for $k > m+1$. These two expressions for $E[D]$ will later result in two different formulas for p_t.

$$\begin{aligned} &E^{(1)}[D] = \\ &= \sum_{i=1}^{k} D^{(1)}(i) \Pr\{B=i\} + p_c^k D^{(1)}(k) = \\ &= \frac{W_0\left(1-p_c\right)\left(1-\left(2p_c\right)^k\right) + \left(1-2p_c\right)\left(1-p_c^k\right)}{2\left(1-2p_c\right)\left(1-p_c\right)}. \end{aligned} \tag{3}$$

$$\begin{aligned} &E^{(2)}[D] = \\ &= \frac{\left(1-2p_c\right)\left(W_0\left(1-2^m p_c^k\right) + \left(1-p_c^k\right)\right) + p_c^k W_0\left(1-\left(2p_c\right)^m\right)}{2\left(1-2p_c\right)\left(1-p_c\right)}. \end{aligned} \tag{4}$$

Finally, with the equations (2), (3) and (4) it is possible to establish p_t as:

$$p_t = \frac{E[B]}{E[D]} = \begin{cases} \frac{E[B]}{E^{(1)}[D]}, & k \le m+1 \\ \frac{E[B]}{E^{(2)}[D]}, & k > m+1 \end{cases}. \tag{5}$$

Interestingly, the first branch of equation (5) was also deduced by Kwak et al. [6] and by Wu et al. [10] (see equation (42) in [6] and equations (8) and (9) in [10]).

B. Proposed Model

Even though describing a more realistic system, equations (3) and (4) only consider unicast traffic. However, users typically transmit both broadcast and unicast traffic. As mentioned above, broadcast relies on one transmission attempt only and, as such, equations (3) and (4) shall be modified. We assume that each user generates a broadcast packet with probability p_b, and a unicast packet with probability $p_u = 1 - p_b$. The $E[B]$ shown in (6) is thus divided into two parts weighted by p_u and p_b respectively:

$$\begin{aligned} &E[B] = \\ &= \overbrace{E_u[B]p_u}^{Unicast} + \overbrace{E_b[B]p_b}^{Broadcast} = \frac{\left(1-p_c^k\right)p_u + (1-p_c)p_b}{1-p_c}. \end{aligned} \tag{6}$$

Likewise what happens with the $E[B]$, the $E[D]$ also changes due to the cycle duration difference. Therefore, there are two expressions for $E[D]$ as in equations (3) and (4). The expression that describes the (re)transmission probability p_t, may still be given by (5) with components determined now by (7) and (8).

$$E^{(1)}[D]=\frac{\left[W_0\left(1-p_c\right)\left(1-\left(2p_c\right)^k\right)+\left(1-2p_c\right)\left(1-p_c^k\right)\right]p_u+\left(W_0+1\right)\left(1-2p_c\right)\left(1-p_c\right)p_b}{2\left(1-2p_c\right)\left(1-p_c\right)}. \tag{7}$$

$$E^{(2)}[D]=\frac{\left[\left(1-2p_c\right)\left(W_0\left(1-2^m p_c^k\right)+\left(1-p_c^k\right)\right)+p_c^k W_0\left(1-\left(2p_c\right)^m\right)\right]p_u+\left(W_0+1\right)\left(1-2p_c\right)\left(1-p_c\right)p_b}{2\left(1-2p_c\right)\left(1-p_c\right)}. \tag{8}$$

C. Backwards Compatibility

This Subsection is dedicated to demonstrate how the proposed p_t expression converges to all the well-known saturation models. This may also be regarded as theoretical model validation. There are two ways to establish the broadcast transmission probability. One is to force $p_u = 0$ and another one is to impose the broadcast parameters on unicast. The broadcast assumes $m = 0$ and $k = 1$, as there are no retransmissions of broadcast packets.

Considering now that $p_b = 0$, equations (7) and (8) converge to equations (3) and (4) respectively. Bianchi's formula [5] can be obtained, when $p_b = 0$ and k tends to infinity. Note that only one constraint $k > m + 1$ may be considered, since in Bianchi's approach k is infinite. Finally, to prove the compatibility with the work of Kwak et al. [6], it is only necessary to consider the $k \leq m + 1$ branch of p_t. This is the special case when k and m both converge to infinity.

VI. Network Heterogeneity

We remind that in a real network there are various types of users with diverse necessities. These users might need different settings to access the wireless channel with prioritized probabilities.

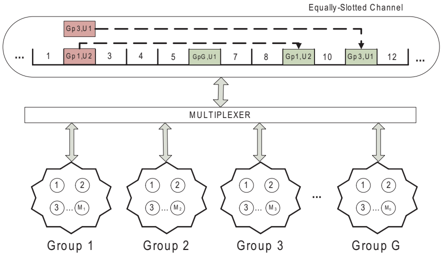


Fig. 6: General topology of a heterogeneous network.

In Figure 6, the stars represent groups number $1, 2, \ldots, G$. Inside each star, there are circles that represent users number $1, 2, \ldots, M_j$, where M_j represents the number of users in the group number j. This kind of approach is not new and was introduced in [20]. Additionally, the Figure demonstrates packet collisions and packets that are send successfully over the channel.

Summarizing, it is assumed that the network population is composed of G different groups as depicted in Figure 6. In what follows, heterogeneity is referred to as network with groups of users that have different channel access probabilities. These groups are heterogeneous between themselves, however, all users belonging to a certain group j are homogeneous. Inside each group, users thus share the same parameters.

Each group j has its own transmission probability $p_t^{(j)}$ and, consequently, observes different conditional collision probability $\mathrm{p}_c^{(j)}$:

$$\mathrm{p}_c^{(j)}=1-\left(1-p_t^{(j)}\right)^{M_j-1}\prod_{i=1;i\neq j}^{G}\left(1-p_t^{(i)}\right)^{M_i}. \tag{9}$$

The equation (9) shows how the conditional collision probability may be obtained for a generic group j. Note that a collision can be intergroup, intragroup or both at the same time. An intergroup is a collision between users belonging to different groups. An intragroup is a collision between users of the same group. The transmission probability $p_t^{(j)}$ is calculated as per (5) according to the parameters of group j:

$$p_t^{(j)}(W_0 = W_0^{(j)}, m = m^{(j)}, k = k^{(j)}, p_b = p_b^{(j)}, p_c = \mathrm{p}_c^{(j)}). \tag{10}$$

Notice that equations (9) and (10) constitute a system of equations with a numerical solution. This solution is unique for any $G \geq 1$, as Li and Battiti describe in [20].

Once the transmission probability for each group is established, the system-wide probabilities can be studied. These probabilities are important because they describe the channel dynamics.

The channel can be regarded in terms of all the slots types it contains. Of all the channel slots, there is a number of busy time slots that occur with probability P_{busy} and a number of idle time slots that occur with probability P_{idle}. The busy slots can be subsequently divided into time slots where a successful transmission occurred with probability P_S^* (conditional system-wide success probability) and time slots where collisions occur with probability $1 - P_S^*$. All the considered probabilities are derived below.

$$P_{idle}=\prod_{j=1}^{G}\left(1-p_t^{(j)}\right)^{M_j}. \tag{11}$$

$$P_{busy}=1-P_{idle}. \tag{12}$$

$$P_S^*=\frac{1}{P_{busy}}\cdot \tag{13}$$
$$\cdot\sum_{j=1}^{G}\left\{M_j p_t^{(j)}\left(1-p_t^{(j)}\right)^{M_j-1}\prod_{i=1,i\neq j}^{G}\left(1-p_t^{(i)}\right)^{M_i}\right\}.$$

$$P_c^* = 1 - P_S^*. \tag{14}$$

We note that in equations (13) and (14) the probabilities are conditioned on the fact that the channel is busy. In order to obtain the system throughput (see the following Section), it is necessary to establish the unconditional system-wide probabilities. The equation (15) represents the unconditional success probability of a generic group j, whereas the equation (16) represents the respective unconditional collision probability.

$$P_{S,j} = M_j p_t^{(j)} \left(1 - p_t^{(j)}\right)^{M_j - 1} \prod_{i=1, i \neq j}^{G} \left(1 - p_t^{(i)}\right)^{M_i}. \tag{15}$$

$$P_c = P_{busy} - \sum_{j=1}^{G} P_{S,j}. \tag{16}$$

VII. Unequally-Slotted System

So far, our mathematical analysis assumed that the wireless channel is divided into equal slots in time. Clearly, this assumption does not hold for IEEE 802.11. The unequally-sized time slots allow the increase in throughput as the channel utilization improves. In order to fit the proposed model into unequally-slotted system (e.g. IEEE 802.11), the slot rescaling is needed. The rescaling technique is often performed by some authors, like in [5] or [20]. As such, this Section proposes an extension of the above results to the realistic unequally-slotted system shown in Figure 7.

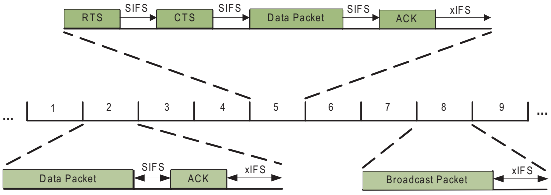


Fig. 7: IEEE 802.11 as an unequally-slotted system.

As we discussed previously, there are essentially three types of packet transmissions in IEEE 802.11: the basic access transmission, the RTS/CTS transmission, and the transmission of a broadcast packet. These three transmission techniques are summarized in Figure 7. Studying the network throughput is easy, when all the packets in a network are sent with the same transmission scheme, at the constant data rate, and have the equal payload length P. Otherwise, the study of the throughput can be a challenge mainly due to the different channel timings.

With the system-wide probabilities deduced previously, it is possible to formulate a simple approximation $\widetilde{S}$ of the network saturation throughput as:

$$\widetilde{S} = \frac{\sum_{j=1}^{G} P_{S,j} E[P]}{P_{idle} T_{idle} + \sum_{j=1}^{G} P_{S,j} T_S + P_c T_c}. \tag{17}$$

The actual slot rescaling is done in (17) by the timings T_{idle}, $T_{S,j}$, and T_c defined by e.g. Li and Battiti in [20]. Notice, that the proposed model can be applied safely in a few cases. These cases are as follows:

1) When clients transmit unicast and broadcast packets with fixed packet payload length and are using the basic access mechanism;
2) When clients transmit unicast packets with RTS/CTS mechanism and the packets have fixed payload lengths;
3) When clients transmit unicast packets with RTS/CTS mechanism and the packets have different payload lengths.

VIII. Model Validation

A. General Settings

In this Section, we validate the proposed model through simulations using the well-known ns-2 simulator [23]. A special traffic generator was developed for ns-2 to generate certain amounts of broadcast/unicast traffic. All validations were done using the basic access transmission scheme, because the collision time difference between the packets is more critical in this case, allowing us to understand how far the model is from the simulation results. However, the same validations can be repeated for RTS/CTS mechanism. The IEEE 802.11b ns-2 implementation was parameterized according to Table I.

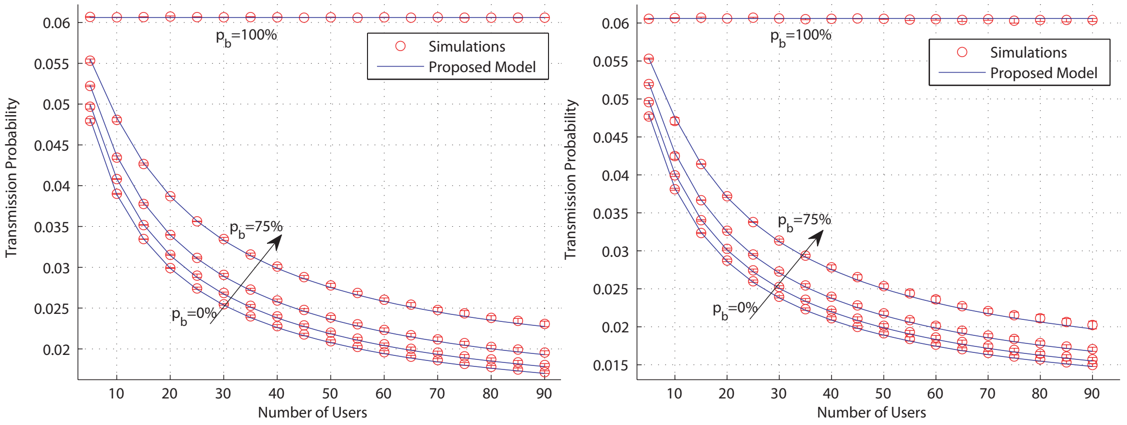
TABLE I: Parameters used in all validation results

ns-2 Settings	
SIFS	10 μs
DIFS	50 μs
EIFS	364 μs
Idle slot time	20 μs
ACK frame duration	27.7 μs
ACK timeout (aprox.)	304 μs
Propagation delay	1 μs
Data rate	11.0 Mbps
Frame size	1500 bytes
Simulation time	750 s

In this Section, only the transmission probability p_t is validated by simulations as it is the main result of this work and there is a direct relationship between it and the system-wide probabilities presented in Section VI. Moreover, the throughput expression given by equation (17) depends indirectly on the transmission probability, because it relies on system-wide probabilities. The practical results are presented by their average values with 95% confidence interval.

For all validations, two different scenarios were adopted:

- Homogeneous scenario: one group of users;
- Simplified EDCA scenario: four heterogeneous groups of users. Each group belongs to a particular AC.



(a) Case 1 validation

(b) Case 2 validation

Fig. 8: Transmission probability validation.

B. Transmission Probability Validation

This Subsection is dedicated to the transmission probability validation. For one group of users, two scenarios were simulated. These scenarios are shown in Table II and correspond to both branches of the proposed p_t equation.

TABLE II: Parameters used for one homogeneous group

	Case 1	Case 2
W_0	32	32
m	5	3
k	5	5
p_b	0;0.25;0.5;0.75;1	

Figure 8 contrasts the theoretical values of p_t against the simulation results.

We observe that the model predictions are very accurate as the transmission probability is directly related to the initial CW length, the number of BEB stages, the packet retry limit, the number of users, and the amount of broadcast traffic, as shown by equation (10). As the number of users in the network increases, the collision probability also grows and, consequently, the p_t decreases due to BEB CW management mechanism.

In order to validate a simplified EDCA scenario, four heterogeneous groups of users were considered. Each group represents one traffic class and its access parameters are shown in Table III. The scenario is simplified as we use several unrealistic assumptions about channel timings following partly by [24]. Therefore, the throughput within our model is expected to diverge from the actual value as proportion of broadcast traffic increases. However, here we focus on the transmission probability validation.

The transmission probabilities for the simplified EDCA evaluation are a special case of four heterogeneous groups of users. The respective validation may be found in Table IV showing absolute difference between analytical and simulation results of p_t. The relative error is small for all the ACs.

TABLE III: Parameters used for simplified EDCA scenario

Parameter	Voice	Video	Background	Best Effort
W_0	8	16	16	32
m	1	1	6	5
k	4	4	7	6
p_b	0			

TABLE IV: Validation of p_t for simplified EDCA scenario

Users	Voice	Video	Background	Best Effort
2	0.8 %	0.3 %	1.5 %	1.1 %
4	0.2 %	0.1 %	2.3 %	1.7 %
6	0.1 %	0.6 %	2.2 %	0.6 %
8	0.1 %	0.4 %	2.2 %	0.5 %
10	0.1 %	0.2 %	1.4 %	0.8 %

IX. Conclusions

This work accomplished a deep analytical study of saturated IEEE 802.11 (Wi-Fi) networks. Currently, several disjoint models exist to address their various features. In his work, we synthesized a novel general model, which includes the most well-known models as special cases. In particular, the proposed model accounts for collision resolution protocol parameters, its retransmission procedure, coexistence of unicast and broadcast traffic, and heterogeneous QoS environment.

In particular, the notorious binary exponential back-off collision resolution protocol was analyzed under realistic assumptions. The regeneration cycle approach was successfully applied to the proposed model, where BEB was described by means of its initial contention window length, maximum number of its stages, packet retry limit, and probability with which a broadcast packet is generated.

Additionally, we proposed a simple and accurate approach to evaluate the main performance metrics, whereas currently more complicated techniques are typically used. The model was validated theoretically by proving the backwards compatibility to previous well-known models. The ns-2 simulations demonstrated that the model predicts very accurately the transmission probability for both homogeneous and heterogeneous scenarios.

References

[1] "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, IEEE standard 802.11, 1999 (revised 2003)."

[2] "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, IEEE standard 802.11, 2007."

[3] S. Lam and L. Kleinrock, "Packet switching in a multiaccess broadcast channel: Dynamic control procedures," *IEEE Transactions on Communications*, vol. com-23, pp. 891–904, 1975.

[4] S. Lam, "Adaptive backoff algorithms for multiple access: A history," *see http://www.cs.utexas.edu/users/lam/NRL/backoff.html*.

[5] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *J. Select. Areas Commun.*, vol. 18, no. 3, pp. 535–547, 2000.

[6] B.-J. Kwak, N. Song, and L. E. Miller, "Performance analysis of exponential backoff," *IEEE Transactions on Networking*, vol. 13, pp. 343–355, 2005.

[7] C. Bordenave, D. McDonald, and A. Proutire, "Random multi-access algorithms - a mean field analysis," *Rapport de Recherche*, vol. 5632, pp. 1–12, 2005.

[8] K. Medepalli and F. A. Tobagi, "Throughput analysis of IEEE 802.11 wireless LANs using an average cycle time approach," *IEEE Globecom*, pp. 3007–3011, 2005.

[9] S. Andreev and A. Turlikov, "Binary exponential backoff algorithm analysis in the lossy system with frames," *In the Proc. of the XII International Symposium on Problems of Redundancy in Information and Control Systems*, pp. 201–210, 2009.

[10] H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma, "Performance of reliable transport protocol over IEEE 802.11 wireless LAN: Analysis and enhancement," *Proceedings of IEEE INFOCOM*, vol. 2, pp. 599–607, 2002.

[11] R. Oliveira, L. Bernardo, and P. Pinto, "Performance analysis of the IEEE 802.11 distributed coordination function with unicast and broadcast traffic," *IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, pp. 1–5, 2006.

[12] X. Ma and X. Chen, "Saturation performance of IEEE 802.11 broadcast networks," *IEEE Communications Letters*, vol. 11, pp. 686–688, 2007.

[13] X. Chen, H. H. Refai, and X. Ma, "Saturation performance of IEEE 802.11 broadcast scheme in ad hoc wireless LANs," *IEEE 66th Vehicular Technology Conference*, pp. 1897–1901, 2007.

[14] X. Ma and X. Chen, "Performance analysis of IEEE 802.11 broadcast scheme in ad hoc wireless LANs," *IEEE Transactions on Vehicular Technology*, vol. 57, N. 6, pp. 3757–3768, 2008.

[15] Z. Wang and M. Hassan, "Analytical evaluation of the 802.11 wireless broadcast under saturated conditions," *School of Computer Science and Engineering, University of New South Wales*, vol. UNSW-CSETR-0801, pp. 1–21, 2008.

[16] R. Oliveira, L. Bernardo, and P. Pinto, "Modelling delay on IEEE 802.11 MAC protocol for unicast and broadcast nonsaturated traffic," *Wireless Communications and Networking Conference*, pp. 463–467, 2007.

[17] R. Oliveira, L. Bernardo, and P. Pinto, "The influence of broadcast traffic on IEEE 802.11 DCF networks," *Elsevier, Computer Comunications*, vol. 32, pp. 439–452, 2009.

[18] J. C.-P. Wang, D. R. Franklin, M. Abolhasan, and F. Safaei, "Characterising the behaviour of IEEE 802.11 broadcast transmissions in ad hoc wireless LANs," *IEEE International Conference on Communications*, pp. 1–5, 2009.

[19] J. C.-P. Wang, D. R. Franklin, M. Abolhasan, and F. Safaei, "Characterising the interactions between unicast and broadcast in IEEE 802.11 ad hoc networks," *Telecommunication Networks and Applications Conference*, pp. 180–185, 2008.

[20] B. Li and R. Battiti, "Performance analysis of an enhanced IEEE 802.11 distributed coordination function supporting service diferentiation," *Springer-Verlag Berlin Heidelberg*, vol. LNCS 2811, pp. 152–161, 2003.

[21] B. Bellalta, M. Oliver, M. Meo, and M. Gerrero, "A simple model of the IEEE 802.11 MAC protocol with heterogeneous traffic flows," *IEEE Region 8 Eurocon*, 2005.

[22] D. Malone, K. Duffy, and D. Leith, "Modeling the 802.11 distributed coordenation function in non-saturated heterogeneous conditions," *IEEE/ACM Transactions on Networks*, vol. 15, No.1, pp. 159–172, 2007.

[23] ns 2, "Network simulator 2," *see http://www.isi.edu/nsnam/ns/*, 2009.

[24] G. Sharma, A. Ganesh, P. Key, and R. Needham, "Performance analysis of contention based medium access control protocols," *IEEE International Conference on Computer Communications*, vol. 0743-166X, pp. 1–12, 2006.

Publication 5

S. Andreev, K. Dubkov, and A. Turlikov, “IEEE 802.11 and 802.16 cooperation within multi-radio stations,” *Wireless Personal Communications Journal (WIRE)*, vol. 58, no. 3, pp. 525–543, 2011.

IEEE 802.11 and 802.16 Cooperation Within Multi-Radio Stations

Sergey Andreev[1], Konstantin Dubkov[2], and Andrey Turlikov[3]

[1] Tampere University of Technology, Finland
[2] Intel Corporation, St. Petersburg, Russia
[3] St. Petersburg State University of Aerospace Instrumentation, Russia

Abstract. In this paper we consider a multi-radio wireless network client that is capable of simultaneous operation in IEEE 802.16 and IEEE 802.11 telecommunication networks. In order to enable the cooperative functioning of both networks we introduce the media access control coordination concept. A set of coordination algorithms is then presented together with a simple approach to their performance analysis. Our performance evaluation shows that the saturation goodput of the proposed coordination algorithm is at least 50% higher than that of the existing coordination algorithms. Moreover, it allows for the considerable reduction in the data packet delay.

1 Introduction and previous work

Wireless technology becomes more widespread as new telecommunication protocols emerge, which enable higher data rates. The parallel evolution of personal, local and metropolitan area networks provides the mobile clients with a wide choice of which infrastructure to use for a given application. Recent advances in the area introduce wireless systems that exploit multiple radios in a collaborative manner. The use of such *multi-radio* devices was shown to dramatically improve the overall system performance and functionality over the traditional single-radio wireless systems.

The first works on multi-radio performance considered IEEE 802.11 (WiFi) [1] telecommunication protocol in the wireless mesh mode. Equipping the mesh routers with multiple radios tuned to non-overlapping channels was studied by many authors. In [2] some common problems in wireless networking were revisited in the multi-radio context, including energy management, capacity enhancement, mobility management, channel failure recovery and last-hop packet scheduling. A novel link layer protocol for multihop community wireless mesh network, where cost of the radios and battery consumption are not limiting factors, was presented and analyzed in [3].

A new metric for routing in multi-radio multihop wireless networks was given by [4]. The authors focused on wireless networks with stationary nodes, such as community wireless networks. The goal of the metric was to choose a high-throughput path between a source and a destination. The authors of [5] mathematically formulated the joint channel assignment and routing problem, taking into account the interference constraints, the number of channels in a network and the number of radios available at each mesh router. They strictly proved that equipping wireless routers with multiple radios improves the capacity by transmitting over multiple radios simultaneously using orthogonal channels.

In [6] specific mechanisms were defined that can transform partially overlapped channels into an advantage, instead of a peril in a wireless network. The work [7] emphasized that the channel assignment presents a challenge because co-located wireless networks are likely to be tuned to the same channels. The resulting increase in interference can adversely affect performance. Therefore, the authors presented an interference-aware channel assignment algorithm for multi-radio wireless mesh networks that addresses this interference problem. The proposed solution intelligently assigned channels to radios to minimize the interference within the mesh network and between the mesh network and co-located wireless networks.

Finally, the design and experimental study of a distributed, self-stabilizing mechanism was conducted by [8] to assign channels to multi-radio nodes in wireless mesh networks. However, with the introduction and further development of IEEE 802.16 (WiMAX) [9] protocol the concept of a multi-radio station was extended to also cover the interworking of several wireless technologies. The multi-radio station may thus operate in several telecommunication networks at the same time in accordance with the inbuilt protocols.

The problems caused by the multi-protocol operation at the *media access control* (MAC) layer has yet received much attention in the scientific literature. The primary focus of [10] is set on the co-existence scenarios between 802.11 and 802.16 in which 802.11 hotspots are inside a 802.16 cell and share the same frequency band. The authors propose new schemes that can control transmit frequency, power, and time of transmission. Three basic schemes are thus proposed: dynamic frequency selection, power control and time agility.

In [11] the above research is continued with the formulation of common spectrum coordination channel etiquette protocol. This protocol is used to exchange control information on transmitter and receiver parameters and hence to cooperatively adapt key variables such as frequency or power. However, the approach of [11] assumes that a common spectrum coordination channel at the edge of available spectrum bands is allocated for announcement of radio parameters. The use of a dedicated channel limits the practical applicability of this research.

Another option to enable 802.16 and 802.11 cooperation is demonstrated in [12] where the capability of 802.11 reuse by 802.16 in the mesh mode is studied. In [13] a general co-existence evaluation approach is shown and [14], particularly, addresses 802.11e and 802.16 interworking, where a concept of the Base Station Hybrid Coordinator is introduced. The use of such a coordinator is possible, when the base station of 802.16 and the hybrid coordinator of 802.11e are co-located. As an alternative, the authors describe some software upgrades to the MAC of the 802.16 base station in [15]. These updates ensure reliable operation of 802.16 when sharing unlicensed spectrum with 802.11.

Some works (see, for example, [16], [17] and [18]) also cover IEEE 802.15.1 (Bluetooth) and 802.11 co-existence issues. In this paper we address the practical case of cooperation between 802.11 and 802.16 standards. But by contrast to the approach of [14] and [15] we consider a more realistic scenario without any modification of the central coordinating node in the telecommunication system. Instead, we discuss the problem of the MAC coordination within a client multi-radio station itself, thus avoiding any restriction on the network topology.

The rest of the text is structured as follows. In Section II we provide a deeper insight into the separate functioning of 802.11 and 802.16. Section III introduces the concept of the MAC coordination and presents a set of coordination algorithms. Section IV analytically evaluates the performance of these algorithms from the MAC goodput viewpoint. In Section V the simulation results are presented and Section VI concludes the paper.

2 Non-cooperative network functioning

2.1 IEEE 802.11 standard

From the MAC layer point of view, the contemporary IEEE 802.11 standard provides both distributed and centralized multiple-access protocols to the shared communications channel. 802.11 supports several operation modes, including the most common *infrastructure* mode, in which an *access point* (AP) becomes the central network node. The AP arbitrates all communication between the stations (STAs) and is mandated to use a contention-based channel access protocol, that is built on top of the truncated binary exponential backoff collision resolution algorithm [19]. The channel access protocol itself is termed carrier sense multiple access with collision avoidance (CSMA/CA). It is fully defined by three parameters: the *arbitration inter-frame space* (AIFS) interval, which each station waits prior to the channel contention and the pair of the minimum (W_{min}) and the maximum (W_{max}) contention windows, that regulate the uniform sampling of random numbers and enable the collision avoidance feature of the protocol.

Starting from 802.11e version [1] the standard introduces quality of service (QoS) enhancements and adopts the concept of a *transmission opportunity* (TXOP), which is illustrated in Fig. 1. A TXOP may be regarded as a bounded time interval during which a sequence of packets encapsulated into frames is transmitted by a source station, while only service messages are received. A TXOP is only obtained if both channel state detection functions of a station indicate that the channel is idle. They are the *clear channel assessment* (CCA) algorithm at the physical (PHY) layer and the *network allocation vector* (NAV) value at the MAC layer. As discussed above, a TXOP is precluded by the deterministic AIFS interval and then by a random number of slots.

Commonly, a source station initiates a frame transaction with a *request to send* (RTS) frame, which is responded by the destination station with a *clear to send* (CTS) frame after a *short inter-frame space* (SIFS) interval. The source station then transmits aggregated data packet (DATA) with a single PHY layer

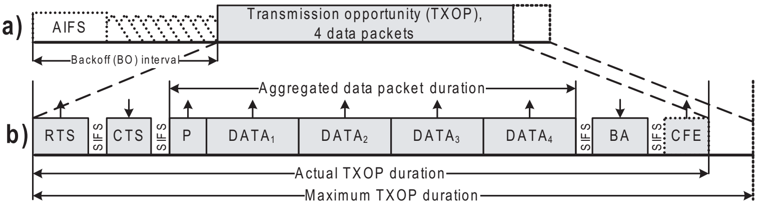


Fig. 1. IEEE 802.11 TXOP: a – example frame transaction; b – typical time structure

preamble, which is subject to the acknowledgment by a *block acknowledgment* (BA) frame. If some unused TXOP time remains, the source station may release it by a *contention-free end* (CFE) frame.

2.2 IEEE 802.16 standard

IEEE 802.16 MAC layer adopts a schedule-based protocol, commonly operating in the mandatory infrastructure mode. A *base station* (BS) arbitrates all activity within the network and broadcasts both service messages and useful data to its *subscribed stations* (SSs) in the *downlink* (DL) sub-frame. The DL sub-frame is composed of a 802.16 MAC header and DL bursts, directed at the SSs (see Fig. 2). In the *uplink* (UL) sub-frame the SSs transmit scheduled UL bursts as well as service messages. 802.16 supports several PHY layer modes, of which the most practical is the *orthogonal frequency division multiple access* (OFDMA) scheme (see Fig. 2).

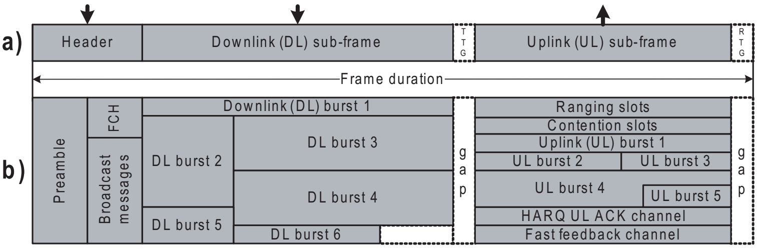


Fig. 2. IEEE 802.16 frame: a – simplified structure; b – detailed OFDMA frame time-frequency structure

IEEE 802.16 was specifically designed to support a variety of traffic types. It should be efficient for high data rate applications (video streaming) as well as for low data rate applications (web surfing). IEEE 802.16 effectiveness should not degrade in case of bursty traffic and delay-critical applications (voice over IP (VoIP), audio). The main challenge in ensuring QoS requirements in 802.16 is that all the traffic types with respective characteristics should be serviced at the same time. For this purpose the standard defines five QoS classes, which are described below.

1. Unsolicited Grant Service (UGS) is oriented at the real-time traffic where fixed-size data packets are generated periodically (CBR input source).
2. Real-Time Polling Service (rtPS) is oriented at the real-time traffic where variable-size data packets are generated periodically (VBR input source).
3. Non Real-Time Polling Service (nrtPS) is similar to rtPS, but data packet generation is not necessarily periodic.
4. Best Effort (BE) is suitable for applications, where no throughput or delay guarantee is provided.
5. Extended Real-Time Variable Rate (ERT-VR) is similar to rtPS, but with more strict delay requirement (guaranteed jitter) to support real-time applications like VoIP with silence suppression. This class is defined only in the recent IEEE 802.16e [9] standard and is often referred to as Extended Real-Time Polling Service (ertPS).

The MAC layer also supports a variety of bandwidth reservation mechanisms, each of which is assigned to a particular service flow. The mechanisms are based upon unicast, multicast or broadcast polling techniques. However, the standard specifies neither scheduling algorithm nor admission control mechanism.

3 Cooperative network functioning

3.1 MAC coordination concept

Currently IEEE 802.11 and 802.16 standards operate in non-overlapping frequency bands [20]. Therefore, the respective telecommunication networks may coexist simultaneously without any significant performance degradation. However, this is the case only when each client station supports the functionality of exactly one protocol. When the functionalities of two or more standards are co-located within a single multi-radio (MR) station (see Fig. 3) the network performance degrades dramatically, even if the simultaneous operation is technically possible. This is explained by the fact that the radio parts of a MR station are close enough and the ongoing transmission in one network prohibits the reception in another one.

Coexistence enhancement research summarized in [21] states that co-located transmissions or receptions via different standards generally do not deteriorate each other. However, when a station receives data, an overlapping transmission of the co-located technology prevents the successful reception. This effect is elaborated on further in [22]. It is shown that 802.11 and 802.16 radio-to-radio interference severely degrades the performance and requires isolation of at least 55 dB. Increasing isolation is costly, large in size and highly platform dependent.

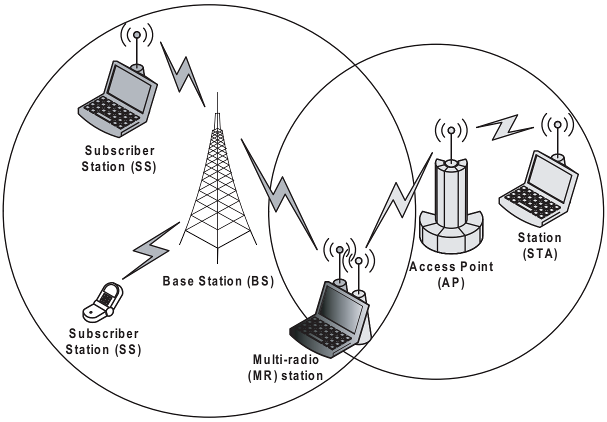


Fig. 3. General cooperative network example

An alternative solution may be pursued in the time domain. In order to mitigate the indicated effect a special module on top of the respective MAC layers may be implemented for the purposes of the *MAC coordination* (MC). This solution is known to be universal, effective and media independent.

The MC module controls scheduling of both network activities within a MR station and thus enables the simultaneous operation of 802.11 and 802.16. As 802.16 is schedule-based, the MC module only monitors its transmit (Tx) and receive (Rx) activity and allows/denies the channel access of 802.11 part depending on the 802.16 schedule.

Two principally different options exist for the MC module implementation within a MR station (see Fig. 4). One of them uses one reconfigurable antenna [23], which becomes *shared* in terms of the channel access. Clearly, this design prohibits the simultaneous operation of two standards (see Table 1). Another possibility is to use two *separate* antennas: one for each of the cooperating standards. As discussed above, the simultaneous Tx-Rx and Rx-Tx operations should be excluded to avoid radio-to-radio interference.

Table 1. MR station technical limitations

IEEE 802.11-802.16	Shared antenna	Separate antennas
Rx-Rx	Denied	Allowed
Rx-Tx	Denied	Denied
Tx-Rx	Denied	Denied
Tx-Tx	Denied	Allowed

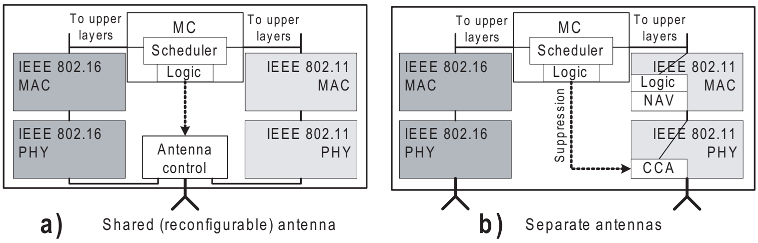


Fig. 4. MR station structure with: a – shared antenna; b – separate antennas

Coordination algorithms restrict the operation of a MR station such that its technical limitations (see Table 1) are accounted for. For the sake of clarity, the below coordination algorithms are demonstrated for the case of only uplink traffic in both networks, that is, 802.11 and 802.16 transmit useful data, while receive only service messages.

3.2 Basic coordination algorithm

Here we present the simplest coordination algorithm, which is referred to as *Basic* in what follows (see Algorithm 1). This algorithm operates under both shared and separate antennas technical limitations (see Table 1) and its main idea is to allow 802.11 utilize only the gaps in 802.16 activity (see Fig. 5).

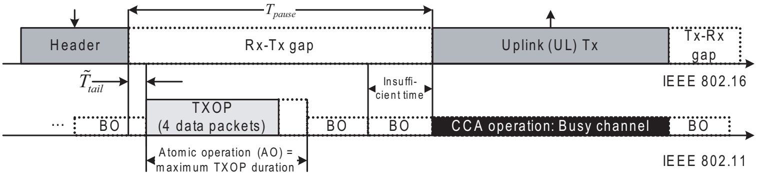


Fig. 5. Basic coordination algorithm operation

Considering the operation of any coordination algorithm we introduce the notion of an *atomic operation* (AO). The AO may be defined as a time interval for a MR station frame transaction such that IEEE 802.11 TXOP the station obtains does not exceed the AO. Thus, AO is the time unit of a MC module and may potentially vary during the coordination algorithm operation. The simplest Basic algorithm, however, utilizes a static AO, which may be reasonably set to the maximum TXOP duration. Therefore, as actual TXOP duration is always less than its maximum, some operation time is unavoidably wasted. This necessarily leads to the less effective performance.

```
1:  Call CCA PHY layer function.
2:  if CCA indicates busy channel then
3:          Go to step 1.
4:  Call NAV MAC layer function.
5:  if NAV indicates busy channel then
6:          Go to step 1.
7:  Obtain current parameters of backoff procedure.
8:  if backoff interval is not over then
9:          Go to step 1.
10: Set maximum IEEE 802.11 TXOP duration $T_{mTXOP}$ as atomic operation duration.
11: Call MC module parametrized by atomic operation duration.
12: Calculate remaining time until closest forthcoming IEEE 802.16 activity.
13: if duration of remaining interval is not less than atomic operation duration then
14:         Send pending data packets during time interval, not exceeding atomic
            operation duration.
15: Begin new backoff interval with minimum contention window value $W_{min}$.
16: Go to step 1.
```

Algorithm 1: Basic coordination algorithm

The implementation of the Basic algorithm is straightforward and involves an additional function call to the MC module. More specifically, once both CCA and NAV of 802.11 indicate that the channel is idle, AIFS interval duration (T_{AIFS}) is spent and backoff time left is 0 the MAC layer requests the time necessary for performing the AO from the MC module. Analyzing the 802.16 schedule, the MC module decides whether there is enough time remaining before the forthcoming 802.16 activity. Further, 802.11 MAC either sends pending TXOP immediately, or initiates a new random backoff with the minimum value of the contention window.

We may therefore formulate the following properties of the Basic coordination algorithm:
+ Simple implementation.
+ Workability in case of both shared and separate antennas.
− Constant atomic operation, resource waste.
− Usage of activity gaps only, non-maximum performance.

3.3 Enhanced coordination algorithm

In order to improve the performance of the Basic coordination algorithm, the *Enhanced* algorithm may be introduced (see Algorithm 2). Its idea is similar to the coexistence-aware TXOP adaptation approach from [16]. It also may operate under both types of technical limitations (see Table 1) and utilizes only gaps in 802.16 activity. However, the Enhanced algorithm varies its atomic operation to adjust to the remaining operation time (see Fig. 6).

The Enhanced coordination algorithm is more complex than its Basic version. In particular, MC module performs more intensive computations of the actual TXOP duration T_{TXOP} with K packets.

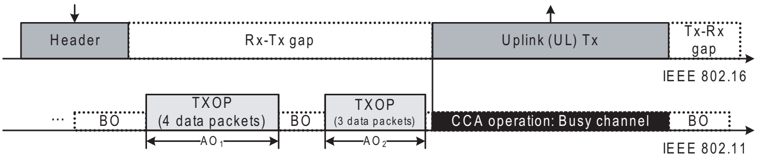


Fig. 6. Enhanced coordination algorithm operation

We may therefore formulate the following properties of the Enhanced coordination algorithm:
+ Dynamic atomic operation, enhanced performance.
+ Workability in case of both shared and separate antennas.
− Higher computational and implementation complexity.
− Usage of activity gaps only, non-maximum performance.

3.4 Suppressing enhanced coordination algorithm

We emphasize, that the previously discussed coordination algorithms utilize only the gaps in 802.16 activity. By relaxing this restriction, higher performance could be achieved. However, enabling simultaneous Tx-Tx and Rx-Rx operation is only possible under the separate antennas technical limitations (see Table 1). We refer to the corresponding coordination algorithm as to *Suppressing* enhanced algorithm in what follows (see Fig. 7 and Algorithm 3).

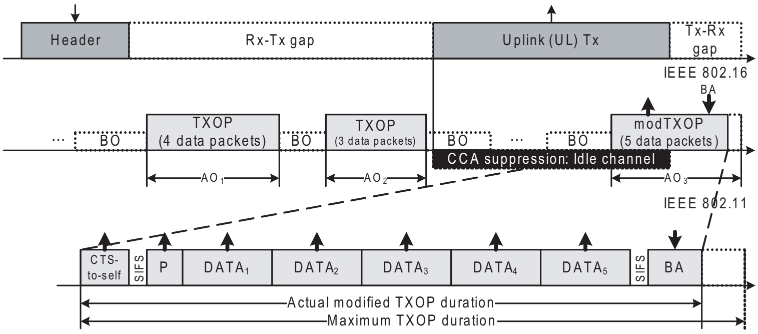


Fig. 7. Suppressing enhanced coordination algorithm operation

We assume, that the channel is sensed busy by 802.11 CCA function of a MR station during any 802.16 Tx activity. We propose the temporary suppression of the CCA signal to enable the simultaneous Tx-Tx operation. This step, however,

```
 1: Call CCA PHY layer function.
 2: if CCA indicates busy channel then
 3:         Go to step 1.
 4: Call NAV MAC layer function.
 5: if NAV indicates busy channel then
 6:         Go to step 1.
 7: Obtain current parameters of backoff procedure.
 8: if backoff interval is not over then
 9:         Go to step 1.
10: Calculate maximum number of data packets $K$ within maximum IEEE 802.11
    TXOP duration $T_{mTXOP}$.
11: while $K > 0$ do
12:         Set actual IEEE 802.11 TXOP duration, which contains exactly $K$ data
            packets, as atomic operation duration.
13:         Call MC module parametrized by atomic operation duration.
14:         Calculate remaining time until closest forthcoming IEEE 802.16 activity.
15:         if duration of remaining interval is not less than atomic operation duration then
16:                 Send $K$ pending data packets.
17:                 $K = 0$.
18:         else
19:                 $K = K - 1$.
20: Begin new backoff interval with minimum contention window value $W_{min}$.
21: Go to step 1.
```

Algorithm 2: Enhanced coordination algorithm

may decrease the robustness of 802.11 busy channel detection mechanism and may lead to the increase in the number of 802.11 collisions.

The Suppressing enhanced algorithm may be regarded as an extension of the Enhanced algorithm for separate antennas (see Fig. 7). Thus, its operation during 802.16 activity gaps remains unchanged. In order to enable the simultaneous operation, the TXOP start time should be scheduled in a way that its Tx part coincides with that of 802.16 (or a gap) and its Rx part – with that of 802.16 (or a gap). Otherwise, according to Table 1 a violation of the technical limitations occurs.

A typical TXOP contains the transmission of RTS, DATA and, optionally, CFE frames, together with the reception of CTS and BA frames (see Fig. 1). In order to simplify Tx/Rx TXOP separation we use a modified TXOP, which consists of CTS-to-self and DATA frames in the Tx part and BA frame in the Rx part (see Fig. 7). Clearly, the complexity of the Suppressing enhanced coordination algorithm is only slightly higher than that of the Enhanced algorithm.

We may therefore formulate the following properties of the Suppressing enhanced coordination algorithm:
+ Simultaneous operation in both networks, better resource utilization.
+ Dynamic atomic operation, enhanced performance.

```
Require: IEEE 802.11 CCA signal suppression is enabled during any ongoing Tx via
    IEEE 802.16.
Ensure: Steps 11 and 13 account for IEEE 802.11 TXOP structure change depending
    on whether CCA signal suppression is enabled.
 1: if CCA signal suppression is *disabled* then
 2:     Call CCA PHY layer function.
 3:     if CCA indicates busy channel then
 4:         Go to step 1.
 5: Call NAV MAC layer function.
 6: if NAV indicates busy channel then
 7:     Go to step 1.
 8: Obtain current parameters of backoff procedure.
 9: if backoff interval is not over then
10:     Go to step 1.
11: Calculate maximum number of data packets $K$ within maximum IEEE 802.11
    TXOP duration $T_{mTXOP}$.
12: while $K > 0$ do
13:     Set actual IEEE 802.11 TXOP duration, which contains exactly $K$ data
        packets, as atomic operation duration.
14:     Call MC module parametrized by atomic operation duration.
15:     Calculate remaining time until closest forthcoming IEEE 802.16 activity.
16:     if duration of remaining interval is *not less* than atomic operation duration then
17:         Send $K$ pending data packets.
18:         $K = 0$.
19:     else
20:         $K = K - 1$.
21: Begin new backoff interval with minimum contention window value $W_{min}$.
22: Go to step 1.
```

Algorithm 3: Suppressing enhanced coordination algorithm

− Workability in case of separate antennas only.
− CCA signal suppression, highest computational and implementation complexity.

4 Performance analysis of coordination algorithms

4.1 System model description

Here we introduce a set of the simplifying restrictions in order to enable the further performance analysis of the described coordination algorithms.
Restriction 1 IEEE 802.16 transmission schedule remains unchanged during the system operation.

Remember, that the MC module operation does not influence the schedule of IEEE 802.16 directly. We, therefore, define the MAC *goodput* of a MR station as the portion of 802.11 PHY layer data rate available for the data transmission at the MAC layer.

Restriction 2 There is only one client station in the system, which is the MR station. It transmits useful data in both 802.11 and 802.16 networks and receives service messages.

Restriction 3 As 802.16 part of the MR station operates in the OFDMA mode, it is assumed to transmit without interruption for the entire UL sub-frame duration, whereas there is no activity in the DL sub-frame except for the header reception.

Restriction 4 802.11 part of the MR station transmits constant-size data packets and is observed under *saturation conditions* [19], that is, it always has a packet ready for transmission.

Restriction 5 The communications channel is noise-free and since no other 802.11 station is present in the system, the MR station always initiates backoff procedure with the minimum contention window size of W_{min} (see description of IEEE 802.11 standard above).

Clearly, MAC goodput under the introduced restrictions is the achievable maximum. For convenience we summarize the principal performance analysis parameters in Table 2[4].

4.2 Single TXOP per frame case

Consider the behavior of the Basic coordination algorithm. Practically, the number of 802.11 TXOPs a MR station obtains per 802.16 frame varies due to the random backoff time. At the same time backoff interval is on average sufficiently shorter than the TXOP duration. One may show that the difference between the maximum number of TXOPs per frame and the respective minimum number is not more than 1. Here we concentrate on the case, when either 0 or 1 TXOP is possible per 802.16 Rx-Tx gap and derive the corresponding goodput value (G_1^B).

Proposition 1. *MAC goodput of the Basic coordination algorithm in case of single TXOP per frame G_1^B may be calculated as:*

$$G_1^B = \frac{LQ_{max}}{T_{frame}} \cdot \Pr\{\tilde{T}_{tail} \leq T\}, \tag{1}$$

where L is the 802.11 data packet length; Q_{max} is the maximum number of packets within 802.11 TXOP; T_{frame} is the 802.16 frame duration.

Proof. In Fig. 5 we observe that an MC module reservation is only possible, when after the first *tagged* backoff in the Rx-Tx gap the remaining time is not less than the maximum TXOP duration (T_{mTXOP}). However, of the tagged backoff

[4] As standard abbreviation (AIFS, TXOP, BO, etc.) is used as lower index of the considered variables, we capitalize the letters and mark random variables with a tilde.

Table 2. Principal performance analysis parameters

Parameter	Description
T_{frame}	IEEE 802.16 frame duration (see Fig. 2)
T_{pause}	Rx-Tx gap duration in IEEE 802.16 schedule (see Fig. 5)
T_{slot}	IEEE 802.11 slot duration (see Fig. 1)
T_{AIFS}	IEEE 802.11 arbitration inter-frame space (AIFS) duration (see Fig. 1)
T_{TXOP}	Actual IEEE 802.11 transmission opportunity (TXOP) duration with maximum number of packets (see Fig. 6)
T_{mTXOP}	Maximum IEEE 802.11 TXOP duration (see Fig. 5)
W_{min}	Minimum contention window value
W_{max}	Maximum contention window value
Q_{max}	Maximum number of packets within IEEE 802.11 TXOP (see Fig. 5)
Q_{mod}	Maximum number of packets within modified IEEE 802.11 TXOP (see Fig. 7)
L	IEEE 802.11 data packet length
$\tilde{T}_{tail}$	IEEE 802.11 tagged backoff interval duration that avoids coincidence with IEEE 802.16 header (see Fig. 5)
$\tilde{T}_{BO}$	IEEE 802.11 backoff interval duration. (see Fig. 1)
$\tilde{W}_{mark}$	Number of slots in tagged backoff interval
$\tilde{Q}_{last}$	Number of packets within last IEEE 802.11 TXOP per IEEE 802.16 frame (see Fig. 6)

interval only the remainder should be accounted for, that does not coincide with the header of 802.16 (denoted by $\tilde{T}_{tail}$ in Fig. 5).

Generally, the backoff time ($\tilde{T}_{BO}$) is a concatenation of a deterministic AIFS interval and a random number of slots, that is, $\tilde{T}_{BO} = T_{AIFS} + \tilde{W}T_{slot}$, where $\tilde{W} \in \{0, 1, \dots, W_{min}\}$. We firstly compute the probability that the number of slots in the tagged backoff ($\tilde{W}_{mark}$) equals to the exact value of j ($\Pr\{\tilde{W}_{mark} = j\}$). We introduce several assumptions that allow for the further simplification of the analysis.

Assumption 1 Assume that the number of consecutive backoff intervals before the tagged one is sufficiently large and regard it as an infinite sequence of the backoff intervals one of which is tagged randomly.

Therefore, accounting for the regeneration properties [24] of the backoff process we obtain the following expression for the sought probability $\Pr\{\tilde{W}_{mark} = j\}$ as:

$$\Pr\{\tilde{W}_{mark} = j\} = \frac{T_{AIFS} + jT_{slot}}{\sum\limits_{i=0}^{W_{min}} T_{AIFS} + iT_{slot}}, \qquad (2)$$

where $j \in \{0, 1, \dots, W_{min}\}$.

Assumption 2 Assume that the interval $\tilde{T}_{tail}$ is discrete. As a discretization unit we select the interval of 1 μs duration, which divides all the standardized intervals (T_{slot}, T_{AIFS}, T_{mTXOP}, etc.).
Assumption 3 Assume that the starting point of the $\tilde{T}_{tail}$ interval is randomly placed at the tagged backoff interval according to the uniform distribution.

Therefore,

$$\Pr\{\tilde{T}_{tail} = i | \tilde{W}_{mark} = j\} = \\ = \begin{cases} (T_{AIFS} + jT_{slot})^{-1}, & \text{if } i \in \{1, 2, \ldots, T_{AIFS} + jT_{slot}\}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

By averaging over the possible values of j we obtain the respective unconditional probability as:

$$\Pr\{\tilde{T}_{tail} = i\} = \sum_{j=0}^{W_{min}} \Pr\{\tilde{T}_{tail} = i | \tilde{W}_{mark} = j\} \cdot \Pr\{\tilde{W}_{mark} = j\}, \tag{4}$$

where $i \in \{1, 2, \ldots, T_{AIFS} + W_{min}T_{slot}\}$.

Let T present the threshold value of the backoff interval remainder duration $\tilde{T}_{tail}$ that still results in one TXOP per frame. This value is given by:

$$T = T_{pause} - T_{mTXOP}, \tag{5}$$

where T_{pause} is the Rx-Tx gap duration in 802.16 schedule (see Fig. 5). Then, the probability that $\tilde{T}_{tail}$ does not exceed T is readily obtained as:

$$\Pr\{\tilde{T}_{tail} \le T\} = \begin{cases} 0, & \text{if } T < 1, \\ 1, & \text{if } T > T_{AIFS} + W_{min}T_{slot}, \\ \sum\limits_{i=1}^{T} \Pr\{\tilde{T}_{tail} = i\}, & \text{otherwise,} \end{cases} \tag{6}$$

which immediately implies (1). ■

4.3 Several TXOPs per frame case

Here we concentrate on the more general case, when the minimum number of TXOPs per frame is k and the maximum number is $k + 1$. Due to the space limitations we consider only the value of $k = 1$ below. The calculations for any natural value $k > 1$ are made similarly.

Proposition 2. *MAC goodput of the Basic coordination algorithm in case of not more than two TXOPs per frame G_2^B may be calculated as:*

$$G_2^B = \frac{LQ_{max}}{T_{frame}} \cdot (1 + \Pr\{\tilde{T}_{tail} + \tilde{T}_{BO} \le T\}). \tag{7}$$

Proof. As before, we derive the threshold value of the random backoff interval duration (T) that now results in two TXOPs per frame. However, this time the random backoff comprises two intervals: the remainder $\tilde{T}_{tail}$ and the full backoff interval $\tilde{T}_{BO}$ between two consecutive TXOPs. The indicated threshold is thus equal to:

$$T = T_{pause} - (T_{TXOP} + T_{mTXOP}), \tag{8}$$

where T_{TXOP} is the actual 802.11 TXOP duration with maximum number of packets. Further, we calculate the probability, that $\tilde{T}_{BO}$ is equal to the exact value of i:

$$\Pr\{\tilde{T}_{BO} = i\} = \begin{cases} (W_{min}+1)^{-1}, & \text{if } i = T_{AIFS} + jT_{slot}, \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where $j \in \{0, 1, \ldots, W_{min}\}$.

The probability that the sum of $\tilde{T}_{tail}$ and $\tilde{T}_{BO}$ is equal to some exact value of j may now be computed as a convolution of the distributions (4) and (9) (see Fig. 8):

$$\Pr\{\tilde{T}_{tail} + \tilde{T}_{BO} = j\} = \sum_{i=1}^{j} \Pr\{\tilde{T}_{tail} = i\} \cdot \Pr\{\tilde{T}_{BO} = j - i\}, \tag{10}$$

where $j \in \{2, 3, \ldots, 2 \cdot (T_{AIFS} + W_{min}T_{slot})\}$.

The value of $\Pr\{\tilde{T}_{tail} + \tilde{T}_{BO} \leq T\}$ is obtained similarly to (6) and may be used to derive the final expression (7). ■

4.4 Enhanced coordination algorithm

MAC goodput of the Enhanced algorithm may be established after an extension of the above approach. Again, a general problem may be formulated for minimum k and maximum $k+1$ number of TXOPs, which we solve below for $k = 1$.

Proposition 3. *MAC goodput of the Enhanced coordination algorithm in case of not more than two TXOPs per frame G_2^E may be calculated as:*

$$G_2^E = \frac{L}{T_{frame}} \cdot (Q_{max} + E[\tilde{Q}_{last}]), \tag{11}$$

where $\tilde{Q}_{last}$ is the number of packets within the last 802.11 TXOP per 802.16 frame.

Proof. We notice, that under the saturation conditions only the duration of the last 802.11 TXOP of those obtained per 802.16 frame may vary, subject to the remaining time in the Rx-Tx gap. We firstly compute a set of thresholds $T(i)$ that result in obtaining the second TXOP, containing exactly i data packets:

$$T(i) = T_{pause} - (T_{TXOP} + T_{TXOP}(i)), \tag{12}$$

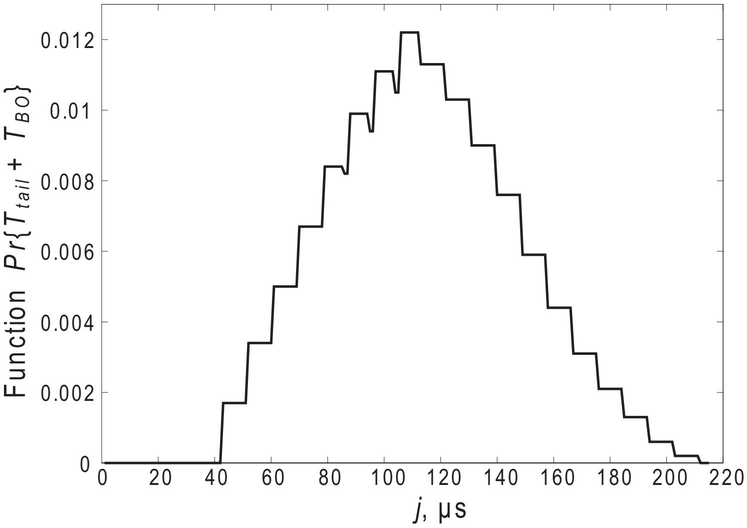


Fig. 8. Example probability function from equation (10): $T_{AIFS} = 43\ \mu s$, $W_{min} = 7$ and $T_{slot} = 9\ \mu s$

where $T_{TXOP}(i)$ is the actual 802.11 TXOP duration, which contains exactly i packets. Once the thresholds are computed, we consider the event E_i that a TXOP contains i packets, conditioning on the fact that $i+1$ packets may not be transmitted. Further, we establish the probabilities $\Pr\{E_i\}$ using (10), (6) and the corresponding thresholds $T(i)$. Denote the random number of packets in the last TXOP by $\tilde{Q}_{last}$. The respective mean value is thus given by:

$$E[\tilde{Q}_{last}] = \sum_{i=1}^{Q_{max}} i \cdot \Pr\{E_i\}, \tag{13}$$

which, in turn, results in (11). ■

4.5 Suppressing enhanced coordination algorithm

Proposition 4. *MAC goodput of the Suppressing enhanced coordination algorithm in case of not more than three TXOPs per frame G_3^S may be calculated as:*

$$G_3^S = \frac{L}{T_{frame}} \cdot (Q_{max} + E[\tilde{Q}_{last}] + Q_{mod}) = G_2^E + \frac{LQ_{mod}}{T_{frame}}, \tag{14}$$

where Q_{mod} is the maximum number of packets within the modified 802.11 TXOP.

Proof. In order to derive the MAC goodput of the Suppressing enhanced algorithm, we should add to the Enhanced algorithm MAC goodput the term corresponding to one modified TXOP per frame (see Fig. 7). As the system operates in the saturation conditions, the modified TXOP contains the maximum number of packets (Q_{mod}), which immediately yields (14). ■

5 Numerical results

5.1 Saturation scenario summary

In order to verify the assumptions of the above coordination algorithms performance analysis, an event-driven simulator was developed, that accounts for the necessary details of the considered system model. The simulator is based on the notorious OPNET Modeler [25] and extends its functionality to the MAC coordination purposes. In particular, to saturate the IEEE 802.16e UL sub-frame, the constant DVD flow of 9.8 Mbps is transmitted. IEEE 802.11n+e part of the MR station also transmits data packets and is observed under the saturation conditions. Each simulation run lasts for 10 s, while the principal simulation parameters are summarized in Table 3.

Table 3. Principal simulation parameters

IEEE 802.16 parameter	Value
DL:UL ratio	60:40
PHY type	OFDMA
Frame duration (T_{frame})	5 ms
Rx-Tx gap duration (T_{pause})	2.5 ms
IEEE 802.11 parameter	**Value**
Maximum IEEE 802.11 TXOP duration (T_{mTXOP})	1.3 ms
Contention window values: W_{min}/W_{max}	7/15
Arbitration inter-frame space (AIFS) duration (T_{AIFS})	43 μs
Slot duration (T_{slot})	9 μs
Data packet length (L)	12 000 bits

5.2 Algorithms saturation performance comparison

We plot both analytical and simulated MAC saturation goodputs for the available set of PHY data rates in Fig. 9. Lines demonstrate the obtained analytical results, while dots represent simulation results. Firstly, we observe that the introduced theoretical approach shows very good accordance with the simulation.

Notice also, that, as expected, the MAC goodput of the Basic coordination algorithm is the lowest comparing with the other algorithms, mainly due

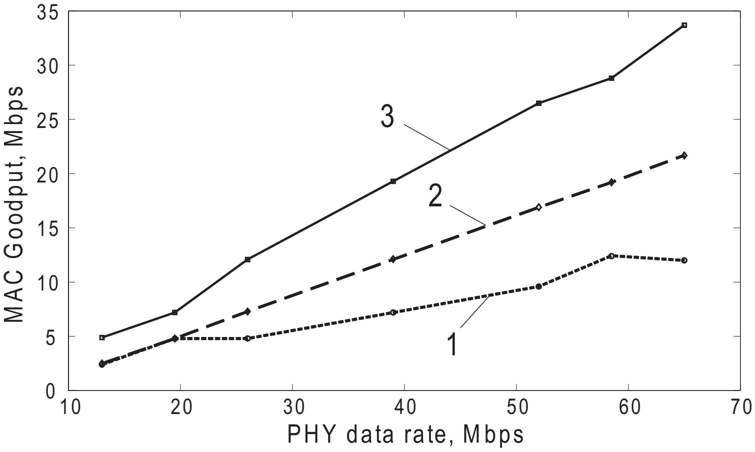


Fig. 9. Coordination algorithms performance comparison: 1 – Basic algorithm; 2 – Enhanced algorithm; 3 – Suppressing enhanced algorithm

to its simplicity. The function for the Enhanced coordination algorithm is almost linear, which is explained by the fact that the dynamic 802.11 TXOP size makes it independent of the variable system parameters. Finally, the Suppressing enhanced coordination algorithm outperforms its competitors for the cost of a more difficult implementation. Additionally, we observe that its effectiveness grows with the increasing data rate, as more data packets fit the additional TXOP per frame.

5.3 Number of MR stations analysis

We continue with the analysis of the presented algorithms with respect to the coexistence issues between MR stations. Remember, that IEEE 802.16 is a schedule-based protocol. Therefore, all the MR stations in the system share the UL sub-frame. As no DL activity is assumed, the 802.16 behavior remains unchanged with the increase in the number of MR stations. By contrast, the MR stations contend for the shared IEEE 802.11 channel. Whenever two or more stations start their transmissions simultaneously, a collision occurs that degrades 802.11 performance. Clearly, with the increasing number of MR stations the collision probability also grows. Consequently, the overall saturation goodput of the 802.11 network drops. The simulation analysis of the indicated problem is shown in Fig. 10 for the fixed PHY data rate of 52 Mbps.

We emphasize the fact that for the Basic coordination algorithm the saturation goodput degradation is almost negligible when the number of MR stations is sufficiently small. This is due to the simplified operation of the Basic algorithm,

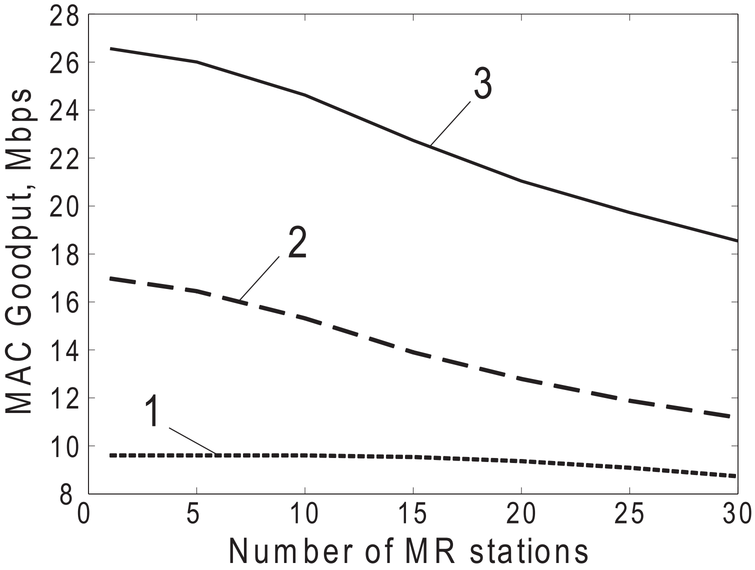


Fig. 10. Saturation goodput vs. number of MR stations for: 1 – Basic algorithm; 2 – Enhanced algorithm; 3 – Suppressing enhanced algorithm

which leaves extra gaps before the forthcoming 802.16 activity. Packet collisions that are short due to the RTS-CTS mechanism do not change this behavior much. By contrast, Enhanced and Suppressing enhanced algorithms utilize the available 802.16 gaps better due to the dynamic atomic operation. Therefore, their performance is more vulnerable to the number of collisions.

5.4 Mean delay analysis

Even though the saturation goodput is the main performance metric of a wireless network, data packet delay analysis is also important to ensure the client QoS requirements are satisfied. For this purpose we extend our simulator with the capability of changing the arrival flow of new packets into the client queue. We compare the delay behavior of the presented coordination algorithms in Fig. 11 for 10 MR station in the system.

Fig. 11 indicates the clear superiority of the Suppressing enhanced coordination algorithm, for which the overall critical arrival rate was established to be 24.5 Mbps in Fig. 10. The Enhanced algorithm is the second best and saturates the system at about 15.5 Mbps. Finally, the Basic algorithm shows the worst mean delay performance having the critical rate of less than 10 Mbps.

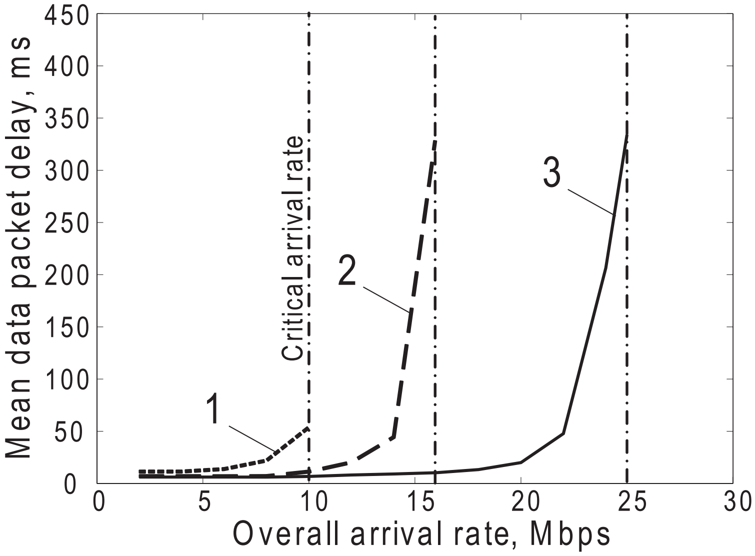


Fig. 11. Mean data packet delay vs. overall arrival rate for: 1 – Basic algorithm; 2 – Enhanced algorithm; 3 – Suppressing enhanced algorithm

6 Conclusion

We presented an approach to enable the simultaneous operation of IEEE 802.11 and IEEE 802.16 telecommunication standards within a multi-radio client station. The MAC coordination concept was introduced and three various coordination algorithms were discussed that demonstrate the performance-complexity trade-off. In particular, we developed a novel Suppressing enhanced coordination algorithm, which increases the MAC goodput of a MR station for more than 50% in comparison to the other algorithms. A simple analytical approach to the performance evaluation of the coordination algorithms was proposed.

The analysis regarding the coexistence with other MR client stations was also performed, as well as the mean data packet delay evaluation. The performance of the considered coordination algorithms was estimated analytically in the framework of the simplified system model, which could be extended further to account for the imperfect channel conditions. It may be shown, that in the noisy channel an appropriate rate adaptation strategy sufficiently improves network performance. The development of coexistence-aware rate adaptation algorithms is thus the prominent research direction.

References

1. *IEEE Std 802.11-2007, New York, USA, June, 2007.*
2. P. Bahl, A. Adya, J. Padhye, and A. Walman, "Reconsidering wireless systems with multiple radios," in *Proc. of the ACM SIGCOMM Computer Communication Review*, vol. 34, pp. 39–46, 2004.
3. A. Adya, P. Bahl, J. Padhye, A. Wolman, and L. Zhou, "A multi-radio unification protocol for IEEE 802.11 wireless networks," in *Proc. of the 1st International Conference on Broadband Networks*, pp. 344–354, 2004.
4. R. Draves, J. Padhye, and B. Zill, "Routing in multi-radio, multi-hop wireless mesh networks," in *Proc. of the 10th Annual International Conference on Mobile Computing and Networking*, pp. 114–128, 2004.
5. M. Alicherry, R. Bhatia, and L. Li, "Joint channel assignment and routing for throughput optimization in multi-radio wireless mesh networks," in *Proc. of the 11th Annual International Conference on Mobile Computing and Networking*, pp. 58–72, 2005.
6. A. Mishra, E. Rozner, S. Banerjee, and W. Arbaugh, "Exploiting partially overlapping channels in wireless networks: Turning a peril into an advantage," in *Proc. of the Internet Measurement Conference*, pp. 311–316, 2005.
7. K. Ramachandran, E. Belding, K. Almeroth, and M. Buddhikot, "Interference-aware channel assignment in multi-radio wireless mesh networks," in *Proc. of the 25th IEEE International Conference on Computer Communications*, 2006.
8. B.-J. Ko, V. Misra, J. Padhye, and D. Rubenstein, "Distributed channel assignment in multi-radio 802.11 mesh networks," in *Proc. of the IEEE Wireless Communications and Networking Conference*, pp. 3978–3983, 2007.
9. *IEEE Std 802.16e-2005, New York, USA, February 2006.*
10. X. Jing, S.-C. Mau, D. Raychaudhuri, and R. Matyas, "Reactive cognitive radio algorithms for co-existence between IEEE 802.11b and 802.16a networks," in *Proc. of the 48th IEEE Global Telecommunications Conference*, vol. 5, 2005.
11. X. Jing and D. Raychaudhuri, "Spectrum co-existence of IEEE 802.11b and 802.16a networks using the CSCC etiquette protocol," in *Proc. of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pp. 243–250, 2005.
12. P. Djukic and S. Valaee, "802.16 MCF for 802.11a based mesh networks: A case for standards re-use," in *Proc. of the 23rd Biennial Symposium on Communications*, pp. 186–189, 2006.
13. S. Mangold, *Analysis of IEEE 802.11e and application of game models for support of quality-of-service in coexisting wireless networks.* PhD thesis, RWTH Aachen University, 2003.
14. L. Berlemann, C. Hoymann, G. Hiertz, and S. Mangold, "Coexistence and interworking of IEEE 802.16 and IEEE 802.11(e)," in *Proc. of the 63rd IEEE Vehicular Technology Conference*, vol. 1, pp. 27–31, 2006.
15. L. Berlemann, C. Hoymann, G. Hiertz, and B. Walke, "Unlicensed operation of IEEE 802.16: Coexistence with 802.11(a) in shared frequency bands," in *Proc. of the 17th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, 2006.
16. J. Zhu, A. Waltho, X. Yang, and X. Guo, "Multi-radio coexistence: Challenges and opportunities," in *Proc. of the 16th International Conference on Computer Communications and Networks*, pp. 358–364, 2007.

17. A. Kamerman, “Coexistence between Bluetooth and IEEE 802.11 CCK solutions to avoid mutual interference,” tech. rep., Lucent Technologies Bell Laboratories (IEEE 802.11-00/162), 1999/2000.
18. F. Wang, A. Nallanathan, and H. Garg, “Introducing packet segmentation for the IEEE 802.11b throughput enhancement in the presence of Bluetooth,” in *Proc. of the 59th IEEE Vehicular Technology Conference*, vol. 4, pp. 2252–2256, 2004.
19. G. Bianchi, “Performance analysis of the IEEE 802.11 distributed coordination function,” *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.
20. B. Walke, S. Mangold, and L. Berlemann, *IEEE 802 Wireless Systems: Protocols, Multi-Hop Mesh/Relaying, Performance and Spectrum Coexistence.* John Wiley & Sons, 2007.
21. *J. Zhu et al., "IEEE 802 Air-Interface Support for Co-Located Coexistence", IEEE 802.19-08/0021r2, July 2008.*
22. *IEEE 802 Plenary Tutorial on WPAN/WLAN/WWAN Multi-Radio Coexistence, November, 2007.*
23. C. Zhang, S. Yang, H. Pan, A. Fathy, S. El-Ghazaly, and V. Nair, “Reconfigurable antenna for simultaneous multi-service wireless applications,” in *Proc. of the IEEE Radio and Wireless Symposium*, pp. 543–546, 2007.
24. L. Kleinrock, *Queueing Systems Volume I: Theory.* New York, 1975.
25. *OPNET Modeler. Discrete Event Simulation API Reference Manual.*

Publication 6

S. Andreev, O. Galinina, and A. Vinel, “Performance evaluation of a three node client relay system,” *International Journal of Wireless Networks and Broadband Technologies (IJWNBT)*, vol. 1, no. 1, pp. 73–84, 2011.

Performance Evaluation of a Three Node Client Relay System

Sergey Andreev, Tampere University of Technology, Finland

Olga Galinina, Tampere University of Technology, Finland

Alexey Vinel, Tampere University of Technology, Finland

ABSTRACT

In this paper, the authors examine a client relay system comprising three wireless nodes. Closed-form expressions for mean packet delay, as well as for throughput, energy expenditure, and energy efficiency of the source nodes are also obtained. The precision of the established parameters is verified by means of simulation.

Keywords: *Cellular Network, Client Relay, Nodes, Performance Evaluation, Queueing System*

INTRODUCTION

Wireless communication networks are becoming more widespread, as novel telecommunication standards emerge (Marks, Nikolich, & Snyder, in press; Nakamura, in press). The future of wireless communication, however, greatly depends on how successfully the disproportion between the required Quality of Service (QoS) and the limited system spectral resource is overcome. Meanwhile, the urge to increase system *spectral efficiency* gradually gives way to the task of the *energy efficiency* improvement. This is particularly true for small-scale handheld wireless devices due to the growing gap between available and required battery capacity (Lahiri, Raghunathan, Dey, & Panigrahi, 2002).

The problem of effective resource utilization is of primary importance for wireless systems, where a large population of users' shares limited spectral resource (Andreev, Koucheryavy, Himayat, Gonchukov, & Turlikov, 2010). Currently, layered system architecture dominates in network design, where each layer is treated independently following the concept of layer abstraction. Among these, *Physical* (PHY) layer is responsible for the transmission of raw data bits, whereas *Media Access Control* (MAC) layer arbitrates access of users to the shared wireless channel.

However, traditional layered architecture appears to be far less flexible and often implies inefficient resource utilization (Andreev, Galinina, & Vinel, 2010). To mitigate this discrepancy, a novel integral and adaptive approach is required. As a consequence, cross-layer techniques receive increasing attention from the research community (Andreev, Koucheryavy, Himayat, Gonchukov, & Turlikov, 2010) with primary focus being set on the joint consider-

DOI: 10.4018/ijwnbt.2011010106

ation of MAC and PHY layers. New channel-aware solutions are introduced to achieve cross-layer benefits by taking advantage of wireless channel state information (CSI). They typically exploit extended MAC-PHY interaction and result in higher QoS, arrival flow, and channel state adaptability (Song, 2005; Miao, 2008; Kim, 2009).

As wireless users are becoming increasingly mobile, the focus of the latest research efforts shifts from throughput optimization (Song & Li, 2006) towards energy efficiency improvement at all layers of a wireless system (Anisimov, Andreev, Galinina, & Turlikov, 2010) from its architecture (Benini, Bogliolo, & de Micheli, 2000) to the adopted communication protocols (Schurgers, 2002). Recently, *cooperative* cross-layer approaches gain increasing international acclaim (Pyattaev, Andreev, Vinel & Sokolov, 2010; Pyattaev, Andreev, Koucheryavy, & Moltchanov, 2010). They exploit variability in CSI of wireless users and, as such, allow for additional performance gains thus constituting a promising research direction.

RESEARCH BACKGROUND

While more and more users are sharing the limited wireless resource and *cellular* networks are gradually shifting towards more aggressive frequency reuse scenarios (Marks, Nikolich, & Snyder, in press; Nakamura, in press), wireless interference becomes one of the major limiting factors that impair network performance growth. Wireless data transmission of a user, being unavoidably broadcast, necessarily impacts the transmission process of other users and consequently degrades the overall system energy efficiency. However, users may gain in their energy efficiencies by acting cooperatively (Cui, Goldsmith, & Bahai, 2004; Jayaweera, 2004). Such a spatial domain resource management is becoming increasingly important to improve the performance of the cell-edge users with a poor communication link (Andreev, Galinina, & Vinel, 2010).

On the other hand, cooperation typically implies extra energy expenditure as more data is transmitted over the air. Moreover, cooperative transmission may negatively impact packet delay, as data packets are sometimes relayed over a longer path. However, increasing delay could sometimes be compensated by reducing transmission data rate; and this is contrastingly known to increase user energy efficiency (Andreev, Koucheryavy, Himayat, Gonchukov, & Turlikov, 2010). As such, it is important to evaluate all the basic trade-offs behind wireless cooperation and indicate scenarios where it actually improves the performance of a cellular network.

Currently, studying the collaboration between *neighboring* users of a wireless system is highly significant. As energy expenditure to guarantee reliable data transmission exponentially grows with distance (Stuber, 2001), it is desirable to relay data over shorter intermediate hops (Rabaey, Ammer, da Silva Jr., & Patel, 2000). Consequently, *client relay* is believed to become a promising concept that would boost the performance of contemporary wireless cellular networks.

Enabling client relay, it is crucial to avoid scenarios when the use of this technology insufficiently increases the performance of the originating user (Haenggi & Puccinelli, 2005). As the result, the task of effective relay selection is often reduced to analyzing the trade-off between the source node benefits and the relay node losses. In this paper, we evaluate the performance of the simplest but nonetheless practical client relay network. We estimate mean packet delay for all the data sources within the considered system, as well as establish their throughput, energy expenditure, and energy efficiency.

SYSTEM MODEL

We consider a wireless cellular network enhanced with client relay capability borrowing the basic methodology from our previous paper

(Andreev, Galinina, & Turlikov, 2010) and extending it. In what follows, we concentrate on the simplest network topology (Figure 1) comprising two source nodes and one sink node. We term user A the *originator*. The originator generates own data packets with the mean arrival rate λ_A. We also term user R the *relay*. The relay generates own data packets with the mean arrival rate λ_R. Additionally, the relay is capable of eavesdropping on the transmissions from the originator and may temporarily store the packets from A for the subsequent retransmission. The node B is termed the *base station* and receives data packets from both the originator and the relay. Below we detail the system model and present the set of main assumptions.

Assumption 1. System time is *slotted*. All the communicated data packets have equal size and the transmission of each one of them takes exactly one slot.

Assumption 2. The numbers of new data packets arriving to either the originator or the relay during consecutive slots are independent and identically distributed (i.i.d) random variables with the means λ_A and λ_R respectively. For the sake of analytical tractability we assume *Poisson* arrival flow of new packets in the rest of the text. The base station has no outgoing traffic.

Assumption 3. Both the originator and the relay have unbounded queues to store own data packets. Additionally, the relay has an extra memory location to keep a *single* data packet from the originator for the subsequent cooperative retransmission. Below we demonstrate that single memory cell is sufficient for the proposed client relay system operation.

Assumption 4. The communication system is centralized and is controlled by the base station. The fair *stochastic* round-robin scheduler operates at the base station and alternates source nodes accessing the wireless channel with equal probability (see Figure 2 as an example behavior of the system detailed in Figure 1, with no additional packet arrivals). In particular, if both the originator and the relay have pending data packets the subsequent slot is given to either of them with probability 0.5. The non-transmitting node stays idle during this slot. If either of the source nodes has no pending data packets the other one is given the subsequent slot with probability 1. This ensures efficient system time utilization. If neither source has pending data packets the entire system stays idle. We also assume that scheduling information about which node transmits in the subsequent slot is immediately available to both source nodes over a separate channel and consumes no system resources.

Assumption 5. The communication channel is error-prone and is based on the multi-packet reception channel model (Rong & Ephremides, 2009). The transmitted data packet is received by the destination successfully with the constant probability dependent only on the link type (direct

Figure 1. Considered three node client relay system

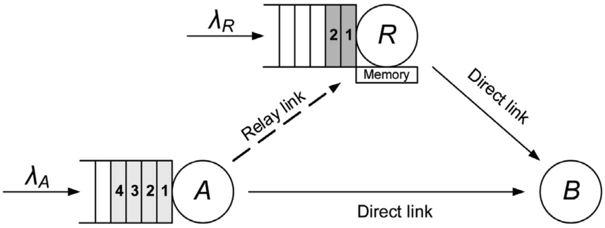
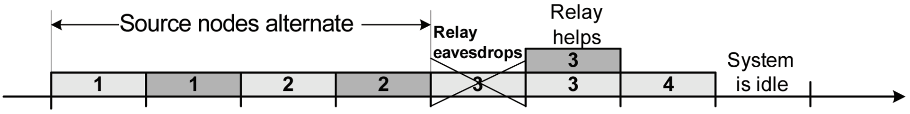


Figure 2. Example client relay system operation



or relay) and on which nodes are transmitting simultaneously.

We define the following non-zero success probabilities:

- $p_{AB} = \Pr\{packet\ from\ A\ is\ received\ at\ B\,|\,only\ A\ transmits\}$
- $p_{RB} = \Pr\{packet\ from\ R\ is\ received\ at\ B\,|\,only\ R\ transmits\}$
- $p_{AR} = \Pr\{packet\ from\ A\ is\ received\ at\ R\,|\,only\ A\ transmits\}$
- $p_{CB} = \Pr\{packet\ from\ A\ is\ received\ at\ B\,|\,A\ and\ R\ cooperate\}$

It is expected that $p_{AR} > p_{AB}$, as well as $p_{CB} > p_{AB}$.

We also assume that feedback information about the success/failure of each reception attempt by the base station is immediately available to both source nodes over a separate channel and consumes no system resources. If the packet is not received successfully, it is retransmitted by its source. The maximum number of allowable retransmission attempts is *unlimited*. The relay is incapable of simultaneous transmission and reception.

Assumption 6. At the *first* packet transmission attempt by the originator, the relay attempts eavesdropping on it with probability 1. According to Assumption 5, this eavesdropping attempt is successful with probability p_{AR}. If the packet is received successfully by the relay it is stored in the single memory location by replacing its previous contents. Also according to Assumption 5, this packet is at the same time successfully received at the base station with probability p_{AB}. If unsuccessful, the originator retransmits the same packet in the next available slot.

Assumption 7. At any *retransmission* attempt by the originator, the relay performs one of the following. If the packet which is being transmitted was already stored in its memory location the relay sends it *simultaneously* with the originator with probability 1 (Figure 2). As such, the relay tries to improve the performance of the originator. According to Assumption 5, this packet is successfully received at the base station with probability p_{CB}. Alternatively, if eavesdropping on the transmitted packet failed previously the relay attempts again with probability 1 (see Assumption 6).

We note that according to Assumptions 6 and 7 a single memory location for the eavesdropped data packets at the relay suffices for the considered client relay system operation. Moreover, the originator is unaware of the cooperative help from the relay and the relay sends no explicit acknowledgements to the originator by contrast to the approach from (Rong & Ephremides, 2009). This enables tailoring the proposed client relay model to the contemporary cellular standards (Marks, Nikolich, & Snyder, in press; Nakamura, in press). The relay improves the throughput of the originator by sacrificing its own energy efficiency. Extra energy is spent by the relay on the eavesdropping, as well as on the simultaneous packet transmissions with the originator.

Generally, the relay may sometimes decide not to eavesdrop on the transmissions from the originator or not to transmit a packet simultaneously subject to a particular client relay policy. In this paper, we restrict our further explorations to the baseline case when the relay is forced to eavesdrop on all the transmissions from the originator and to transmit a packet simultaneously whenever it is kept in the memory (see Assumptions 6 and 7). We leave any *opportunistic* cooperation for the future work.

Main Notations

The proposed analytical approach to the performance evaluation of the considered three node client relay system is based on the notion of the packet service time. The packet *service time* is the time interval from the moment when the tagged packet is ready for service to the moment its service ends (Jaiswal, 1968). More specifically, in the considered system the service time of the tagged packet from a node starts when this packet becomes the first one in the queue of this node and ends when its successful transmission ends.

We denote the service time of a packet from node A as $T_{AR}(\lambda_A, \lambda_R) \triangleq T_{AR}$, where '$\triangleq$' reads as "equal by definition". Additionally, we introduce the *mean* service time of a packet from node A as $\tau_{AR}(\lambda_A, \lambda_R) \triangleq \tau_{AR} = E[T_{AR}]$. Further, we denote by $\tau_{AR}(\lambda_A, 0) \triangleq \tau_{A0}$ the mean service time of a packet from node A conditioning on the fact that $\lambda_R = 0$.

Symmetrically, we denote the service time of a packet from node R as $T_{RA}(\lambda_R, \lambda_A) \triangleq T_{RA}$ and the corresponding mean service time as $\tau_{RA}(\lambda_R, \lambda_A) \triangleq \tau_{RA} = E[T_{RA}]$. Analogously, the conditional mean service time is introduced as $\tau_{RA}(\lambda_R, 0) \triangleq \tau_{R0}$ for $\lambda_A = 0$.

Note that as T_{R0} is distributed geometrically, for both system with cooperation (when $p_{AR} > 0$) and system without cooperation (when $p_{AR} = 0$) it holds the following:

$$\tau_{R0} = \frac{1}{p_{RB}} \tag{1}$$

whereas only for the system without cooperation it holds:

$$\tau_{A0} = \frac{1}{p_{AB}} \tag{2}$$

The derivation of τ_{A0} for the system with cooperation is a more complicated task and will be addressed below.

Denote the numbers of data packets in the queues of the nodes A and R at the beginning of a particular slot t by $Q_A^{(t)}$ and $Q_R^{(t)}$ respectively. As we observe the client relay system in stationary conditions, we omit the upper index t of variables $Q_A^{(t)}$ and $Q_R^{(t)}$.

Finally, we denote the queue *load coefficient* (Kleinrock, 1975) of node A as $\rho_{AR}(\lambda_A, \lambda_R) \triangleq \rho_{AR}$. By definition we have:

$$\rho_{AR} = \Pr\{Q_A \neq 0\} = \lambda_A \tau_{AR} \tag{3}$$

In particular, queue load coefficient of node A conditioning on the fact that $\lambda_R = 0$ may be established as: $\rho_{AR}(\lambda_A, 0) \triangleq \rho_{A0} = \lambda_A \tau_{A0}$. Accounting for (2), for the system without cooperation ρ_{A0} further simplifies to $\rho_{A0} = \dfrac{\lambda_A}{p_{AB}}$.

Analogously, queue load coefficient of node R is denoted as $\rho_{RA}(\lambda_R, \lambda_A) \triangleq \rho_{RA}$. Also by definition we have:

$$\rho_{RA} = \Pr\{Q_R \neq 0\} = \lambda_R \tau_{RA} \tag{4}$$

Symmetrically, queue load coefficient of node R conditioning on the fact that $\lambda_A = 0$ may be established as:

$$\rho_{RA}(\lambda_R, 0) \triangleq \rho_{R0} = \lambda_R \tau_{R0}.$$

Accounting for (1), for both systems with and without cooperation ρ_{R0} further simplifies to

$\rho_{R0} = \dfrac{\lambda_R}{p_{RB}}$. The main notations we consistently use throughout this paper are summarized in Table 1.

General Statements

Consider the queue at node A. We remind that by definition $\rho_{AR} = \Pr\{Q_A \neq 0\}$ and set $\rho_{A0} > \rho_{R0}$ as an example. The following propositions may thus be formulated.

Proposition 1. For the queue load coefficient of node A it holds the following:

$$\rho_{AR} \leq \frac{\rho_{A0}}{1-\rho_{R0}} \tag{5}$$

Another important proposition may be formulated considering normalization condition of the respective system generating function or balance equations of the corresponding embedded Markov chain.

Proposition 2. For the queue load coefficients of nodes A and R it holds the following:

$$\rho_{AR} - \rho_{RA} = \rho_{A0} - \rho_{R0} \tag{6}$$

The proofs of Propositions 1 and 2 are not included into this text due to space constraints.

Proposition 3. For the queue load coefficient of node R it holds the following:

$$\rho_{RA} = \rho_{AR} - \rho_{A0} + \rho_{R0} \leq \frac{\rho_{A0}}{1-\rho_{R0}} - \rho_{A0} + \rho_{R0} \tag{7}$$

The proof of Proposition 3 follows immediately from (5) and (6).

The established upper bounds on ρ_{AR} and ρ_{RA} hold for both systems with and without cooperation. In what follows, we firstly study the system without cooperation and then extend the proposed analytical approach to the system with cooperation.

Non-Cooperative System Performance Evaluation

We study the behavior of node A within the framework of the queueing theory. As such, consider the queueing system associated with node A. Due to the fact that the queues of nodes A and R are mutually dependent, the notorious Pollazek-Khinchine formula (Kleinrock, 1975) may not be used to obtain the *exact* mean queue length of node A. We, however, apply this formula to establish the *approximate* value of the mean queue length of node A as:

$$q_A \cong \lambda_A E[T_{AR}] + \frac{\lambda_A^2 E[T_{AR}^2]}{2(1-\lambda_A E[T_{AR}])} = \lambda_A \tau_{AR} + \frac{\lambda_A^2 E[T_{AR}^2]}{2(1-\lambda_A \tau_{AR})} \tag{8}$$

where $\tau_{AR} = E[T_{AR}]$ is the mean service time of a packet from node A (the first moment of random service time T_{AR}) and $E[T_{AR}^2]$ is the respective second moment of the service time. Accounting for (3) and (8), we may write:

$$q_A \cong \rho_{AR} + \frac{\lambda_A^2 E[T_{AR}^2]}{2(1-\rho_{AR})} \tag{9}$$

We now demonstrate how to derive the unknown components of equation (9). Consider the service time of the tagged packet from node A. We remind that the packet scheduler at the base station is stochastic, that is, it assigns the subsequent slot to node A with probability 0.5 if both source nodes are loaded. Therefore, in every slot for which $Q_R \neq 0$ and $Q_A \neq 0$ the packet from A is included into the system schedule with probability 0.5. We introduce the following auxiliary probability:

$$\gamma_A \triangleq \Pr\{Q_R \neq 0 \mid Q_A \neq 0\} = \frac{\Pr\{Q_R \neq 0, Q_A \neq 0\}}{\Pr\{Q_A \neq 0\}}.$$

Table 1. Main notations

Notation	***Parameter description***
λ_A	Mean arrival rate of packets in node A
λ_R	Mean arrival rate of packets in node R
p_{AB}	Probability of successful reception from A at B when A transmits
p_{RB}	Probability of successful reception from R at B when R transmits
p_{AR}	Probability of successful reception from A at R when A transmits
p_{CB}	Probability of successful reception from A at B when A and R cooperate
τ_{AR}	Mean service time of a packet from node A
τ_{RA}	Mean service time of a packet from node R
ρ_{AR}	Queue load coefficient of node A
ρ_{RA}	Queue load coefficient of node R
q_A	Mean queue length of node A
q_R	Mean queue length of node R
δ_A	Mean packet delay of node A
δ_R	Mean packet delay of node R
η_A	Mean departure rate of packets from node A (throughput of A)
η_R	Mean departure rate of packets from node R (throughput of R)
ε_A	Mean energy expenditure of node A
ε_R	Mean energy expenditure of node R
φ_A	Mean energy efficiency of node A
φ_R	Mean energy efficiency of node R

Clearly, the scheduler either assigns the subsequent slot to node R with probability $0.5\gamma_A$ or assigns it to node A with the complementary probability $1-0.5\gamma_A$.

Consider the probability of the event that $Q_R \neq 0$ and $Q_A \neq 0$ simultaneously. By the complete probability formula, we may write $\Pr\{Q_R \neq 0, Q_A \neq 0\} = \Pr\{Q_R \neq 0\} - \Pr\{Q_R \neq 0, Q_A = 0\}$. On the other hand, by definition we have $\Pr\{Q_R \neq 0\} = \rho_{RA}$. Further, for the probability $\Pr\{Q_R \neq 0, Q_A = 0\}$ we obtain the following expression:

$$\Pr\{Q_R \neq 0, Q_A = 0\} = \Pr\{Q_A = 0\} - \Pr\{Q_R = 0, Q_A = 0\}.$$

Using the definition of ρ_{AR}, we note that $\Pr\{Q_A = 0\} = 1 - \rho_{AR}$. Moreover, we also note that $\Pr\{Q_R = 0, Q_A = 0\} = 1 - \rho_{A0} - \rho_{R0}$.

Summarizing the above, $\Pr\{Q_R \neq 0, Q_A \neq 0\} = \rho_{AR} + \rho_{RA} - \rho_{A0} - \rho_{R0}$. Additionally, from Proposition 2 it immediately follows that:

$$\Pr\{Q_R \neq 0, Q_A \neq 0\} = 2 \cdot (\rho_{AR} - \rho_{A0}).$$

Finally, we obtain:

$$0.5\gamma_A = 0.5 \cdot \frac{\Pr\{Q_R \neq 0, Q_A \neq 0\}}{\Pr\{Q_A \neq 0\}} = 1 - \frac{\rho_{A0}}{\rho_{AR}}.$$

We may establish the following distribution for the service time of a packet from node A:

$$\Pr\{T_{AR} = n\} = p_{AB}(1 - 0.5\gamma_A)(1 - p_{AB}(1 - 0.5\gamma_A))^{n-1}.$$

The above expression accounts for the fact that out of n slots spent to serve a packet from node A the last slot was assigned to node A and its transmission in this slot was successful. The previous $n-1$ slots were either not assigned to node A or its transmissions in these slots were unsuccessful.

Calculating the first and the second moment of the service time ($E[T_{AR}]$ and $E[T_{AR}^2]$), accounting for (9) and also using Little's formula in the form $q_A = \lambda_A \delta_A$, it is now easy to approximate the mean packet delay of node A as:

$$\delta_A \cong \frac{\rho_{AR}}{\lambda_A} + \frac{\lambda_A(2 - p_{AB}(1 - 0.5\gamma_A))}{2(1 - \rho_{AR})p_{AB}^2(1 - 0.5\gamma_A)^2}.$$

The performance metrics of node R may be calculated analogously, due to the symmetric nature of the respective direct links. Accounting for $0.5\gamma_R = 1 - \frac{\rho_{R0}}{\rho_{RA}}$, where $\gamma_R \triangleq \frac{\Pr\{Q_R \neq 0, Q_A \neq 0\}}{\Pr\{Q_R \neq 0\}}$, the approximate mean packet delay of node R is given by:

$$\delta_R \cong \frac{\rho_{RA}}{\lambda_R} + \frac{\lambda_R(2 - p_{RB}(1 - 0.5\gamma_R))}{2(1 - \rho_{RA})p_{RB}^2(1 - 0.5\gamma_R)^2}.$$

The proposed analytical approach to the performance evaluation of the considered three node client relay system is also applicable for establishing the *exact* mean departure rate of packets from (throughput of) nodes A and R. In particular, the throughput of A is given by:

$$\eta_A = \begin{cases} \lambda_A, \ no \ \ saturation \\ \dfrac{1 \ - \ \lambda_R \tau_{R0}}{\tau_{A0}}, \ saturation \ \ for \ \ A \\ \dfrac{1}{2\tau_{A0}}, \ saturation \ \ for \ \ A, R \end{cases}.$$

Similarly, the throughput of R may be derived by:

$$\eta_R = \begin{cases} \lambda_R, \ no \ \ saturation \\ \dfrac{1 \ - \ \lambda_A \tau_{A0}}{\tau_{R0}}, \ saturation \ \ for \ \ R \\ \dfrac{1}{2\tau_{R0}}, \ saturation \ \ for \ \ A, R \end{cases}.$$

The above expressions may be further simplified accounting for equations (1) and (2). Here, the saturation conditions are defined as follows:

- Saturation for A: $\left(\lambda_A \tau_{A0} + \lambda_R \tau_{R0} > 1\right)$ and at the same time $\left(\lambda_R \tau_{R0} < 0.5\right)$.
- Saturation for R: $\left(\lambda_A \tau_{A0} + \lambda_R \tau_{R0} > 1\right)$ and at the same time $\left(\lambda_A \tau_{A0} < 0.5\right)$.
- Saturation for A and R: $\left(\lambda_A \tau_{A0} > 0.5\right)$ and at the same time $\left(\lambda_R \tau_{R0} > 0.5\right)$.

Additionally, we may obtain the exact value of the mean energy expenditure of node A as:

$$\varepsilon_A = P_{TX} \eta_A \tau_{A0} \ + \ P_I (1 \ - \ \eta_A \tau_{A0}),$$

together with the mean energy expenditure of node R as:

$$\varepsilon_R = \ P_{TX} \eta_{\mathrm{R}} \tau_{R0} \ + \ P_I \left(1 \ - \ \eta_{\mathrm{R}} \tau_{R0}\right).$$

Here, P_{TX} is the average power that is spent by a node in the packet transmission state, whereas P_I is the average power that is spent by the same node in the idle state. As such, the mean energy efficiencies of nodes A and R readily follow and are given by expressions $\varphi_A = \dfrac{\eta_A}{\varepsilon_A}$ and $\varphi_R = \dfrac{\eta_R}{\varepsilon_R}$ respectively.

Cooperative System Performance Evaluation

In order to mathematically describe the system with cooperation, we firstly consider an important special case when the queue at node R is always empty. We establish the distribution of the number of slots required to serve a packet from node A. By using the obtained distribution, we then generalize the proposed approach for the case of non-empty queue at node R. All the respective performance metrics for the system with cooperation are marked by symbol '*' in the rest of the text.

Case 1. The queue at node R is always empty ($\lambda_R = 0$).

Analogously to the derivations in the previous section, we may express the sought distribution for the service time of a packet from node A as:

$$\Pr\{T^*_{A0} = n\} \\ = X(1 - p_{CB})^{n-1} - Y[(1 - p_{AB})(1 - p_{AR})]^{n-1},$$

where $X = \dfrac{p_{AR}(1 - p_{AB})p_{CB}}{1 - p_{CB} - (1 - p_{AB})(1 - p_{AR})}$ and $Y = X - p_{AB}$.

Coming now to the mean service time, we have the following:

$$\tau^*_{A0} = \frac{p_{CB} + (1 - p_{AB})p_{AR}}{p_{CB}[p_{AB} + (1 - p_{AB})p_{AR}]} \tag{10}$$

Case 2. The queue at node R is *not* always empty ($\lambda_R > 0$).

Here we generalize the above analytical expressions for the most complex cooperative case with $\lambda_R > 0$. Omitting lengthy but straightforward derivations, we give the respective distribution for the service time of a packet from node A as:

$$\begin{aligned}&\Pr\{T^*_{AR}=n\}\\&=X(1-0.5\gamma^*_A)(1-p_{CB}(1-0.5\gamma^*_A))^{n-1}\\&-Y(1-0.5\gamma^*_A)(1-p_A(1-0.5\gamma^*_A))^{n-1}\quad,\end{aligned}$$

where $0.5\gamma^*_A = 1-\dfrac{\rho^*_{A0}}{\rho^*_{AR}}$ and also $p_A = p_{AB}+p_{AR}-p_{AB}\cdot p_{AR}$ for brevity. Here p_A is the probability to successfully receive a packet from A at either node R or at the base station.

Queue load coefficients of nodes A and R (ρ^*_{AR} and ρ^*_{RA}) may be calculated similarly to the respective parameters for the system without cooperation accounting for the fact that $\rho^*_{A0} \triangleq \lambda_A\tau^*_{A0}$, where the expression for τ^*_{A0} is given by (10).

Finally, calculating the second moment of the service time we derive the resulting expression for the approximate mean packet delay of node A as:

$$\delta^*_A \cong \frac{\rho^*_{AR}}{\lambda_A}+\frac{\lambda_A}{2(1-\rho^*_{AR})(1-0.5\gamma^*_A)^2}\cdot\left[X\cdot\frac{2-p_{CB}(1-0.5\gamma^*_A)}{p^3_{CB}}-Y\cdot\frac{2-p_A(1-0.5\gamma^*_A)}{p^3_A}\right],$$

where X and Y were given above.

Accounting for (10), the resulting approximation for the mean packet delay δ^*_R of node R, as well as expressions for the throughput η^*_A and η^*_R of nodes A and R in the system with cooperation are similar to the respective metrics in the system without cooperation from the previous section. Analogously, the mean energy expenditure of node A in the considered case is given by:

$$\varepsilon^*_A = P_{TX}\eta^*_A\tau^*_{A0}+P_I(1-\eta^*_A\tau^*_{A0}),$$

whereas the mean energy expenditure of node R may be calculated as:

$$\begin{aligned}\varepsilon^*_R = {}& P_{TX}\left(\eta^*_R\tau_{R0}+\eta^*_A\cdot\frac{1-\mathrm{p}_{AB}\tau^*_{A0}}{\mathrm{p}_{CB}-\mathrm{p}_{AB}}\right)\\&+P_{RX}\left(1-\eta^*_R\tau_{R0}-\eta^*_A\cdot\frac{1-\mathrm{p}_{AB}\tau^*_{A0}}{\mathrm{p}_{CB}-\mathrm{p}_{AB}}\right),\end{aligned}$$

where P_{RX} is the average power that is spent by a node in the packet reception state. As before, the mean energy efficiencies of nodes A and R are given by expressions $\varphi^*_A=\dfrac{\eta^*_A}{\varepsilon^*_A}$ and $\varphi^*_R=\dfrac{\eta^*_R}{\varepsilon^*_R}$ respectively.

Figure 3. Dependency of throughput (left) and mean packet delay (right) of source nodes on mean packet arrival rate λ_A

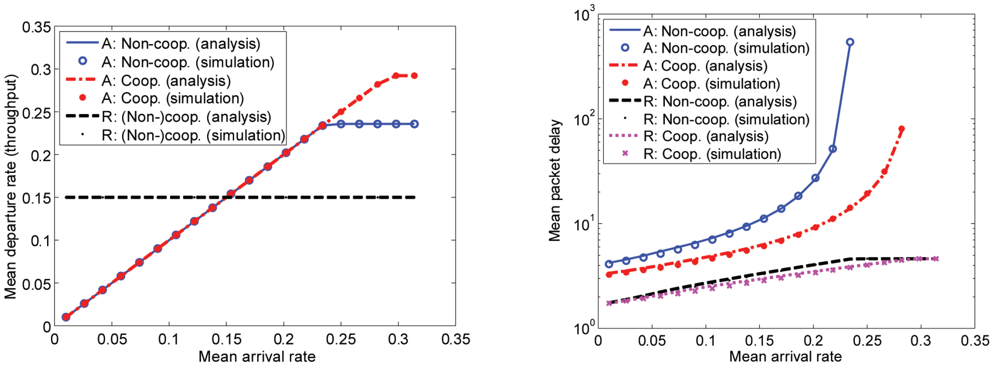
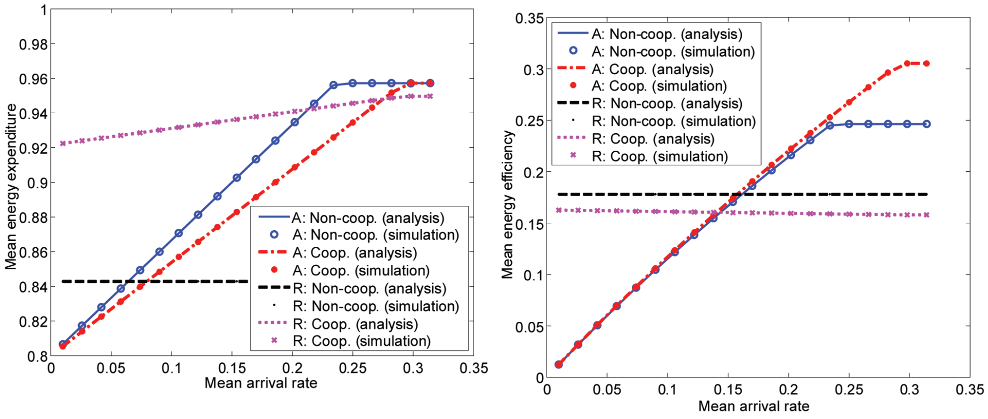


Figure 4. Dependency of energy expenditure (left) and energy efficiency (right) of source nodes on mean packet arrival rate λ_A



Numerical Results and Conclusions

In this section, we discuss some simulation results of the considered three node client relay system (Figures 3 and 4). Particularly, we focus on throughput, mean packet delay, energy expenditure, and energy efficiency of the source nodes. Partly following (Rong & Ephremides, 2009), the simulation parameters are set as: $p_{AB} = 0.3$, $p_{RB} = 0.7$, $p_{AR} = 0.4$, $p_{CB} = 0.5$, $P_{TX} = 1.0$, $P_{RX} = 0.9$, $P_I = 0.8$, $\lambda_R = 0.15$, whereas λ_A is varied across the system stability region. The plots compare the two scenarios: with cooperation (*cooperative*) and without cooperation (*non-cooperative*).

Evidently, the proposed analytical approach to the performance evaluation of the considered three node client relay system shows excellent agreement with simulation data. The obtained results allow for concluding upon the feasibility of the client relay technology. In particular, the originator throughput gain is up to 24% in saturation region. As such, client collaboration makes a promising technique to improve cell-edge user performance in the contemporary and future wireless cellular networks.

By contrast to known approaches where research is significantly simulation-based, this paper primarily introduces a formal mathematical model to assess the performance of a client relay network. The addressed client collaboration mechanisms may be implemented as part of networking equipment produced by Motorola, Intel, Nokia, etc. The proposed cooperation protocols could be tailored to next-generation telecommunication standards IEEE 802.16m (Marks, Nikolich, & Snyder, in press) and LTE-Advanced (Nakamura, in press).

ACKNOWLEDGMENTS

This work is supported by Graduate School in Electronics, Telecommunications and Automation, Tampere Graduate School in Information Science and Engineering, Nokia Foundation, and HPY Research Foundation.

REFERENCES

Andreev, S., Galinina, O., & Turlikov, A. (2010). Basic client relay model for wireless cellular networks. In *Proceedings of the International Congress on Ultra Modern Telecommunications and Control Systems and Workshops* (pp. 909-915).

Andreev, S., Galinina, O., & Vinel, A. (2010). Cross-layer channel-aware approaches for modern wireless networks. In A. Vinel, B. Bellalta, C. Sacchi, A. Lyakhov, M. Telek, & M. Oliver (Eds.), *Proceedings of the Third International Workshop on Multiple Access Communications* (LNCS 6235, pp. 163-179).

Andreev, S., Koucheryavy, Y., Himayat, N., Gonchukov, P., & Turlikov, A. (2010, December). *Active-mode power optimization in OFDMA-based wireless networks*. Paper presented at the 6th IEEE Broadband Wireless Access Workshop, Miami, FL.

Anisimov, A., Andreev, S., Galinina, O., & Turlikov, A. (2010). Comparative analysis of sleep mode control algorithms for contemporary metropolitan area wireless networks. In S. Balandin, R. Dunaytsev, & Y. Koucheryavy (Eds.), *Proceedings of the 10th International Conference on Smart Spaces and Next Generation Wired/Wireless Networking* (LNCS 6294, pp. 184-195).

Benini, L., Bogliolo, A., & de Micheli, G. (2000). A survey of design techniques for system-level dynamic power management. *IEEE Transactions on Very Large Scale Integration*, *8*(3), 299–316. doi:10.1109/92.845896

Cui, S., Goldsmith, A., & Bahai, A. (2004). Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks. *IEEE Journal on Selected Areas in Communications*, *22*, 1089–1098. doi:10.1109/JSAC.2004.830916

Haenggi, M., & Puccinelli, D. (2005). Routing in ad hoc networks: A case for long hops. *IEEE Communications Magazine*, *43*, 112–119. doi:10.1109/MCOM.2005.1522131

Jaiswal, N. (1968). *Priority queues*. New York, NY: Academic Press.

Jayaweera, S. K. (2004). An energy-efficient virtual MIMO architecture based on V-BLAST processing for distributed wireless sensor networks. In *Proceedings of the First Annual IEEE Communications Conference on Sensor and Ad Hoc Communications and Networks* (pp. 299-308).

Kim, H. (2009). *Exploring tradeoffs in wireless networks under flow-level traffic: Energy, capacity and QoS*. Unpublished doctoral dissertation, University of Texas, Austin.

Kleinrock, L. (1975). Queueing systems: *Vol. 1. Theory*. New York, NY: John Wiley & Sons.

Lahiri, K., Raghunathan, A., Dey, S., & Panigrahi, D. (2002). Battery-driven system design: A new frontier in low power design. In *Proceedings of the 15th IEEE International Conference on Design Automation and the 7th International Conference on Very Large Scale Integration Design* (pp. 261-267).

Marks, R. B., Nikolich, P., & Snyder, R. (in press). *IEEE Std 802.16m, Amendment to IEEE standard for local and metropolitan area networks – Part 16: Air interface for broadband wireless access systems – Advanced air interface*. Retrieved from http://ieee802.org/16/pubs/80216m.html

Miao, G. (2008). *Cross-layer optimization for spectral and energy efficiency*. Unpublished doctoral dissertation, Georgia Institute of Technology, School of Electrical and Computer Engineering, Atlanta.

Nakamura, T. (in press). *LTE release 10 & beyond (LTE-Advanced)*. Retrieved from http://www.3gpp.org/article/lte-advanced

Pyattaev, A., Andreev, S., Koucheryavy, Y., & Moltchanov, D. (2010, December). *Some modeling approaches for client relay networks*. Paper presented at the 15th IEEE International Workshop on Computer Aided Modeling Analysis and Design of Communication Links and Networks, Miami, FL.

Pyattaev, A., Andreev, S., Vinel, A., & Sokolov, B. (2010). *Client relay simulation model for centralized wireless networks*. Paper presented at the Federation of European Simulation Societies Congress, Prague, Czech Republic.

Rabaey, J., Ammer, J., da Silva, J., Jr., & Patel, D. (2000). PicoRadio: Ad-hoc wireless networking of ubiquitous low-energy sensor/monitor nodes. In *Proceedings of the IEEE Computer Society Annual Workshop on Very Large Scale Integration* (pp. 9-12). Washington, DC: IEEE Computer Society.

Rong, B., & Ephremides, A. (2009). On opportunistic cooperation for improving the stability region with multipacket reception. In *Proceedings of the 3rd Euro-NF Conference on Network Control and Optimization* (pp. 45-59).

Schurgers, C. (2002). *Energy-aware wireless communications*. Unpublished doctoral dissertation, University of California, Los Angeles.

Song, G. (2005). *Cross-layer optimization for spectral and energy efficiency*. Unpublished doctoral dissertation, Georgia Institute of Technology, School of Electrical and Computer Engineering, Atlanta.

Song, G., & Li, Y. (2006). Asymptotic throughput analysis for channel-aware scheduling. *IEEE Transactions on Communications*, *54*(10), 1827–1834. doi:10.1109/TCOMM.2006.881254

Stuber, G. (2001). *Principles of mobile communication*. Boston, MA: Kluwer Academic Publishers.

Publication 7

S. Andreev, E. Pustovalov, and A. Turlikov, “A practical tree algorithm with successive interference cancellation for delay reduction in IEEE 802.16 networks,” in *Proc. of the 18th International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA)*, pp. 301–315, 2011.

A Practical Tree Algorithm with Successive Interference Cancellation for Delay Reduction in IEEE 802.16 Networks

Sergey Andreev[1], Eugeny Pustovalov[2], and Andrey Turlikov[2]

Tampere University of Technology[1] (TUT), FINLAND
sergey.andreev@tut.fi
State University of Aerospace Instrumentation[2] (SUAI), St. Petersburg, RUSSIA
eugeny@vu.spb.ru, turlikov@vu.spb.ru

Abstract. This paper thoroughly studies a modification of tree algorithm with successive interference cancellation. In particular, we focus on the algorithm throughput and account for a single signal memory location, as well as cancellation errors of three types. The resulting scheme is robust to imperfect interference cancellation and is tailored to the uplink bandwidth request collision resolution in an IEEE 802.16 cellular network. The mean packet delay is shown to be considerably reduced when using the proposed approach.

Keywords: tree algorithm, successive interference cancellation, throughput, mean packet delay.

1 Introduction and background

Contemporary communication networks adopt multi-access techniques to arbitrate the access of the user population to the shared communication link. Multiple access algorithms are specifically designed to effectively control the resource allocation and reside at the Medium Access Control (MAC) layer. They constitute an important component of the widespread wireless protocols, such as IEEE 802.11 (Wi-Fi) and IEEE 802.16 (WiMAX). Random Multiple Access (RMA) algorithms are often used due to their simple implementation and reasonably high performance.

We remind that every MAC algorithm comprises a Channel Access Algorithm (CAA) and a Collision Resolution Algorithm (CRA). Whereas the former arbitrates user access to the shared medium, the latter is responsible for the collision resolution, whenever two or more users transmit their packets simultaneously. The most widespread ALOHA-based family of algorithms includes diversity slotted ALOHA, binary exponential backoff, and other popular mechanisms. The main idea of these approaches is to specify CAA and defer the packet retransmission after a collision took place for some random future time.

By contrast, tree algorithms independently proposed in [1] and [2], focus on CRA and thus demonstrate higher efficiency. The family of conventional tree algorithms is represented by Standard Tree Algorithm (STA) and Modified Tree

Algorithm (MTA). During the operation of the conventional STA and MTA, it is implicitly assumed that no meaningful information is extracted at the receiver after a collision. However, recent advances in physical (PHY) layer techniques allow using Successive Interference Cancellation (SIC) techniques [3], [4]. During SIC operation the packets involved into a collision may be restored successively. In [3], it was argued that SIC is naturally applicable to the uplink packet transmission in centralized communication networks. As such, in this paper we consider the prominent IEEE 802.16 [5] wireless cellular protocol.

The pioneering research work [6] proposes a combination of SIC and a tree algorithm (SICTA) to improve the performance of the conventional tree algorithms. Briefly, the acquired collision signals are stored in the signal memory of the receiver, which is assumed to be unbounded and then processed by SIC. Consequently, the performance of SICTA algorithm is shown to double the performance of the conventional STA. In the subsequent years, several modifications of the baseline SICTA algorithm were proposed [7], [8], [9], including the solutions that take advantage of the bounded signal memory. In particular, [10] proposes a SICTA-based algorithm to replace the standard algorithm at the bandwidth requesting stage in IEEE 802.16 networks.

All the existing SICTA-based solutions may be classified into two categories. Firstly, there are algorithms that assume perfect SIC operation and therefore are susceptible to cancellation errors falling into a deadlock. Secondly, there are algorithms that are robust to imperfect SIC operation, but at the same time are unstable when the number of users grows unboundedly. In our previous work [11], we proposed a robust SICTA algorithm that tolerates cancellation errors and demonstrates nonzero performance even when the user population is infinite. In this paper, we extend our algorithm to account for a more realistic SIC operation and conduct its thorough throughput analysis. Finally, we tailor the proposed solution for the uplink bandwidth requesting in the prominent IEEE 802.16 protocol.

2 System model and algorithms

2.1 Conventional tree algorithms

Consider tree multi-access algorithms proposed independently by [1] and [2] in the framework of classical RMA model with infinite user population and Poisson arrivals. We remind that each tree algorithm defines both CAA and CRA, which arbitrate user channel access and collision resolution process respectively. CRA is conveniently illustrated by a binary tree (see Figure 1, a), where a collided user selects right slot with probability p and selects left slot with probability $(1-p)$. Here the root corresponds to the set of users that collided initially, whereas the remaining nodes correspond to the subsets (possibly, empty) of users that decided to transmit in particular slots of the Collision Resolution Interval (CRI).

We note that in the example collision resolution tree (see Figure 1, a) a collision in slot 5 is inevitable, as none of the users which collided in slot 3

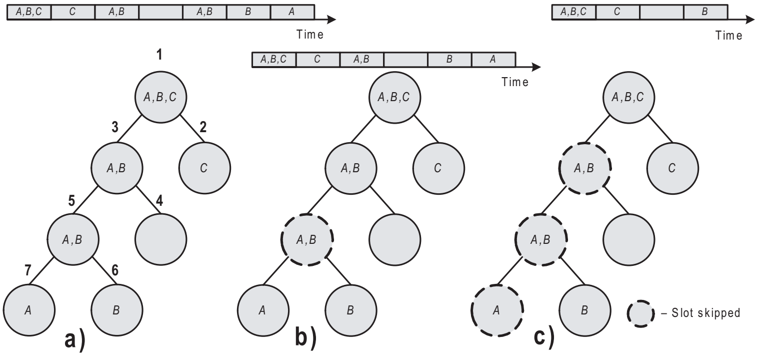


Fig. 1. Tree algorithms operation: a – STA; b – MTA; c – SICTA

select empty right slot 4. As such, it is reasonable to skip the collision slot 5 and proceed immediately to the next tree level. This simple modification is adopted by the conventional MTA (see Figure 1, b) to increase the performance of the conventional STA.

Maximum stable throughput is one of the most important performance parameters of tree algorithms. It may be defined as the highest arrival rate, which still results in the bounded value of the mean packet delay. The algorithm is stable for the given arrival rate if its mean packet delay is finite. Conventional STA has the throughput of 0.346, whereas conventional MTA improves throughput up to 0.375 in the framework of the classical model.

2.2 Successive interference cancellation

Recent advances in telecommunication equipment allow using SIC at the PHY layer [12], [4]. Generally, SIC is an approach to process a combination of wireless signals having some additional information. Following [7], we show how SIC may improve the performance of a tree algorithm in Figure 2. Assume for simplicity that the channel is error-free. Denote by y_s the signal received by the end of slot s. Similarly, denote by x_A and x_B the signals corresponding to packets A and B respectively. Let two users transmit their packets A and B in the first slot and collide. As such, the receiver acquires the combined signal $y_1 = x_A + x_B$ and decides that a collision occurred. The initial combined signal y_1 is then stored in the signal memory of the receiver.

After acquiring the signal $y_2 = x_A$ at the end of slot 2, the receiver successfully extracts signal x_A and decodes packet A. Further, SIC procedure processes signal y_1 and cancels the extracted signal x_A from the stored combination, that is, $\tilde{y}_1 = y_1 - x_A$. Then it is also possible to extract signal $x_B = \tilde{y}_1$ and to decode

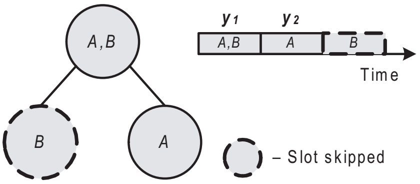


Fig. 2. Simple SIC example

packet B. Therefore, the subsequent collision resolution is not necessary. In the considered example the CRI duration is one slot less for any tree algorithm.

SIC-based algorithms typically exploit extended feedback from the PHY layer. This extra feedback is the result of the SIC operation. The baseline SICTA algorithm proposed in [6] requires the *K–EMPTY–COLLISION* feedback at MAC, where K is the number of successfully restored packets together with the number of left slots of the collision resolution tree labeled as empty after the SIC operation.

Consider example SICTA operation in Figure 1, c, where the CRI duration is only 4 slots. As left subtree may be skipped completely, the throughput is 0.693, that is, twice the STA throughput. More formal analysis is conducted below in subsection 3.2. In Figure 3, we detail the simplified SIC transceiver, which may be used to implement the baseline SICTA algorithm from [6]. Solid lines indicate transmission of data, whereas dashed lines indicate transmission of the control information. As follows from the figure, the unbounded signal memory at the receiver is practically infeasible. To mitigate this limitation, we propose our modification of SICTA that uses only a single signal memory location. As such, the implementation and operation complexity may be considerably reduced.

2.3 Imperfect interference cancellation

We note that in practical wireless devices using SIC, the interference cancellation is not perfect [13]. Cancellation errors may occur due to the residual signals after canceling the received signal in the stored combination. For example, after canceling the extracted signal x_A from the combined signal $x_A + x_B$ (see e.g. Figure 2) in slot s, the resulting signal contains $\tilde{y}_s = x_B + n_A$, where n_A is the residual signal x_A. After subsequent cancellation of x_B we analogously obtain $\tilde{\tilde{y}}_s = n_A + n_B$.

If the power of the residual signal $n_A + n_B$ is sufficiently high, the receiver mistakenly decides that the corresponding slot is not empty, that is, detects a non-existent collision between the users. For simplicity, we assume that this event occurs with some constant probability that depends on the receiver implementation. As such, due to the imperfect SIC operation, the PHY-MAC feedback is error-prone.

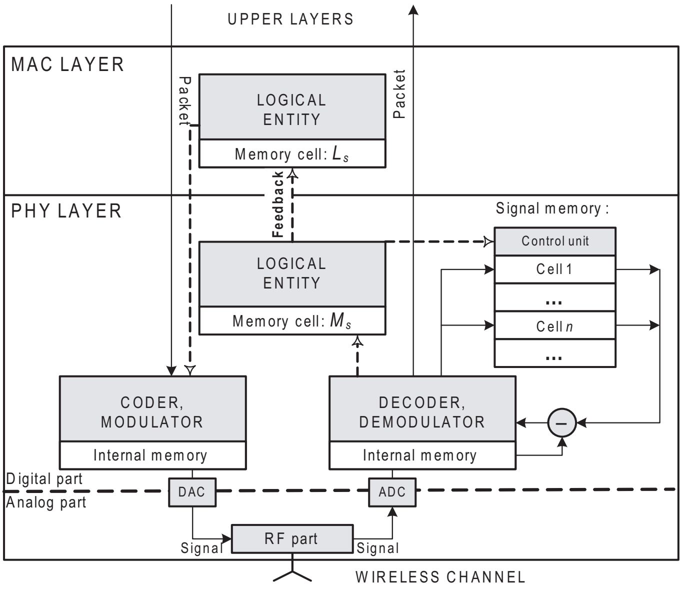


Fig. 3. SIC transceiver structure

Despite its high throughput, the baseline SICTA algorithm is vulnerable to cancellation errors. Indeed, assume that in the considered example in Figure 1, c the last cancellation operation of the signal x_B in the initial stored signal implies the residual signal n_B, that is, $\tilde{\tilde{y}}_1 = y_1 - x_C - x_B + n_B$. If now the power of n_B is high enough, the signal x_A may be extracted unsuccessfully and the collision resolution process would continue. Sooner or later, when packet A is decoded successfully, the residual signal power may be too high and the receiver would detect another non-existent collision in the left slot. According to the SICTA rules, the collision resolution process would then continue indefinitely unless it is aborted externally. As such, SICTA falls into a deadlock.

Our overview of SICTA-based algorithms indicates that currently there is no algorithm that is robust to imperfect interference cancellation and, at the same time, stable in the framework of the classical RMA model. The proposed robust SICTA (R-SICTA) algorithm tolerates cancellation errors for the cost of some reduction in its throughput. Additionally, it takes advantage of the single signal memory location at the receiver side. The increase in the amount of the available memory will result in growing throughput (still below the throughput of SICTA) and complexity [9].

The main idea of the proposed algorithm (see Figure 4, a) is to refrain from skipping particular collision slots (such as slot 3), which otherwise might result in a deadlock. In Figure 4, a, the time diagram for the best case corresponds to two successful cancellation operations, whereas the time diagram for the worst case corresponds to two unsuccessful operations. The formal description of our R-SICTA algorithm is given below in subsection 3.3. We note that due to cancellation errors the rules 3, 5, and 7 (see Table 1) must not allow skipping a slot of the collision resolution tree. Otherwise, deadlock effect may not be controlled.

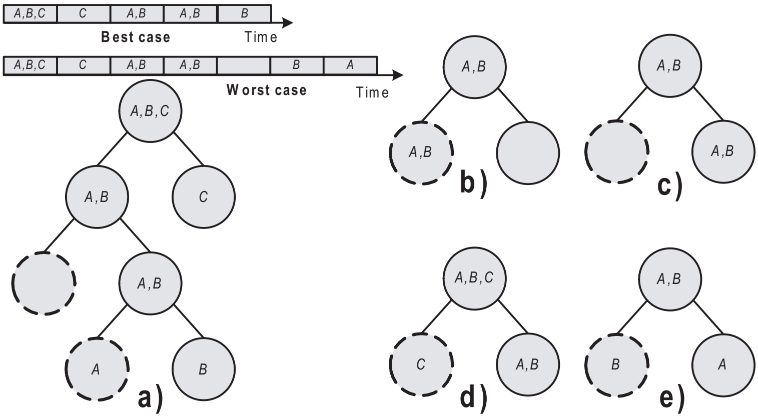


Fig. 4. Proposed R-SICTA algorithm: a – example operation of R-SICTA; b, c, d, e – some subtrees of R-SICTA

3 Proposed throughput calculation technique

3.1 General procedure

Here we describe an approach to obtain the throughput of a tree algorithm with successive interference cancellation. As such, we develop the re-calculation method for the mean collision resolution time from [14] and apply it to SICTA-based algorithms. Denote by v the duration of CRI in slots (time to resolve a collision of size k, or the number of nodes in the respective collision resolution tree), which is a discrete random variable. The conditional mean $E[v|$collision of size k is being resolved$]$ determines the CRI duration for a collision between k users. Our approach to the throughput calculation of tree algorithms using SIC is based on the following auxiliary assumptions.

Proposition 1. *Consider tree algorithm A using SIC. Denote by T_k^A the mean time to resolve a collision of size k for this algorithm. Taking $\frac{k}{T_k^A}$ into account, the following bounds for the throughput of A denoted by R_A may be established:*

$$\liminf_{k\to\infty} \frac{k}{T_k^A} < R_A < \limsup_{k\to\infty} \frac{k}{T_k^A}. \tag{1}$$

The proof of this proposition follows immediately from [1]. Consider now STA denoting the respective mean collision resolution time by T_k. Analogously to (1), we may derive the bounds for the throughput of STA omitting the lower index and denoting it simply by R:

$$\liminf_{k\to\infty} \frac{k}{T_k} < R < \limsup_{k\to\infty} \frac{k}{T_k}. \tag{2}$$

We note that the bounds for R were established in [15] and are equal to:

$$0.34657320 < R < 0.34657397. \tag{3}$$

Proposition 2. *Upper and lower bounds for the throughput R of STA may be obtained as:*

$$\left(\frac{2}{\ln 2} + c\right)^{-1} < R < \left(\frac{2}{\ln 2} - c\right)^{-1}, \tag{4}$$

where $c = 3.127 \cdot 10^{-6}$.

The interrelation between the bounds from (3) and expression (4) was established in [14].

Proposition 3. *Denoting the number of nodes in the collision resolution tree of STA by v, we readily obtain $T_k = E[v]$. The number of success, collision, and empty slots within CRI is denoted by v_s, v_c and v_e respectively. Then $v_s+v_c+v_e = v$ and at the same time:*

$$v_s = k; \quad v_c = \frac{v-1}{2}; \quad v_e = \frac{v+1}{2} - k. \tag{5}$$

The proof of the above relations may be found in [14].

Proposition 4. *Consider the collision resolution tree of algorithm A as the collision resolution tree of STA, where the time to pass some tree nodes is zero due to the SIC operation. We establish the number of thus skipped nodes r denoting by u the number of remaining nodes. Coming to the expected values, we may write:*

$$E[u] = E[v] - E[r] \quad \textit{or} \quad T_k^A = T_k - E[r]. \tag{6}$$

The proof of this proposition follows from Proposition 1 and equation (2). Substituting (6) into (1) and after some transformations, we derive the bounds for the throughput R_A as functions of known bounds for the throughput of STA (4). Below we use Proposition 3 to obtain $E[r]$ for the baseline SICTA algorithm, as well as for the proposed R-SICTA algorithm. Also we account for Proposition 4 to establish bounds for their throughputs.

3.2 Baseline SICTA algorithm

Below we calculate the throughput of the baseline SICTA algorithm from [6] as an example. The following feedback must be available from the SIC receiver for the proper operation of SICTA:

1. *COLLISION.*
2. *EMPTY.*
3. K – the number of successfully restored packets together with the number of left slots of the collision resolution tree labeled as empty after SIC operation ($K \geq 1$).

The operation of the respective PHY layer is detailed in Figure 3.

Proposition 5. *The mean number of non-skipped nodes in the collision resolution tree of SICTA (T_k^S) is established as:*

$$T_k^S = T_k - \frac{1}{2}E[v_s] - \frac{1}{2}(E[v_c] - 1) - \frac{1}{2}E[v_e] = T_k - \frac{1}{2}T_k + \frac{1}{2} = \frac{T_k + 1}{2}, \quad (7)$$

where T_k is the mean number of nodes in the collision resolution tree of STA.

The thorough proof of this proposition may be conducted using the approach from [14] and describing the performance of SICTA in the framework of the graph theory.

Consequently, accounting for (1) and (4), we establish the following bounds for the throughput R_S of SICTA:

$$\left(\frac{1}{\ln 2} + c\right)^{-1} < R_S < \left(\frac{1}{\ln 2} - c\right)^{-1}. \quad (8)$$

Concluding our analysis, we note that upper and lower bounds for the throughput of SICTA are sufficiently close to each other due to small value of c. Therefore, SICTA throughput may be written as:

$$R_S \approx \ln 2 \approx 0.693, \quad (9)$$

which gives the previously known result from [6]. However, our approach if far simpler and has reduced computational burden.

3.3 Proposed R-SICTA algorithm

Consider now the proposed R-SICTA algorithm, which is robust to imperfect interference cancellation. This algorithm stores a single collision signal and cancels both success and collision signals (see Figure 4, a). It is also designed to tolerate cancellation errors (see subsection 2.3).

The following feedback must be available from the SIC receiver for the proper operation of R-SICTA:

1. *COLLISION* and slot skip (contents of the left slot extracted) (C/skip).
2. *COLLISION* and no slot skip (C/–).
3. *SUCCESS*/*EMPTY* and no slot skip (SE/–).
4. *SUCCESS* and slot skip (contents of the left slot extracted) (S/skip).
5. *EMPTY* and slot skip (inevitable collision in the following slot) (E/skip).

Algorithm 1 describes the operation of R-SICTA MAC. Denote the received signal by cs, the stored signal by ss, and the extracted signal by ms. The PHY operation is summarized in Table 1.

```
Ensure: During algorithm operation, account for a particular CAA.
 1: Reset position $L$ of a user in collision resolution tree.
 2: Generate new arrivals to the user according to a particular arrival flow.
 3: if user has pending packets then
 4:         if position of user $L = 0$ then
 5:                 Transmit a packet.
 6: Wait for end of the current slot.
 7: Receive feedback from PHY.
 8: if C/skip feedback is received then
 9:         if $L = 1$, then
10:                 Delete the pending packet.
11:         else if $L = 0$, then
12:                 $L = \begin{cases} 0 & \text{with probability } p, \\ 1 & \text{with probability } 1-p. \end{cases}$
13: else if C/– feedback is received then
14:         if $L > 0$, then
15:                 $L = L + 1$.
16:         else
17:                 $L = \begin{cases} 0 & \text{with probability } p, \\ 1 & \text{with probability } 1-p. \end{cases}$
18: else if SE/– feedback is received then
19:         if $L > 0$, then
20:                 $L = L - 1$.
21:         else
22:                 Delete the pending packet.
23: else if S/skip feedback is received then
24:         if $L \geq 2$, then
25:                 $L = L - 2$.
26:         else
27:                 Delete the pending packet.
28: else if E/skip feedback is received then
29:         if $L = 1$, then
30:                 $L = \begin{cases} 0 & \text{with probability } p, \\ 1 & \text{with probability } 1-p. \end{cases}$
31: Go to step 2.
```

Algorithm 1: R-SICTA MAC operation

Table 1. R-SICTA PHY operation

Rule	Channel – PHY	Memory contents	PHY – MAC	Store
1	*COLLISION*	$ss - cs = 0$	C/skip	cs
2	*COLLISION*	$ss - cs = ms$	C/skip	cs
3	*COLLISION*	otherwise	C/-	cs
4	*SUCCESS*	$ss - cs = ms$	S/skip	0
5	*SUCCESS*	otherwise	SE/-	0
6	*EMPTY*	$ss \neq 0$	E/skip	ss
7	*EMPTY*	$ss = 0$	SE/-	0

Below we express T_k^{RS} according to Proposition 4. As such, the following tree nodes are subtracted from T_k as they are skipped due to SIC operation:

- Figure 4, b, when a collision slot is followed by an empty slot. The MTA rules allow to skip a collision slot with probability 1.
- Figure 4, c, when a collision slot is followed by a collision slot. The SIC operation allows to skip an empty slot with probability $1 - q_{ce}$.
- Figure 4, d, when a collision slot is again followed by a collision slot. The SIC operation allows to skip a success slot with probability $1 - q_{cs}$.
- Figure 4, e, when a collision slot is followed by a success slot. The SIC operation allows to skip a success slot with probability $1 - q_{ss}$.

We note that q_{ce}, q_{cs}, and q_{ss} are the parameters of SIC and depend on its implementation. The following proposition may thus be formulated.

Proposition 6. *The mean number of non-skipped nodes in the collision resolution tree of R-SICTA (T_k^{RS}) is established as:*

$$T_k^{RS} = \left(\frac{1}{2} + \frac{q_{ce}}{4}\right) T_k - \frac{1}{2} + \frac{q_{ce}}{4} + \frac{k}{2}(1 + q_{cs} - q_{ce}) + \frac{N_k}{2}(q_{ss} - q_{cs}), \qquad (10)$$

where T_k is the mean number of nodes in the collision resolution tree of STA and N_k is the number of collisions of size 2 in the STA collision resolution tree of initial size k, whereas probabilities q_{ce}, q_{ss}, and q_{cs} were discussed above.

We omit the proof of this proposition due to the space limitations.

Finalizing our analysis, it is important to evaluate the relation $\frac{N_k}{k}$ for the increasingly large values of k. Clearly, $N_0 = N_1 = 0$ as there are no collision nodes in the respective collision resolution trees. Also it is easy to show that $N_2 = 2$. Generally, accounting for the properties of collision resolution trees, for $k > 2$ it follows that:

$$N_k = \frac{\sum\limits_{i=1}^{k-1} \binom{k}{i} N_i}{2^{k-1} - 1}. \qquad (11)$$

Equation (11) may be calculated recursively for any chosen value of k.

Consider the Poisson transform [16] of the sequence $N_0, N_1, \ldots, N_i, \ldots$, which is denoted by:

$$N(s) \triangleq \sum_{k \geq 0} N_k \cdot \frac{s^k}{k!} e^{-s}, \; s \in \mathcal{R}. \tag{12}$$

After some derivations, we may obtain the following recursive expression to establish $N(s)$:

$$N(s) = 2N\left(\frac{s}{2}\right) + \frac{s^2}{2} e^{-s}. \tag{13}$$

Consider also the normalized Poisson transform analogously to [15], which is denoted by:

$$M(s) \triangleq \frac{N(s)}{s}. \tag{14}$$

Using (13), we may rewrite (14) as:

$$M(s) = M\left(\frac{s}{2}\right) + \frac{s}{2} e^{-s}. \tag{15}$$

The function $M(s)$ is periodic for large values of its argument, which may be used to evaluate it. We consider the normalized Poisson transform for sufficiently large values of its argument $2^n r$, where $n \in \mathcal{Z}$ and $r \in \mathcal{R}$. Formally substituting $2^n r$ into (15), we obtain:

$$M(2^n r) = M(2^{n-1} r) + \frac{2^n r}{2} e^{-2^n r}. \tag{16}$$

For considerably large n, the variation of $M(2^n r)$ from 2^n to 2^{n+1} corresponds to one period of function $M(2^n r)$. Therefore, for $1 \leq r \leq 2$ and some value of n we obtain the highest and the lowest values of the function for all the subsequent values of its argument. Consider the equality (16) in more detail and execute the recursion for $n-1$ times:

$$M(2^n r) = M(r) + \sum_{i=1}^{n} \frac{2^i r}{2} e^{-2^i r} = M(r) + H_n(r). \tag{17}$$

The series $H_n(r)$ converge fast and easy to evaluate with any required precision. The values of $M(r)$ for low r are easy to obtain accounting for (12) and (14). Further, using (17), we study the behavior of the initial function $M(2^n r)$ over one period, when $n \geq 20$. It may be shown that the precision of the obtained values is then at least 10^{-8}. We give these values in the integer points k, that is, $M(2^n r) = M(k)$:

$$\max_{2^{20} \leq k \leq 2^{21}} M(k) = \limsup_{k \to \infty} M(k) < 0.72135464 + 1 \cdot 10^{-8} \tag{18}$$

and

$$\min_{2^{20} \le k \le 2^{21}} M(k) = \liminf_{k \to \infty} M(k) > 0.72134039 - 1 \cdot 10^{-8}.$$

Following the approach from [17] (Theorem 1), it may be shown that the lower and the upper bounds for $M(k)$ (18) also hold for the relation $\frac{N_k}{k}$. We note that as the limit of $\frac{N_k}{k}$ does not exist, we inevitably have an interval, where the behavior of $\frac{N_k}{k}$ is not determined. The length of this interval, however, is only 0.00001425.

Simplifying the representation of the final result, we note that $\limsup\limits_{k \to \infty} \frac{N_k}{k} = \liminf\limits_{k \to \infty} \frac{N_k}{k} = \gamma$ with the precision of at least three decimal digits and $\gamma = 0.721$. Also for the sake of simplicity, we avoid finding the upper and the lower bounds for the throughput of R-SICTA (R_{RS}) as they are sufficiently close to each other. Then the resulting approximation for the throughput of the proposed algorithm may be obtained as:

$$R_{RS} \approx \frac{2 \ln 2}{2 + q_{ce} + 2 \ln 2 (1 + q_{cs} - q_{ce} + (q_{ss} - q_{cs})\gamma)}. \tag{19}$$

In particular, when $q_{ce} = q_{ss} = q_{cs} = 0$, that is, when there are no cancellation errors, $R_{RS} \approx 0.515$.

4 IEEE 802.16 performance improvement and conclusions

The combination of successive interference cancellation at PHY and tree algorithms at MAC constitutes a promising direction toward the improvement of the contemporary communication protocols. In particular, it allows for the considerable throughput increase for the moderate cost of implementation and operation complexity. Currently, the family of SIC-based algorithms is known, where the baseline SICTA demonstrates the highest throughput of 0.693 under the classical set of assumptions. However, SICTA requires unbounded signal memory at the receiver side, which is practically infeasible. Moreover, its performance degrades significantly due to the imperfect interference cancellation.

We proposed a practical SICTA-based algorithm that mitigates the limitations of the baseline SICTA and has the throughput of 0.515 in case of no cancellation errors. Moreover, our R-SICTA is robust even in case of high cancellation error probability and demonstrates graceful performance degradation. In the worst case, when SIC operation is not possible, R-SICTA converges to MTA with the throughput of 0.375. In order to conclude upon the feasibility of the proposed algorithm, we consider its usage for the uplink bandwidth requesting in the prominent IEEE 802.16 protocol.

Below we evaluate the gain after the replacement of the standardized Binary Exponential Backoff (BEB) algorithm [5] with the proposed R-SICTA. We use our own IEEE 802.16 simulator extensively verified over the recent years and referenced by our previous works, e.g. [18]. The simulation parameters are summarized in Table 2. In Figure 5, we plot the reservation delay as the function of the overall arrival rate. Here M stands for the total number of users and N is the number of bandwidth request slots per frame. As expected, the baseline SICTA algorithm with unbounded signal memory demonstrates the lowest delay. However, the proposed R-SICTA algorithm with single signal memory performs closely to SICTA in case of no cancellation errors.

Table 2. Basic simulation parameters

IEEE 802.16 network parameter	Value
DL:UL	60:40
PHY type	OFDM
Frame duration	5 ms
Sub-channel bandwidth	7 MHz
Contention slot length	170 μs
Data packet size	4096 bits

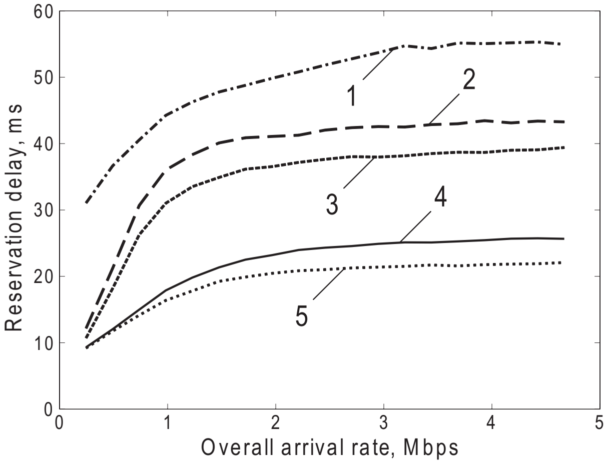


Fig. 5. IEEE 802.16 reservation delay for $M = 6$ and $N = 1$: 1 – BEB, $W = W_0$, $m = 0$; 2 – STA; 3 – MTA; 4 – R-SICTA, $q_{ce} = q_{ss} = q_{cs} = 0$; 5 – SICTA

In the worst case of imperfect interference cancellation, when $q_{ce} = q_{ss} = q_{cs} = 1$, the R-SICTA reservation delay approaches that of MTA. As such, the gap between curves 3 and 4 in Figure 5 demonstrates the range of possible gains from the proposed solution depending on the cancellation error probability. Finally, STA has higher reservation delay than MTA, whereas the standardized BEB algorithm is clearly the worst case even with the optimal operation parameters W and m [18].

Figure 6 plots the overall packet delay in IEEE 802.16. The results generally follow the respective trends as in Figure 5, but show different delay values. In particular, for the case without cancellation errors, the proposed R-SICTA has the gain of 50-60% comparatively to the standardized BEB algortihm and depending on the arrival rate. In the worst case of imperfect interference cancellation, when $q_{ce} = q_{ss} = q_{cs} = 1$, the gain reduces to 25-56%, but still remains considerable.

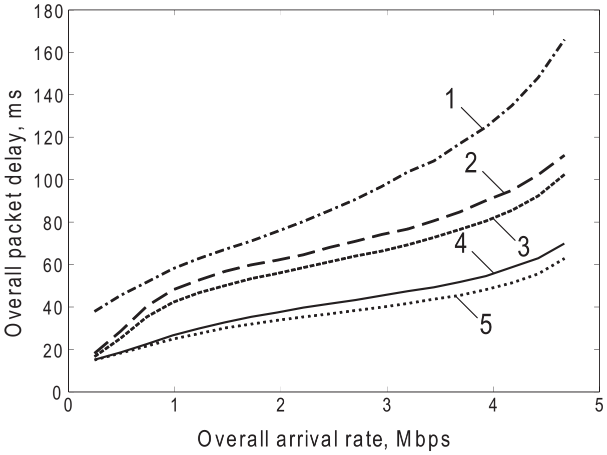


Fig. 6. IEEE 802.16 overall delay for $M = 6$ and $N = 1$: 1 – BEB, $W = W_0$, $m = 0$; 2 – STA; 3 – MTA; 4 – R-SICTA, $q_{ce} = q_{ss} = q_{cs} = 0$; 5 – SICTA

Summarizing, we have thoroughly analyzed the proposed practical R-SICTA algorithm with successive interference cancellation. We established its throughput, as well as tailored it to the uplink bandwidth requesting in the contemporary IEEE 802.16 protocol. Our results indicate significant delay gains after the replacement of the standardized BEB algorithm with the proposed R-SICTA. As such, the proposed solution is attractive to improve the performance of modern cellular networks.

Acknowledgments

This work is supported by the Russian Foundation for Basic Research (project # 10-08-01071-a).

References

1. B. S. Tsybakov and V. A. Mikhailov, “Free synchronous packet access in a broadcast channel with feedback,” *Problems of Information Transmission*, vol. 14, no. 4, pp. 32–59, 1978.
2. J. I. Capetanakis, “Tree algorithms for packet broadcast channels,” *IEEE Transactions on Information Theory*, vol. 25, no. 4, pp. 505–515, 1979.
3. K. Pedersen, T. Kolding, I. Seskar, and J. Holtzman, “Practical implementation of successive interference cancellation in DS/CDMA systems,” in *Proceedings of IEEE ICUPC*, vol. 1, pp. 321–325, 1996.
4. S. Weber, J. Andrews, X. Yang, and G. Veciana, “Transmission capacity of wireless ad hoc networks with successive interference cancellation,” *IEEE Transactions on Information Theory*, vol. 53, no. 8, pp. 2799–2814, 2007.
5. *IEEE Std 802.16m (D9), Amendment to IEEE Standard for Local and metropolitan area networks. Advanced Air Interface.*
6. Y. Yu and G. B. Giannakis, “SICTA: A 0.693 contention tree algorithm using successive interference cancellation,” *Proceedings of IEEE INFOCOM*, vol. 3, pp. 1908–1916, 2005.
7. Y. Yu and G. B. Giannakis, “High-throughput random access using successive interference cancellation in a tree algorithm,” *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4628–4639, 2007.
8. X. Wang, Y. Yu, and G. B. Giannakis, “A robust high-throughput tree algorithm using successive interference cancellation,” *IEEE Transactions on Communications*, vol. 55, no. 12, pp. 2253–2256, 2007.
9. G. T. Peeters and B. V. Houdt, “Interference cancellation tree algorithms with k-signal memory locations,” *IEEE Transactions on Communications*, vol. 58, no. 11, pp. 3056–3061, 2010.
10. X. Wang, Y. Yu, and G. B. Giannakis, “Combining random backoff with a cross-layer tree algorithm for random access in IEEE 802.16,” *IEEE Wireless Communications and Networking Conference*, vol. 2, pp. 972–977, 2006.
11. S. Andreev, E. Pustovalov, and A. Turlikov, “SICTA modifications with single memory location and resistant to cancellation errors,” *Proceedings of NEW2AN*, vol. 5174/2008, pp. 13–24, 2008.
12. A. Agrawal, J. Andrews, J. Cioffi, and T. Meng, “Iterative power pontrol for imperfect successive interference cancellation,” *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 878–884, 2005.
13. J. Andrews and A. Hasan, “Analysis of cancellation error for successive interference cancellation with imperfect channel estimation,” tech. rep., EE-381K: Multiuser Wireless Communications, 2002.
14. G. S. Evseev and A. M. Turlikov, “A connection between characteristics of blocked stack algorithms for random multiple access system,” *Problems of Information Transmission*, vol. 43, no. 4, pp. 345–279, 2007.
15. L. Gyorfi, S. Gyori, and J. L. Massey, “Principles of stability analysis for random accessing with feedback,” *NATO Security through Science Series: Information and Communication Security*, vol. 10, pp. 214–250, 2007.
16. W. Szpankowski, *Average Case Analysis of Algorithms on Sequences.* Wiley, 2001.
17. L. Gyorfi and S. Gyori, “Analysis of tree algorithm for collision resolution,” in *Proceedings of AofA*, pp. 357–364, 2005.
18. S. Andreev, A. Turlikov, and A. Vinel, “Contention-based polling efficiency in broadband wireless networks,” *Proceedings of ASMTA*, vol. 5055/2008, pp. 295–309, 2008.

Publication 8

S. Andreev, O. Galinina, and Y. Koucheryavy, “Energy-efficient client relay scheme for machine-to-machine communication,” in *Proc. of the 54th IEEE Global Communications Conference (GLOBECOM)*, 2011.

Energy-Efficient Client Relay Scheme for Machine-to-Machine Communication

Sergey Andreev, Olga Galinina, and Yevgeni Koucheryavy
Tampere University of Technology, Tampere, FINLAND
E-mails: {sergey.andreev, olga.galinina}@tut.fi, yk@cs.tut.fi

***Abstract*—In this paper, we consider a wireless cellular network capable of supporting Machine-to-Machine (M2M) applications. According to the recent IEEE 802.16p proposals, a wireless M2M device may act as an aggregation point and communicate data packets on behalf of the other M2M devices, which may lack a cellular interface or have a poor communication link to the network. We propose a client relay scheme to improve the link reliability and energy efficiency for devices with weak links. Performance of the proposed scheme is evaluated through analysis and simulation across several metrics covering client throughput, latency, and energy consumption. Our analytical approach is a novel queueing model that captures realistic traffic arrival patterns borrowed from the evaluation methodology. It is shown that the obtained analytical results demonstrate excellent agreement with simulation. We also conclude that the proposed client relay scheme may save power for devices with poor communication link.**

I. Introduction and Background

According to [1], *Machine-to-Machine* (M2M) communication may be defined as information exchange between a Subscriber Station (SS) and a Server in the core network through a Base Station (BS), which may be carried out without any human interaction. Industry reports indicate considerable potential of this market, with millions of devices connected within the following five years resulting in predicted revenues of $300 billion [2]. Due to its huge market potential, several cellular standards are now focusing on developing air interface enhancements to support M2M communication.

For example, emerging IEEE 802.16p proposals [1] address enhancements for IEEE 802.16m standard to support M2M applications. 3GPP LTE also has several work items defined on M2M communications, primarily with respect to overload control [3]. The IEEE 802.16 M2M study report covers several M2M use cases under broader categories of Metering, Secured Access and Surveillance, Remote Maintenance and Control, and to a limited extent under Tracking, Tracing & Recovery. For further details on these use cases, see [1].

A key M2M use case is smart metering that involves meters autonomously reporting *usage* and *alarm* information to grid infrastructure to help reduce operational cost, as well as regulate customer's utility usage based on load-dependent pricing signals received from the grid. We expect that wireless technologies, such as IEEE 802.16 and LTE, will play a key role in enabling smart metering applications.

This paper studies a typical smart metering M2M application scenario in the context of IEEE 802.16 wireless cellular network, which features a large number of devices connecting to the network. We focus on enhancing the performance of cell-edge M2M devices with poor communication link and propose a simple and feasible client relay scheme to improve link performance. An analytical approach is formulated to predict performance improvement of the proposed scheme. Key performance metrics, including throughput, latency (packet delay), and energy efficiency [4] are addressed in greater detail to highlight important issues.

Results of our analysis indicate that latency and energy expenditure of cell-edge M2M devices may be dramatically lowered even when there is a surge in near simultaneous network entry attempts by a large number of meters. Such surge in network access attempts may occur, for example, in a power outage scenario where a large number of smart meters attempts to connect to the network to report the outage event and again when they reconnect to the network upon restoration of power. Our recent contributions to IEEE 802.16p [5] characterize the *network overload* resulting from such alarm events. However, the currently proposed client relay scheme can help ensure that the performance of other cellular devices is not adversely impacted by a large number of uncontrolled network access attempts from M2M devices.

Although we study the client relay approach in the context of smart metering M2M applications, the concepts are equally applicable to other M2M use cases, such as in-building sensors or surveillance equipment that may have weak connection to the network. Also, the general conclusions of this paper are applicable to other M2M-enhanced wireless systems, such as 3GPP LTE.

II. System Model and Analysis

A. M2M Architecture

Figure 1 captures the considered system architecture, which is based on the IEEE 802.16-based M2M communication architecture shown in [1]. The *IEEE 802.16 M2M device* is an IEEE 802.16 SS with M2M functionality. The *M2M Server* is an entity that communicates to one or more IEEE 802.16 M2M devices through an IEEE 802.16 BS. It has an interface which can be accessed by an M2M service consumer (e.g., utility company). Note that the M2M system architecture allows for an IEEE 802.16 M2M device to act as an *aggregation point* for *non-IEEE 802.16 M2M devices* (sensors or meters) without a cellular interface. These non-IEEE 802.16 M2M devices may use different radio interfaces, such as IEEE 802.11, IEEE 802.15, PLC, etc.

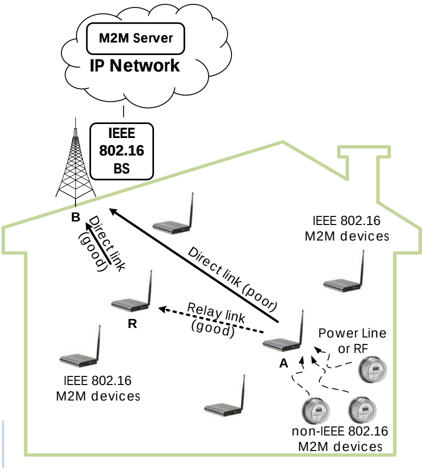


Fig. 1: Relay-enhanced M2M architecture

Importantly, an IEEE 802.16 M2M device can also act as a cooperator for another IEEE 802.16 M2M device. That is, an IEEE 802.16 M2M device R may relay traffic on behalf of e.g. device A with poor communication link and thus improve its performance. In this case, air interface changes to IEEE 802.16 may be expected to handle the *client relay* functionality. Particularly, the operation of R should enable eavesdrop mode to capture traffic from A. In our earlier work [6], we proposed a simple client relay protocol that may be used in cellular networks. In what follows, we tailor our client relay scheme from [6] to the considered M2M use case.

B. System Model

We consider an M2M-compatible wireless cellular network enhanced with client relay capability following the basic methodology from [6] and extending it in what follows to mimic IEEE 802.16p operation. We concentrate on the above realistic network topology (see Figure 1) comprising multiple source M2M and non-M2M nodes and one sink node (the BS). We focus on the performance of the tagged M2M aggregation point A and term it the *originator*. The originator aggregates data packets from N non-IEEE 802.16 M2M devices with the overall mean arrival rate λ_A. We term another M2M source node R the *relay*. The relay generates own data packets with the mean arrival rate λ_R. Additionally, the relay is capable of eavesdropping on the transmissions from the originator and may temporarily store the packets from A for the subsequent retransmission. The node B is termed the *base station* and receives data packets from both the originator and the relay, as well as from the other clients. The base station has no outgoing traffic. Below we detail the system model.

Assumption 1. **The system.** System time is discrete and the unity of system time is termed the *frame*. All the communicated data packets from M2M devices are sufficiently short to be transmitted within their *random-access* requests [5]. As such, it is not necessary to explicitly schedule the data packet transmissions and they adhere to the contention-based procedure specified by the standard [7]. Reducing control overhead, each frame has exactly one random-access opportunity to contend for (a contention slot).

As an example, see Figure 2, where transmissions from node R are always successful in frames no. 2, 4, and 6. We, however, focus on the unfortunate events for node A. Since the relay is incapable of simultaneous transmission and reception, we assume that the source nodes alternate accessing the channel with probability 0.5 (e.g., they may be split into two multicast groups). Initially, node A fails random-access procedure in frame no. 1. When it later accesses the channel in frames no. 3 and 5, it fails to transmit successfully due to its poor direct link. Node R firstly fails to eavesdrop in frame no. 3, but then succeeds to do so in frame no. 5. Finally, node R transmits *simultaneously* with node A in frame no. 7 (creating a virtual MIMO link) and the transmission is successful due to the cooperative gain (see [6] for more details). As such, our scheme is different from other research in the field [8] and is easy to implement.

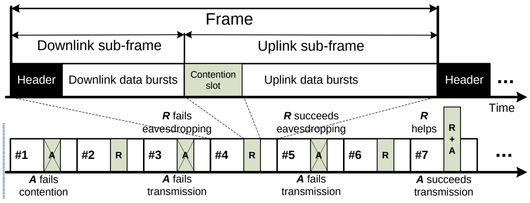


Fig. 2: Example of client relay operation

Assumption 2. **The traffic.** The aggregate traffic from N meters to the originator A has the following realistic properties (see Figure 3). Every meter has a specific flow of packets and we assume for simplicity that flows from different meters are mutually independent. Each flow belongs to a particular type i, where $i = \overline{1, L}$. Here L is the total number of source types (usage meters, alarm meters, etc.). A type is defined by two parameters: the period length T_i and the transmission duration T_i^{ON} (or, the ON-period length). See Figure 3 as an example for $N = 3$, where all three types are different. Note that a meter has a single ON-period over a period of T_i. As suggested by the recommendation [9], the random time interval between two consecutive ON-periods (T_i^{OFF}) follows either Uniform $U[0, T_i - T_i^{ON}]$, or Beta $Be(3, 4)$ distribution. One of L types is associated with a meter randomly following the distribution $\{p_i\}_{i=1}^{L}$, where p_i is the probability that a meter belongs to type i.

The own traffic to the relay R is Poisson. Note that our previous work [6] assumed Poisson arrival process to both the originator and the relay. However, in this paper we extend the model for the realistic M2M traffic arrival patterns at A. Our methodology allows traffic at R be aggregated as well,

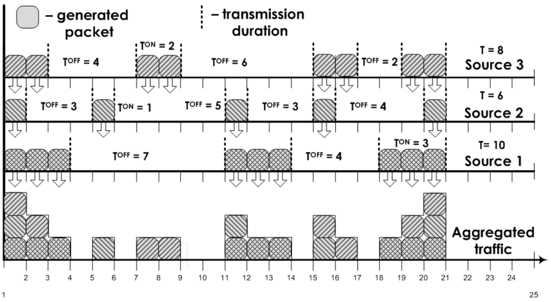


Fig. 3: Example of aggregated arrival process

TABLE I: Analytical model notations

Notation	Parameter description
λ_X	Mean arrival rate of packets to node X (A or R)
p_{AB}	Probability of successful reception at B when A transmits
p_{RB}	Probability of successful reception at B when R transmits
p_{AR}	Probability of successful reception at R when A transmits
p_{CB}	Probability of successful reception from A and R at B when nodes cooperate
τ	Mean service time of a packet
ρ	Queue load coefficient
q_X	Mean queue length of node X
δ_X	Mean packet delay of node X
η_X	Mean throughput of node X
ϵ_X	Mean energy expenditure of node X
ϕ_X	Mean energy efficiency of node X
H_A	Hurst parameter of self-similar process at node A
P_S	Contention success probability

and we assume Poisson distribution here only to simplify the presentation of analytical results. Interestingly, when the population N of meters is sufficiently high, the aggregated traffic demonstrates strong *self-similar* properties. We, therefore, account for Hurst parameter [10] of self-similar process at node A. If the major part of N sources has the on-period T_i^{ON} other than 1, the Hurst parameter is different from 0.5 (for Poisson traffic). For example, assuming that all the sources have $T^{ON} = 2$, we obtain traffic with Hurst parameter H of about 0.65. In case of $T^{ON} = 3$, H would be close to 0.75. It means that the observed traffic is clearly not Poisson.

The specific arrival flow types to other source nodes in the system are not relevant for the subsequent analysis, as these nodes would only impact nodes A and R through the contention success probability P_S. This is the probability of a node to succeed in accessing the contention slot, given that it has transmitted its random-access request, and it is considered constant. In practice, this probability is influenced by numerous factors, including system configuration and parameters, and its estimation is a separate research task. Instead, we study how performance metrics of interest depend on this probability.

Assumption 3. **The clients.** Both the originator and the relay have unbounded queues to store own data packets. We assume FIFO discipline for both queues. Additionally, the relay has an extra memory location to keep a *single* data packet from the originator for the subsequent cooperative retransmission. In our previous work [6], it was demonstrated that single memory cell is sufficient for the proposed client relay system operation.

Assumption 4. **The channel.** The communication channel is error-prone and is based on the multi-packet reception channel model [11]. The transmitted data packet is thus received by the destination successfully with the constant probability dependent only on the link type (direct or relay) and on which nodes are transmitting simultaneously. The non-zero success probabilities, as well as other relevant system model parameters are summarized in Table I. It is expected that $p_{AR} > p_{AB}$, as well as $p_{CB} > p_{AB}$. If the packet is not received successfully, it is retransmitted by its source. The maximum number of allowed transmissions is unlimited.

C. Analytical Results

Firstly, our analytical approach is applicable for establishing the exact mean departure rate of packets from (throughput of) nodes A and R. In particular, the throughput of A is given by:

$$\eta_A = \begin{cases} \lambda_A, \text{ no saturation} \\ (1 - \lambda_R \tau_{R0})\tau_{A0}^{-1}, \text{ saturation for } A \\ (2\tau_{A0})^{-1}, \text{ saturation for } A, \ R. \end{cases}$$

The throughput of R may be derived similarly. The saturation conditions are defined as follows:

- For A: $(\lambda_A \tau_{A0} + \lambda_R \tau_{R0} > 1)$ and $(\lambda_R \tau_{R0} < 0.5)$.
- For R: $(\lambda_A \tau_{A0} + \lambda_R \tau_{R0} > 1)$ and $(\lambda_A \tau_{A0} < 0.5)$.
- For A and R: $(\lambda_A \tau_{A0} > 0.5)$ and $(\lambda_R \tau_{R0} > 0.5)$.

Here, the mean service time of a packet from node A conditioning on the fact that $\lambda_R = 0$ is:

$$\tau_{A0} = \frac{p_{AR}(1-p_{AB})}{1-p_{CB}-(1-p_{AB})(1-p_{AR})} \frac{1}{p_{CB}} \\ - \left(\frac{p_{AR}(1-p_{AB}) p_{CB}}{1-p_{CB}-(1-p_{AB})(1-p_{AR})} - p_{AB} \right) \frac{1}{[1-(1-p_{AB})(1-p_{AR})]^2}.$$

Note that the mean service time of a packet from node R conditioning on the fact that $\lambda_A = 0$ may be established simply as $\tau_{R0} = {p_{RB}}^{-1}$.

Secondly, we may obtain the exact value of the mean energy expenditure of node A as:

$$\varepsilon_A = P_{TX} \eta_A \tau_{A0} + P_I (1 - \eta_A \tau_{A0})$$

and of node R as:

$$\varepsilon_R = P_{TX} \left(\eta_R \tau_{R0} + \eta_A \frac{1-\tilde{p}_{AB} \tau_{A0}}{\tilde{p}_{CB} - \tilde{p}_{AB}} \right) \\ + P_{RX} \left(\eta_A \tau_{A0} - \eta_A \frac{1-\tilde{p}_{AB} \tau_{A0}}{\tilde{p}_{CB} - \tilde{p}_{AB}} \right) \\ + P_I \left(1 - \eta_R \tau_{R0} - \eta_A \tau_{A0} \right).$$

Here, P_{TX} is the average power that is spent by a node in the packet transmission state, P_{RX} is the average power that is spent by a node in the packet reception state, whereas P_I is the average power that is spent by a node in the idle state. The mean energy efficiencies of nodes A and R are given by expressions $\varphi_A = \eta_A {\varepsilon_A}^{-1}$ and $\varphi_R = \eta_R {\varepsilon_R}^{-1}$ respectively.

In the above analysis, $\tilde{p}_{AB} = P_S p_{AB} \rho_{A0} {\rho_{AR}}^{-1}$, whereas $\tilde{p}_{CB} = P_S p_{CB} \rho_{A0} {\rho_{AR}}^{-1}$. The queue load coefficient of node

A conditioning on the fact that $\lambda_A = 0$ is $\rho_{A0} = \lambda_A \tau_{A0}$, whereas the queue load coefficient of node R conditioning on the fact that $\lambda_R = 0$ is $\rho_{R0} = \lambda_R \tau_{R0}$. If we now set $\rho_{A0} > \rho_{R0}$ as an example, then the queue load coefficient of node A is $\rho_{AR} \cong \rho_{A0}(1 - \rho_{R0})^{-1}$, whereas the queue load coefficient of node R is $\rho_{RA} = \rho_{AR} - \rho_{A0} + \rho_{R0} \cong \rho_{A0}(1 - \rho_{R0})^{-1} - \rho_{A0} + \rho_{R0}$ (see [6] for more details).

Finally, we address the average packet delay of the originator A. Due to the self-similar nature of traffic, it is infeasible to use our previous approach [6] based on Pollazek-Khinchine formula, since it assumes Poisson arrival flow. However, taking into account Hurst parameter H_A, improves the analytical model [10]. For the cooperative system, the approximate mean packet delay of node A is given by:

$$\delta_A \cong \tau_{AR} + \frac{{\rho_{AR}}^{\frac{0.5}{1-H_A}}}{\tau_{AR}} \\ \times \frac{1}{2(1-\rho_{AR})^{\frac{H_A}{1-H_A}}} \left[\frac{X}{\tilde{p}_{CB}} \frac{2-\tilde{p}_{CB}}{\tilde{p}^2_{CB}} - \frac{Y}{\tilde{p}_A} \frac{2-\tilde{p}_A}{\tilde{p}^2_A} \right],$$

where $\tau_{AR} = \rho_{AR} {\lambda_A}^{-1}$, as well as the auxiliary variables are $X = \frac{p_{AR} P_S (1-\tilde{p}_{AB}) \tilde{p}_{CB}}{1-\tilde{p}_{CB}-(1-\tilde{p}_{AB})(1-p_{AR})}$ and $Y = X - \tilde{p}_{AB}$. Also for brevity $\tilde{p}_A = \tilde{p}_{AB} + p_{AR} P_S - \tilde{p}_{AB} p_{AR} P_S$.

The average packet delay of the relay R may be established similarly to the average delay of the originator A in the non-cooperative case (when $p_{AR} = 0$). Hurst parameter H_A could be estimated by a well-known procedure [12]. As expected, for the special case of $T_i^{ON} = 1$, the aggregated flow becomes Poisson (with $H_A = 0.5$) and our statistical tests (Pearson's chi-square and Kolmogorov-Smirnov) confirmed the exponential distribution of inter-arrival times. As such, the above formula for δ_A reduces to the respective expression from [6].

III. Performance Evaluation

A. Simulation Methodology

We use the extended system-level simulator described in [13] to verify the obtained analytical results. Partly following [11], the simulation parameters are set as: $p_{AB} = 0.3$, $p_{RB} = 0.7$, $p_{AR} = 0.4$, $p_{CB} = 0.5$, whereas λ_R is fixed to the moderate value of 30 packets per second. Additionally, we borrow power consumption values from [14] as: $P_{TX} = 1.65$ W, $P_{RX} = 1.40$ W, and $P_I = 1.15$ W. The aggregate traffic from a large number of meters N is simulated according to [9] (for $N = 1000$, 3000, 5000). We also assume realistic $T_i^{ON} = 1, 2, 3$ and take into account the following values of T_i as proposed in [15]: $T_1 = 900$ s, $T_2 = 300$ s, $T_3 = 60$ s, and $T_4 = 10$ s.

We set the number of source types $L = 7$. Summarizing, the periods are:

$$T_i = \{T_1, T_1, T_1, T_2, T_2, T_3, T_4\},$$

the lengths of ON-period are:

$$T_i^{ON} = \{3, 2, 1, 2, 1, 1, 1\},$$

and the type probabilities are:

$$\{p_i\}_{i=1}^{L} = \{0.20, 0.01, 0.01, 0.75, 0.01, 0.01, 0.01\}.$$

For a particular wireless system topology, we can estimate the number of meters by the given arrival rate or vice versa. Let us also fix the moderate arrival rate λ_A of 30 packets per second for the originator. For a numerical example, we estimate the number of sources using:

$$\lambda_A = \frac{N}{W} \sum_{i=1}^{L} \frac{T_i^{ON} p_i}{T_i},$$

where T_i^{ON}, p_i, and T_i are system topology parameters corresponding to the source types, W is the expectation of the specified type distribution. Since we consider $\alpha = 3$ and $\beta = 4$ as Beta distribution parameters from [9], $W = \frac{1}{2}$ for Uniform and $W = \frac{3}{7}$ for $Be(3, 4)$ distribution.

B. Simulation Results

In Figure 4a and Figure 5a, we fix the value of the arrival rate of $\lambda_A = 30$ packets per second and then vary the probability of collision $P = 1 - P_S$. In Figure 4b and Figure 5b, we vary the arrival rate or the number of data sources fixing the probability of collision to 0.1.

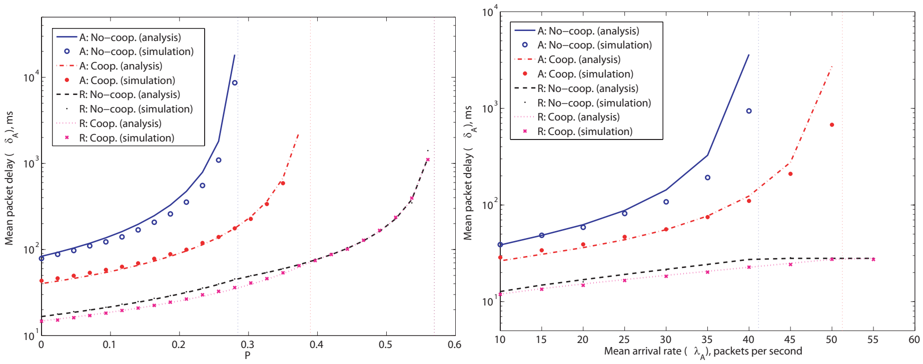
Figure 4a shows the increasing mean delay for packets at the originator and the relay due to growing probability of collision in the channel. Each value tends to the respective asymptote with the growth of the probability P. The asymptotes for A are 0.28 (non-cooperative) and 0.39 (cooperative). The asymptote for R is 0.57 (both non-cooperative and cooperative), as cooperation does not prevent R from transmitting own traffic.

Let us divide Figure 5a into four segments along the horizontal axis. The first segment [0, 0.28] contains four monotonically increasing functions (as the channel is getting highly populated, node A attempts to retransmit more and its expenditure increases). The second segment [0.28, 0.39] is explained by the fact that the contention success probability degrades and the originator has less chances to transmit than before. In the next segment, [0.39, 0.57], one can see the same situation for the relay-enhanced originator expenditure. Beyond the last asymptote for the relay, the system goes into full-buffer state. Interestingly, trends for the originator in both modes and for the relay in no-cooperation mode converge. It is the result of fair channel access between A and R.

Similarly, we may examine the dependence of the mean packet delay value (see Figure 5b) on the number of sources.

IV. Conclusion

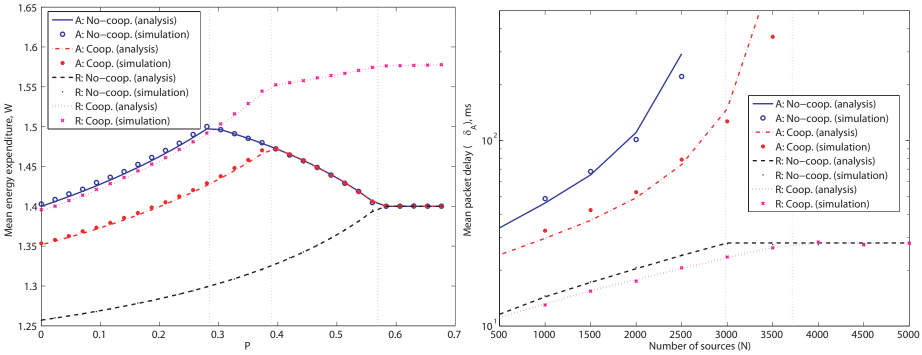
In this work, we proposed a simple client relay scheme to improve delay and energy efficiency of cell-edge M2M devices with poor communication link. Our analytical approach indicates significant performance gains that are verified by extensive simulations. It is expected that the novel scheme would become an important consideration for the future development of emerging IEEE 802.16p standard. In turn, its success is beneficial for smart metering market supported by international governmental organizations, utility companies, and equipment manufacturers.



(a) Mean packet delay vs. collision probability

(b) Mean packet delay vs. mean arrival rate

Fig. 4: Simulation results – I



(a) Mean energy expenditure vs. collision probability

(b) Mean packet delay vs. number of meters

Fig. 5: Simulation results – II

ACKNOWLEDGMENT

The authors are grateful to Dr. Nageen Himayat (Wireless Commun. Lab./Commun. Tech. Lab., Corp. Tech. Group, Intel Corp., Santa Clara, CA, USA) for valuable discussions and insightful comments that allowed to improve this paper.

REFERENCES

[1] H. Cho and J. Puthenkulam, *Machine to Machine (M2M) Communication Study Report, IEEE 802.16ppc-10/0002r6*, May 2010.

[2] Harbor Research Report, *Machine-To-Machine (M2M) & Smart Systems Forecast 2010-2014*, 2009.

[3] 3GPP Technical Report, *System Improvements for Machine-Type Communications (Release 10), TR 23.888*, July 2010.

[4] G. Miao, N. Himayat, and G. Y. Li, “Energy-efficient link adaptation in frequency-selective channels,” *IEEE Transactions on Communications*, vol. 58, no. 2, pp. 545–554, 2010.

[5] N. Himayat, S. Talwar, K. Johnsson, S. Andreev, O. Galinina, and A. Turlikov, *Proposed IEEE 802.16p Performance Requirements for Network Entry by Large Number of Devices, IEEE 802.16p-10/0006*, November 2010.

[6] S. Andreev, O. Galinina, and A. Vinel, “Performance evaluation of a three node client relay system,” *International Journal of Wireless Networks and Broadband Technologies*, vol. 1, pp. 73–84, 2011.

[7] *IEEE Std 802.16m (D9), Amendment to IEEE Standard for Local and metropolitan area networks. Advanced Air Interface*.

[8] R. Tannious and A. Nosratinia, “Spectrally efficient relay selection with limited feedback,” *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1419–1428, 2008.

[9] QUALCOMM Incorporated, *Simulation Assumptions for MTC and RACH Load Simulation Results for UMTS, R2-105619*, October 2010.

[10] I. Norros, “A storage model with self-similar input,” *Queueing Systems*, vol. 16, no. 3-4, pp. 387–396, 1994.

[11] B. Rong and A. Ephremides, “On opportunistic cooperation for improving the stability region with multipacket reception,” in *Proc. of the NET-COOP Conference*, pp. 45–59, 2009.

[12] M. S. Taqqu, V. Teverovsky, and W. Willinger, “Estimators for long-range dependence: An empirical study,” *Fractals*, vol. 3, pp. 785–798, 1995.

[13] A. Pyattaev, S. Andreev, Y. Koucheryavy, and D. Moltchanov, “Some modeling approaches for client relay networks,” in *Proc. of the IEEE CAMAD Workshop*, 2010.

[14] K. D. Turck, S. Andreev, S. D. Vuyst, D. Fiems, S. Wittevrongel, and H. Bruneel, “Performance of the IEEE 802.16e sleep mode mechanism in the presence of bidirectional traffic,” in *Proc. of the IEEE ICC Conference*, 2009.

[15] N. Himayat, K. Johnsson, S. Talwar, and X. Wang, *Functional Requirements for Network Entry and Random Access by Large Number of Devices, IEEE 802.16ppc-10/0049r1*, August 2010.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-2881-1
ISSN 1459-2045