FUTURE WIRELESS NETWORK ARCHITECTURE

by

Alia Asheralieva

A thesis by publication submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (PhD)

School of Electrical Engineering and Computer Science
Faculty of Engineering and Built Environment
University of Newcastle

August 2014

... to my best teacher, Prof Tapio Erke

DECLARATION

I hereby certify that this thesis is submitted in the form of a series of published papers of which I am a joint author. I have included as part of the thesis a written statement from each co-author; and endorsed by the Faculty Assistant Dean (Research Training), attesting to my contribution to the joint publications.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository[1], subject to the provisions of the Copyright Act 1968.

I hereby certify that the work embodied in this thesis contains a published and submitted for publication papers of which I am a joint author. I have included as part of the thesis a written statement, endorsed by my supervisor, attesting to my contribution to the joint publications.

Signature

Name

Date 21 August 2014

[1] Unless an Embargo has been approved for a determined period.

ACKNOWLEDGEMENTS

I would like to express my gratitude to all those who gave me the possibility to complete this thesis.

I am very grateful to Prof Jamil Yusuf Khan whose help, stimulating suggestions, knowledge, experience and encouragement helped me in all the times of study and analysis of the project.

Special thanks to my mother and Dr Kaushik Mahata whose strong believe was very meaningful for me during this research …

And to God, who made all things possible.

LIST OF PUBLICATIONS BY THE CANDIDATE

1. **A. Asheralieva**, J. Khan, K. Mahata and E.-Hw. Ong, "A Predictive Network Resource Allocation Technique for Cognitive Wireless Networks", in Proc. ICSPCS, Dec. 2010, pp. 1 – 9.
2. **A. Asheralieva**, J. Khan, K. Mahata, "Traffic Prediction Based Packet Transmission Priority Technique in an Infrastructure Wireless Network", in Proc. WCNC, Mar. 2011, pp. 404 – 409.
3. **A. Asheralieva**, J. Khan, K. Mahata, "Performance Analysis of VoIP Services on the LTE Network", in Proc. ATNAC, Nov. 2011, pp. 1 - 6.
4. **A. Asheralieva**, J. Khan, K. Mahata, "Performance of LTE for VoIP Users", International Journal of Internet Protocol Technology, Volume 7, Issue 1, 2012, pp. 3 – 14.
5. **A. Asheralieva**, J. Khan, K. Mahata, "Dynamic Resource Allocation in a LTE/WLAN Heterogeneous Network", in Proc. ICUMT, Oct. 2012, pp. 1 - 6.
6. **A. Asheralieva**, "Prediction Based Bandwidth Allocation for Cognitive LTE Network", in Proc. WCNC, Apr. 2013, pp. 801 - 806.
7. **A. Asheralieva** and K. Mahata, "Resource Allocation Algorithm for Cognitive Radio Network with Heterogeneous User Traffic", in Proc. GLOBECOM, Dec. 2013, pp. 4852 – 4857.
8. **A. Asheralieva** and K. Mahata, "A Two-Step Resource Allocation Procedure for LTE-based Cognitive Radio Network", Computer Networks, Vol. 59, 2014, pp. 137 – 152.
9. **A. Asheralieva** and K. Mahata, "Delay Aware Resource Allocation for Secondary Users in Cognitive LTE Network", to be published in Proc. MASS, Oct. 2014.

LIST OF ABBREVIATIONS AND ACRONYMS

3GPP	- 3d Generation Partnership Project
ACK	- ACKnowledgement
AIC	- Akaike Infromation Criterion
AMC	- Adaptive Modulation and Coding
AP	- Access Point
AR	- Autoregressive
ARIMA	- Autoregressive Integrated Moving Average
ARMA	- Autoregressive Moving Average
ARQ	- Automatic Repeat reQuest
AWGN	- Additive White Gaussian Noise
B3G	- Beyond 3rd Generation
BER	- Bit Error Ratio
BPSK	- Binary Phase Shift Keying
BS	- Base Station
BSR	- Buffer Status Report
CC	- Chase Combining
CCE	- Control Channel Element
CCH	- Control CHannel
CDMA	- Code Division Multiple Access
CR	- Cognitive Radio
CRC	- Cyclic Redundancy Check
CRN	- Cognitive Radio Network
CSMA/CA	- Carrier Sense Multiple Access with Collision Avoidance
DCF	- Distributed Coordination Function
DL	- DownLink
DL-SCH	- Downlink Shared CHannel
DSA	- Dynamic Spectrum Access
eNB	- Evolved Node B
EPC	- Evolved Packet Core
EPC	- Evolved Packet System
E-UTRA	- Evolved Universal Terrestrial Radio Access
E-UTRAN	- Evolved Universal Terrestrial Radio Access Network
FARIMA	- Fractional Autoregressive Integrated Moving Average
FD	- Fully Dynamic

FDD	- Frequency Division Duplex
FEC	- Forward Error Correction
FIFO	- First-In-First-Out
FTP	- File Transfer Protocol
GERAN	- Groupe spécial mobile Enhanced data rates Radio Access Network
GSM	- Groupe Spécial Mobile
GPRS	- General Packet Radio Service
GW	- GateWay
HARQ	- Hybrid Automatic Repeat reQuest
HCF	- Hybrid Coordination Function
HSPA	- High Speed Packet Access
HTTP	- HyperText Transfer Protocol
IEEE	- Institute of Electrical and Electronics Engineers
IR	- Incremental Redundancy
IP	- Internet Protocol
ITU	- International Telecommunication Union
ITU-T	- International Telecommunication Union – Telecommunication Standardization Sector
L1	- Layer 1
L2	- Layer 2
LTE	- Long-Term Evolution
LTE-A	- Long-Term Evolution Advanced
MA	- Moving Average
MAC	- Medium Access Control
MCS	- Modulation and Coding Scheme
MIMO	- Multiple Input Multiple Output
MME	- Mobility Management Entity
NACK	- Negative ACKnowledgement
NMSE	- Normalized Mean Squared Error
OFDM	- Orthogonal Frequency Division Mode
OFDMA	- Orthogonal Frequency Division Multiple Access
PDCCH	- Physical Downlink Control CHannel
PDCP	- Packet Data Convergence Protocol
p.d.f.	- probability density function
PDSCH	- Physical Downlink Shared Channel
PER	- Packet Error Ratio
PER	- Prediction Error Ratio

PF	- Proportional Fair
P-GW	- Packet-data network GateWay
PHY	- PHYsical
PLR	- Pseudo-Linear Regression
PS	- Packet Scheduler
PU	- Primary User
PUCCH	- Physical Uplink Control Channel
QAM	- Quadrature Amplitude Modulation
QoS	- Quality of Service
QPSK	- Quadrature Phase Shift Keying
RA	- Random Access
RACH	- Random Access Channel procedure
RAN	- Radio Access Network
RAT	- Radio Access Technology
RB	- Resource Block
RBP	- Resource Block Pair
RIV	- Recursive Instrumental Variable
RLC	- Radio Link Control
RLS	- Recursive Least Squares
RPEM	- Recursive Prediction Error Method
RR	- Round Robin
RRC	- Radio Resource Control
RTP	- Real-time Transport Protocol
Rx	- Receiver
SAE	- System Architecture Evolution
SC-FDMA	- Single Carrier Frequency Division Multiple Access
SCM	- Spatial Channel Model
SDP	- Session Description Protocol
SDR	- Software Defined Radio
SE	- Spectral Efficiency
S-GW	- Serving GateWay
SINR	- Signal-to-Interference-and-Noise Ratio
SIP	- Session Initiation Protocol
SJF	- Shortest-Job-First
SMP	- SeMi-Persistent
SNR	- Signal-to-Noise Ratio
SP	- Service Provider
st.dev.	- standard deviation

SU	- Secondary User
TCP	- Transmission Control Protocol
TTI	- Transmission Time Interval
Tx	- Transmitter
UDP	- User Datagram Protocol
UE	- User Equipment
UL	- UpLink
UL-SCH	- Uplink Shared CHannel
UMTS	- Universal Mobile Telecommunications System
UTRAN	- Universal Terrestrial Radio Access Network
VoIP	- Voice over Internet Protocol
WiMAX	- Worldwide interoperability for Microwave Access
WLAN	- Wireless Local Area Network
WRAN	- Wireless Regional Area Network

TABLE OF CONTENTS

ABSTRACT

With widespread use of wireless networks and the emergence of multiple Radio Access Technologies (RATs), the present-day network architecture is currently being transformed into the one global infrastructure vision, called Beyond 3rd Generation (B3G) [1]. B3G is a heterogeneous Internet Protocol (IP) based wireless access infrastructure, which aims to provide higher capacity and quality of service (QoS) to the users even considering the limited radio spectrum through support of a cooperative diversity [2] and reconfigurability [3].

In a system with a cooperative diversity each node in the network can act both as an information source and a relay. Such information relay may increase the capacity and diversity gain in wireless networks, leading to the improved performance in terms of both area coverage and QoS [4]. In B3G the cooperative communication assumes that the network infrastructure will rely on more than one RAT: depending on encountered specific conditions (e.g., hot-spot requirements, traffic demands, etc.) at different times in different areas the RATs will cooperate with each other to achieve the maximization of QoS levels offered to users. To support the cooperative communications in B3G, the advanced management functionality is required to deal with the reallocation of traffic to different RATs and sub-networks, as well as the mapping of applications to QoS levels [5-8].

The move towards the reconfigurability concept was initiated by the development of the Cognitive Radio Network (CRN) – the network, where the nodes with fixed licensed spectrum (so-called primary nodes) can share their spectrum resources with nodes without fixed licensed spectrum (secondary nodes) [9]. In B3G the reconfigurability aims to provide essential mechanisms for terminals and sub-networks, to enable them to adapt dynamically and transparently to the most appropriate RAT depending on encountered situation (hot-spot requirements, traffic demands, etc.). The reconfigurability allows for the dynamic allocation of resources (such as bandwidth, service rate, etc.) to RATs, and invokes a variety of new possibilities with respect to the more efficient utilization of available spectrum [1, 9-10].

With regard to the diverse challenges arising upon the development and deployment of B3G, this thesis aims to:

1. explore the potential ways of implementing the future wireless infrastucture based on existing wireless networking standards and co-existance of air such features ;
2. study the main principles of cooperative and cognitive communication which lie in:
   (a) cooperation and information exchange between all member subnetworks;
   (b) support of reconfiguration capabilities of all nodes/user terminals within the network;
   (c) coexistence of the nodes/user terminals belonging to different RATs comprising the network ;
   (d) intelligent resource planning involving cognitive reactive and proactive management of the network resources based on external (environmental) aspects, as well as on goals, capabilities, experience and knowledge.
3. develop the efficient radio resource management platform in order to provide increased spectrum utilization and enhanced end-to-end QoS for users of different RATs with and without fixed spectrum allocation.
4. investigate the problems of co-existance, intra- and cross-layer control between different RATs comprising the network, including:
   (e) PHY layer channel modeling, including noise and interference models, log-distance path loss, shadow and multipath induced fading, physical layer transmission techniques (MCS, AMC);
   (f) MAC/RLC layer design, including traffic generation models, packet scheduling, ARQ/HARQ, DCF/HCF, buffer status reporting, etc.;
   (g) Cross-layer control: necessary parameters (such as packet arrival rate, buffer occupancy, SINR) are observed on MAC and PHY; control of available resources (such as bandwidth, data rate, buffer capacity) on PHY layer;
   (h) Application layer QoS for users as a result of undertaken control on PHY/MAC layer.

OVERVIEW

1 Introduction

With widespread use of wireless services the wireless networking design paradigm is currently being transformed into the one global infrastructure vision, called Beyond 3rd Generation (B3G) [1, 101]. In future, the wireless services will be provided using a multiple number of radio access technologies (RATs) rather than using a single standard wireless network [102, 103]. The emergence of software defined radio (SDR) [104] will allow the customers to connect to any network based on the capacity and quality of service (QoS) requirements [102]. In B3G, various access points (APs) and base stations (BSs) belonging to different RATs will connect to the spectrum manager (SM) using an internet protocol (IP) network. Wireless users with their SDR terminals will connect to any of the APs/BSs within their coverage area (Figure 1) [102].

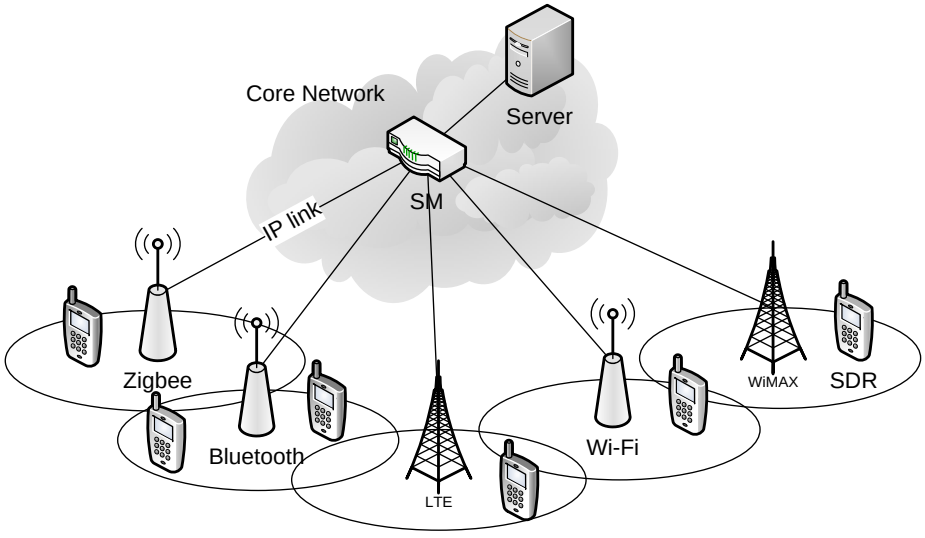


Fig. 1. Future heterogeneous cognitive wireless network architecture

Recall, that the traditional concept of cognitive user behavior has been formulated as follows. Each unlicensed (secondary) node senses the spectrum to find (on its own) the available unused bandwidth

(spectrum hole) from some primary node, and utilizes it according to the requirements of this primary node. At any time the connection of the secondary node can be blocked by the primary node (which usually happens in case if the bandwidth used by the primary node at the current state is not enough to satisfy the requirements of the primary node exceed the contemporary level) [33-34, 40-50]. Such concept has many disadvantages mainly because of the absence of cooperation and information exchange between all member nodes. As a result, secondary nodes will spend more time on unnecessary channel sensing and competing for access to the licensed spectrum bands, which eventually will lead to connection loss, poor service quality and increased power consumption of the user terminals.

In contrast to this concept, in B3G the wireless services will be provided using a cooperative approach [2] in which all member nodes will cooperate with each other by exchanging their network status information and sharing available capacity in an orderly manner. Information exchange and coordination between the nodes will allow maximizing the overall capacity and QoS of the network. In this way secondary nodes will be able to (temporarily) borrow network resources in a more efficient way, and minimize the loss of connection. However, to support the cooperation, the advanced management functionality is required to deal with the reallocation of traffic to different RATs and sub-networks, as well as the mapping of applications to QoS levels [5-8].

In this research project we explore the potential ways of implementing the main principles of cooperative and cognitive communication in existing wireless networking standards, 3GPP LTE [69] and IEEE 802.11 (Wi-Fi) [70]. These principles lie in the effective management of the available resources, i.e. (i) more efficient utilization of available spectrum, (ii) improved end-to-end QoS for users of different RATs with and without fixed spectrum allocation, and (iii) an intelligent network planning process.

Thesis makes of the following major research contributions:

1. Three most common scenarios of cognitive network behavior in the network architecture illustrated on Figure 1 have been investigated:
   (i) In the first scenario all users, sub-networks and RATs have equal priorities in accessing the network resources. The network resources, represented by the total available bandwidth are shared

according to some predefined flexible spectrum usage policy. This network deployment scenarios has been proposed by the IEEE 802.22 working group in November 2004 [51, 52].

(j) In the second scenario the network comprises a number of licensed (primary) and unlicensed (secondary) APs/BSs. Primary APs/BSs operate on their licensed spectrum bands (primary channels), whereas secondary APs/BSs don't have fixed licensed spectrum. Each primary BS can share its channel with one or more secondary BSs. In this case the primary station is given a prioritized access to its licensed spectrum band, whereas the secondary stations are served on the best-effort (non-prioritized) basis. Similar network deployment scenarios have been considered in [83 - 89].

(k) In the third scenario a network comprises a number of BSs operating on their fixed licensed spectrum bands. Each BS serves a number of primary and secondary users. Primary users (PUs) are the licensed network users who pay some price for accessing the wireless services, and therefore have priority in accessing the spectrum. Secondary users (SUs) are unlicensed network users who can access the wireless services for free on best-effort basis. Similar scenarios of the network deployment have been considered in [91, 92].

2. To increase the efficiency of resource allocation, prevent potential network congestions, decrease packet delay and connection loss for the wireless users, we deploy traffic prediction in the considered network architecture. Considering the known difficulty of parameter estimation for time-varying wireless channels and heterogeneous nature of the wireless traffic, comprising large number of different network applications (such as data, voice or video), we propose to use recursive estimation techniques applied with time-series models for traffic prediction. Unlike off-line estimation methods, these techniques do not require a long observation history, are highly adaptive and have modest memory requirements [106].
3. Different algorithms for resource allocation in cognitive wireless network architecture have been proposed. The objectives and the complexity of these algorithms vary depending on the considered cognitive user behavior and network deployment scenario. Most of these algorithms use a so-called cognitive cycle for resource

allocation, consisting of observation, information gathering and traffic prediction. Proposed algorithms have been implemented in cognitive network architecture based on IEEE 802.11g (Wi-Fi) and 3GPP LTE standard networks. The algorithms efficiency has been evaluated using simulations conducted in OPNET platform [112].

This thesis is organized as follows. Further in the Overview we review the latest research in the area of cognitive and cooperative networking in the section entitled Literature Review. The summary of main thesis contributions is provided in the Research Methodology section.

Parameter estimation techniques and the time-series models proposed for traffic prediction in CRN are described in Chapter 1, which also provides the performance of these techniques based on results published in Proceedings of IEEE International Conference on Signal Processing and Communication Systems (ICSPCS) in 2010.

A priority based packet transmission technique for an infrastructure based network WLAN is presented in Chapter 2. This technique can be used (together with one of the proposed resource allocation algorithms) to support future wireless network infrastructure to improve the capacity and the QoS for the users under all network deployment scenarios considered in the thesis. The corresponding paper has been published in Proceedings of IEEE Wireless Communications and Network-ing Conference (WCNC) in 2011.

Chapter 3 is based on the contributions in two papers on the analysis of VoIP services performance in LTE network. The first paper who published in Proceedings of Australasian Telecommunication Networks and Applications Conference (ATNAC) in 2012. The other paper appeares in the International Journal of Internet Protocol Technology (IJIPT) in 2012. Performance analysis of LTE is a logical first step toward deployment of this network in B3G infrastructure. Performance analysis of LTE is carried out using VoIP user applications. VoIP services have the strictest (compared to other network applications) delay requirements: VoIP can only tolerate packet end-to-end delay of up to 100 ms and packet loss of up to 1% [116]. Thus, the ability of LTE to achieve good performance for voice applications will automatically guarantee a satisfactory QoS for other user applications.

In Chapter 4 a resource allocation technique for LTE-based CRN in the first scenario is presented. Here we outline the algorithm for

resource allocation, show its implementation in LTE-based network infrastructure and analyze its performance based on some simulations in OPNET environment [112]. The corresponding paper appears in Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), 2013.

Chapter 5 is based on a paper published in Proceedings of IEEE International Congress on Ultra-Modern Telecommunications and Control Systems (ICUMT) in 2012. In this chapter a resource allocation technique for a combined LTE/WLAN CRN in the first scenario is presented. Here we focus on some specific challenges of resource allocation in the complex networks comprising more than one RAT propose the algorithm for spectrum access in combined LTE/WLAN architecture, and evaluate its performance based on results of simulations in OPNET environment [112].

In Chapter 6 we present a novel approach for resource allocation in cognitive LTE network in the first scenario, derive a resource allocation algorithm, and present the results of algorithm performance based on simulation model developed in OPNET environment [112]. The corresponding paper is published in Proceedings of IEEE GLOBal COMmunications Conference (GLOBECOM) in 2013.

A resource allocation technique for a cognitive LTE network in the second scenario is described in Chapter 7. The corresponding paper is published in Elservier Computer Networks Journal in 2014.

In Chapter 8 the spectrum access algorithm for LTE-based CRN is the third scenario is summarized. The corresponding paper will be published in Proceedings of IEEE International Conference on Mobile Ad hoc and Sensor Systems (MASS) in 2014.

The conluding remarks describing the practical implication of the proposed network architectire and resource management platform are provided in Conclusions.

2 Literature Review

Existing literature in the area of cooperative and cognitive networking can be arbitrary divided into three large groups. The first group of papers combines some cooperative coding methods to realize the cooperative diversity (for instance, [12, 16-18]), and various lower layer techniques for spectrum sensing and spectrum mobility (e.g., [27

– 30, 177 - 179]). The second group of papers investigate analytical models of the user behavior and traffic load in cooperative and cognitive radio networks using either game-theoretic approach or some results in queuing theory (examples are [20, 21, 32, 40, 41]). The third group of papers examine application of different resource allocation techniques for cognitive access in OFDMA-based networks (e.g., [19, 22, 23, 45 – 47, 50]).

The studies on cooperative coding [12, 16-18] are mainly focused on physical layer signal processing and coding for cooperative networks. The cooperative system is represented by a multiple-input-multiple-output (MIMO) system, formed by multiple source and relay antennas. All relaying methods use the same general procedure, in which a cooperation cycle consists of two phases. In the first phase, each user transmits parts of its own data and receives data from the other users. In the second phase, the users help each other by relaying the data they received in the first phase. Various cooperative diversity schemes define different ways of performing the second phase and representing the data of partner. Classification of existing relaying approaches is provided in Table 1 [4, 15].

Table 1. Classification of Existing Relaying Approaches [4, 15]

Approach	*Data Regeneration*	*Cooperative Diversity*	*Coding Scheme*
Store & Forward (S&F)	Yes	No	-
Amplify & Forward (A&F)	No	Yes	-
Compress & Forward (C&F)	No	Yes	Compression
Decode & Forward (D&F)	Yes	Yes	Repetition
Coded Cooperation (CC)	Yes	Yes	Forward Error Correction (FEC)
Space Time Coded Cooperation (STCC)	Yes	Yes	Space-time & FEC

Good examples of spectrum sensing and spectrum mobility techniques for cognitive radio networks can be found in [177 - 179]. The opportunistic spectrum access (OSA) for interference minimization has been investigated in [177], where it has been shown that the implementation of OSA enhances the overall system performance by intelligent aggregation of the unutilized spectrum. Relay selection and resource allocation in cognitive relay network has been studied in [178]. It has been assumed that the primary stations communicate via a relay assisted network, some of the secondary stations play the role of the network relays, and the remaining nodes interact using a centralized algorithm in the licensed spectrum band. Simulation results have demonstrated that the proposed resource allocation algorithm shows increased throughput compared to the conventional random relay selection and uniform power allocation method. A cross-layer protocol for spectrum mobility and handover in cognitive networks has been presented in [179]. This protocol assumes a Poisson distributed model for the spectrum resources. An empirical performance study has illustrated that the proposed hand-off protocol significantly reduces the expected transmission time and the spectrum mobility ratio within the network.

All above techniques are very effective for coded cooperation between the users, and identifying and reducing the interference in the physical channels, but do not improve the overall user-perceived quality of service (QoS) which is mainly expressed in terms of the packet end-to-end delay and loss for the network users. The results of these works can be applied only in combination with other techniques working on MAC and higher layers to provide the reduced delay and loss for the end-users [26]. Therefore, we will further focus only on the second and third group of papers where the implementation of cooperative and cognitive radio networks is investigated using network layer models to maintain qualitative service performance for the users.

2.1 Cooperative Networks

The concept of a cooperative diversity was first introduced by the works of Van der Meulen [11], Cover and El Gamal [12], and Gallager [13] on the relay channel. Relay channel is the simplest scenario of user cooperation in which a nearby terminal (called relay or partner)

forwards information from a source to the destination. Based on the works on relay channel, Sendonaris et. al. proposed a user cooperation diversity, in which the users were allowed to share their resources by acting both as a source and as a relay [14]. In a cooperation diversity, each user is represented by a distributed multiple antenna system. In contrast to relaying, the information from one source is forwarded via multiple channels between these antennas. The work of Sendonaris et al. attracted a lot of interest to the cooperative diversity, and lead to the development of a new research area called cooperative networking.

In cooperative networking resource allocation and partner selection for cooperative users (i.e. selection of a best relay, called "partner") is made based on certain optimization criterion. It is assumed that cooperative coding schemes are integrated into wireless networks to optimize the service performance of various multi-user systems. Clearly, the optimization objective depends on scenario, factors and system parameters of the optimization problem. Moreover, since it is not possible to consider all network parameters, only those scenario factor/system parameters which can be observed/modified at the corresponding node and related to the corresponding layer are useful for optimization. The classification of possible observable scenario factors (called observed parameters) and controllable system parameters (called controlled parameters) is shown on Figure 2. In optimization the observed parameters are used to select the optimal values of controlled parameters based on certain optimization objective. Consequently, the literature on resource allocation and partner selection techniques for cooperative users can also be classified according to the Figure 2.

Most existing works on resource allocation for cooperative networks focus on various issues at the physical layer, where the advantages are often demonstrated using some information-theoretic approach. The only relevant papers are [19-24], where the MAC and higher layer issues of QoS provisioning in cooperative networks has been carefully addressed. Shan et al. [19] investigated the influence and integration of physical-layer cooperation with the MAC layer to increase the throughput and reliability of communication, and proposed a cross-layer design for a cooperative MAC protocol. With channel and payload length adaptation, this protocol was used to support multiple

transmission rates and transmission modes, and outperformed the traditional non-cooperative MACs.

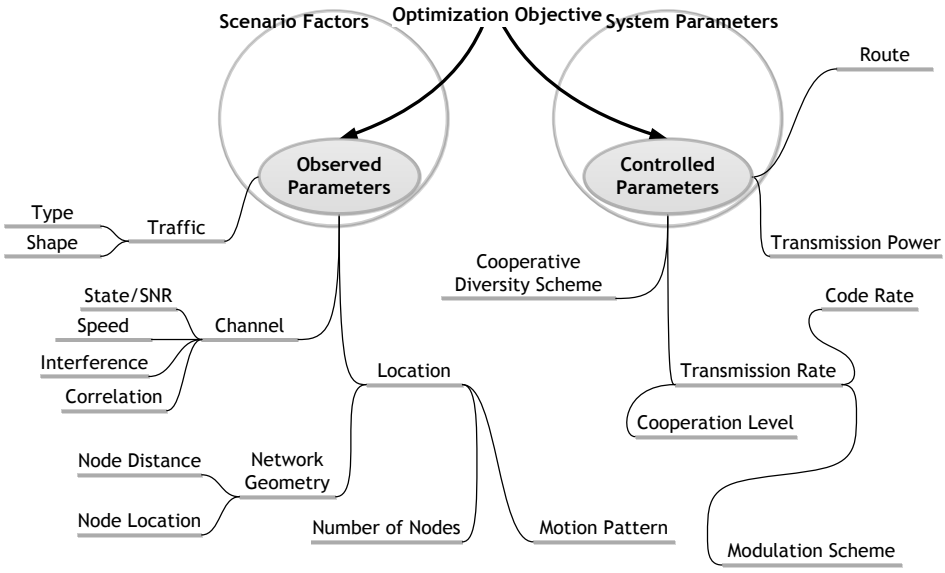


Fig. 2. Classification of possible resource allocation strategies

A game-theoretic approach to solve the cooperation problems was applied in [20-21]. In [20] the Zhang et al. have analyzed the cooperative behavior of the nodes in a wireless network, and presented a cooperation bandwidth allocation strategy based on the Nash bargaining solution to solve two basic problems - when to cooperate and how to cooperate. Using simulations, the authors demonstrated that when cooperation takes place, users benefit from the proposed strategy in terms of utility, and those with longer distance to the access point (AP) should spend more bandwidth to cooperate with others. In [21] Huang et al. proposed two distributed algorithms, with the signal-to-noise ratio (SNR) auction and the power auction, to determine relay selection and relay power allocation. It was shown that the power auction achieves the efficient allocation by maximizing the total rate increase, and the SNR auction is flexible in trading off fairness and efficiency. Authors also have shown the convergence of both algorithms to the unique Nash Equilibrium, and verified their effectiveness and robustness.

The problem of partner selection for non-altruistic node cooperation was studied in [22-23]: in [22] the authors proposed three schemes of partner selection with power control to balance the transmit power and system performance; in [23] the partner selection schemes were deployed to minimize the average outage probability. However, the MAC-layer service differentiation was not addressed in these papers. In [24], Zhang et al. proposed a simple two-step scheme for the system throughput maximization problem with physical-layer QoS assurance. However, the application layer QoS support was not taken into consideration.

2.2 Cognitive Radio Networks

User cooperation can increase the capacity of the radio networks to a certain extent, but cannot fully solve the problem of spectrum scarcity introduced by the traditional fixed bandwidth allocation strategy. To deal with this issue, a new spectrum access concept, called Cognitive Radio Network (CRN), has been proposed in the pioneering work of Simon Haykin [25]. In CRN available spectrum can be shared among the users. In this context, CRN is defined as an intelligent system that has the ability to perceive its environment, and then learn and adapt to the current network conditions [25, 26] with three potential tasks:

- to sense the spectrum and model users' behaviour;
- to manage and share available radio resources;
- to maintain qualitative service performance for users during the channel transition (end-to-end QoS) [25, 26].

Prior research works in CRNs (for example, [27-30]) were mainly focused on finding efficient ways to sense primary users, and let secondary users access the spectrum with minimal interference to the primary users. Even though this framework on physical layer is very important, the goals of CRN cannot be fully achieved without information gathering, processing, and control on higher layers [31]. Therefore, we will further review only the works where MAC and higher layer issues have been considered.

Spectrum sharing enables the secondary users (SUs) to access the licensed band in a primary system without its modification [33]. In this case actions of SU are transparent for the primary system. The SU should vacate the frequency in primary system on arrival of the PU, and scan the frequency range to detect available band at the same time (spectrum handoff procedure). Centralized solutions for spectrum sharing problem mainly focused on centralized solutions ([34-35]) which should be avoided due to the non-centralized nature of wireless networks and potential propagation delays ([36-37]). Decentralized solutions were proposed in [38-39], but only for a homogenous environment.

Different spectrum access schemes in open spectrum wireless networks were investigated in [32]. Two different types of radio systems with different channel requirements (3 and 9 overlapping frequency bands respectively) are assumed to operate in the same band. The offered traffic is modelled with two random processes per radio system: Poisson arrival traffic and negative-exponentially distributed radio system access. The authors use average airtime (which is referred as a ratio of allocation time to a certain reference time) per radio system to evaluate the fairness and blocking to evaluate efficiency. Applying these two parameters, they compare different spectrum sharing models: equal traffic without queuing, equal traffic with queuing, general traffic load, and random access models, and show that random access model achieves near-optimal weighted fairness. In the random model proposed by the authors radio systems do not access the unlicensed band in a greedy manner, but with some certain probabilities. Even though this model shows satisfactory parameters, it can be applied only for unlicensed band, and not applicable in the case when SUs entering the primary system.

In another work the priority of primary users (PUs) is considered [40]. Instead of blocking, the authors introduce a finite queue to store the SU requests on arrival. Spectrum handoff is regarded as a priority policy for SU. Additionally, they develop a Markov approach to analyse the proposed spectrum sharing policies (with and without buffering) with generalized bandwidth size in both primary and secondary systems, and evaluate performance for SU using blocking, interrupted, forced termination, non-completion probability and waiting

time. Simulation results had shown that the buffer is able to significantly reduce the SU blocking and non-completion probability. Therefore buffering proposed in this work may be one possible solution of the spectrum sharing problem, but can be prohibitive for delay-sensitive applications, and does not consider fairness.

Unlike the work presented above, in [41] the authors investigate the aggregate throughput and proportional fairness of two independent secondary user groups in CRN. The PU spectrum is completely shared among two SUs groups. On the arrival of a PU connection, the interrupted SUs can move to other vacant subchannels (spectrum handoff). The new SU connections from each SU group are served on the First-In-First-Out (FIFO) basis. Using a continuous time Markov chain, the authors determine the individual and combined forced termination, blocking probabilities and aggregate throughput. The results show that fairness is ideal in case of equal service rates of SUs groups, and deviates from its ideal value as the different between service rates increases. For the fixed arrival rate of SU groups there exists an optimal service rate pair which maximizes the aggregate throughput. The latter conclusion can be possibly applied to solve the spectrum management problem for different groups of SUs.

The most extensive investigation is presented in [26] where the authors develop a dynamic channel-selection scheme for autonomous wireless users transmitting delay-sensitive applications over the CRN. Considering heterogeneous network where users might have different channel requirements the authors propose a novel priority virtual queue interface that determines the required information exchanges and evaluate the expected delays. PU has priority to preempt the transmission of the SUs, while SUs are divided into groups with different pre-assigned priorities (depending on their QoS requirements and right to access the channel). Each SU may have one of the following two types of utility functions: the delay-based utility (for delay-sensitive applications), and the throughput-based utility (for delay-insensitive applications). The goal of each SU is to maximize his utility function. Modelling dynamic resource management problem as a multi-agent interaction, the authors present the following resource management scheme: (i) SU collect information from other SUs through the priority virtual queue interface; (ii) interface estimates a channel selection strategy of the SU, and performs priority queuing

analysis based to evaluate expected utility function; (iii) based on utility function, SU adapts its strategy; (iv) SU assigns each packet an action, selecting frequency channel; (v) the packet waits in queue to be transmitted. Simulation shows that proposed solution significantly reduces the packet loss rate and outperforms the conventional single-channel dynamic resource allocation by almost 2 dB in terms of video quality. The results achieved in this work show that the proposed method to solve the spectrum sharing problem as a multi-agent interaction can be extended and applied in the future development of the spectrum sharing policy in CRN.

The Architecture and End-to-End QoS for the Users of CRN

The final goal of any research investigation in most cases will be the achievement of satisfactory results and practical solutions applicable in "real world". In CRN this goal is one of the most challenging because of the complexity of this network. Therefore, development of clear architecture and internetworking mechanisms that will let to obtain guaranteed end-to-end QoS and implement this architecture in practice, is very important. Even though there exists several framework proposals for Cognitive Radio Architecture (for example, [25, 42-43]), most of them narrowly focused on some certain tasks, and lack a complete perspective on the problem. In this section the most relevant works in this area will be presented.

In [31] the authors propose the architecture of the CogNet System following a fully distributed cognitive networking approach. This architecture presents a cognitive function through CogNet nodes collaboration. In each layer (from physical to application) it has a cognitive agent (CA) gathering information and controlling protocol parameters within that layer. Information and data exchange is carried out through an internal CogNet Cross layer Bus (CogBus). Coordination of CAs within one node and coordination with other nodes is carried by CogPlane lying across all layers. CogPlane contains a module called Cognitive Executive Function (CEF) that helps to understand data interaction and performance models across the layers and develop user requirement translation. Once an end-to-end user goal is defined, CEF is responsible for translating the end-goals to executable action items required for each layer. The CogPlane provides

an opportunity for dynamic resource allocation and management with the help of the past history of the user, the device and network. CogNet Access Point (AP) is an autonomous form of the CogNet architecture. It has the capability of using higher layer traffic information for efficient management of the network resources. AP has two interfaces: the service interface to provide network services to the users or client nodes which are associated to this AP, and the monitoring interface for constant monitoring of the channels. CogNet AP builds statistical models based on its traffic observation. These models (MAC and higher layers) are built from the packets received at the CogPlane, and exploit the temporal behavior of the network traffic activity in the given area. The proposed CogNet AP has been deployed in two test-bed environments: residential and office. The experiments had shown that the throughput achieved by end-user devices provided improvement of about 5-10% and 10-15%, which shows that this architecture can be deployed for future research. Unfortunately, the results were derived for the system operating in a homogenous environment, and performance of the proposed model in real heterogeneous conditions was not measured [31].

A component-based approach to construction of the control and management in CRN is applied in [44]. The proposed architecture consists of: 1) Cognitive Resource Management (CRM) core composed of a set of a pluggable behaviour components, 2) generic interfaces components to support transparent access to underlying system, 3) distributed control and coordination modules, and 4) a policy engine. Components are encapsulated units of functionality and deployment, which may interact only through well defined interfaces and receptacles. The CRM core provides the administrative framework to coordinate the construction of the system, and the control framework to coordinate the actions of the components. The control framework is composed of behaviours, generic interfaces, action brokers, action resolvers, distributed control components and a police manager. The listed components are described as follows:

- A behaviour is simply a unit of goal directed functionality, which has the following properties: i) it collects some input for processing, ii) it evaluates the input(s) collected, and iii) it suggests some action(s) to

perform based on evaluation. Therefore, the behaviour is simply a coupling of some set of input, processing, and output mechanisms.

- Generic interfaces abstract the varied and complex implementation of various protocol layers and present them in a consistent and systematic manner. The authors use Universal Link Layer API (ULLA) for radio and link level, GENI – for transport and network layers, and Common Application Requirement Interface (CAPRI) to interface between applications and the CRM.
- Action brokers collect together sets of dependent behaviours and provide a means to choose between their proposed actions – in effect the behaviours dispatch their decisions to action brokers.
- Action resolvers are contained within the action brokers and implement the resolution method to be used by the broker. In effect action resolvers define how a broker chooses between the candidate actions represented by the behaviours it groups.
- Distributed control and coordination components support provide the mechanism and abstractions to support control and coordination in a transparent manner.
- Policies are required to provide operation constraints to the CRM, and they may be static or dynamic with respect to time and geographical location.

In the paper the authors presented an example of cross-layer application aiming to jointly optimise video and MAC parameters in order to maximize the video throughput. This early prototype was implemented as a part of ARAGORN project so as to test architectural primitives. Unfortunately, clear results and conclusions about the effectiveness of this system are not shown in the paper, and the performance of the proposed architecture is not known. Nevertheless, the component-based approach applied in the paper worse to be mentioned, because it might help to solve the fundamental design problems arising due to the large numbers of involved hardware and software modules [44].

Nowadays, there is little research in the area of QoS provisioning in CRN. Some papers use cognitive probe packets to guarantee end-to-end QoS. This can increase network load and affect capacity of the network [45-47]. The others propose cognitive algorithms that can solve the problem but are not very effective at the current moment [48-49]. The

most relevant investigation in this area was conducted in [50], where the authors present a QoS provisioning scheme. Based on the feedback on current network conditions, this scheme finds and even predicts the bottleneck of the network performance, takes some reactions in advance, and provides end-to-end QoS guarantee. The proposed architecture of CRN in [50] is the following: all network elements (mobile terminals, APs, routers etc.) are classified as reconfigurable nodes (RNs) (terminals of different access networks) and cognitive intelligence nodes (CINs) (other network elements able to control and manage RNs). The RNs are reconfigurable and equipped with cognitive radio technology which has the properties of sensing, awareness, adaptation, and learning capabilities. Cognitive core network is composed of heterogeneous CINs located in different places. According to the network conditions reported by RNs, the CINs orient the priority attached to any observation, plan action(s), and decide the most appropriate modification(s). After a modification, the actions and corresponding changes of the overall network are stored in repository and can be used as reference for future. The QoS components have the function of QoS measurement, analysis, adaptation, and feedback. According to the division of heterogeneous network, the end-to-end QoS can also be divided into several cognitive QoS components. The cognitive QoS components can measure QoS performance in local domain. QoS analysis is performed to analyse the conditions and make decisions based on the observed service requirements and network status. QoS downgrade events can be found/ predicted in partial or end-to-end path. In this case QoS feedback notifies QoS components about the downgrade event, and components cooperate with others to adapt QoS policy based on the current network status and previous stored experience. Besides, each QoS component can negotiate SLAs with the neighbouring domain, map QoS policies, and execute traffic conditioning as the non-cognitive QoS components. To integrate each host on the end-to-end path the QoS provisioning model uses end-to-end signalling supported by Next Steps in Signalling (NSIS) mechanism. The reason behind choosing NSIS for signalling is that it can configure resource for services with high priprity of QoS requirements in advance. However, NSIS is still under research, and is a big challenge for heterogeneous CRN. The authors also mention that although QoS guarantee schemes for CRN has been proposed, the end-

to-end QoS provisioning for cognitive heterogeneous network is still in blank. More problems are still in the future, such as intelligent and effective CRN discovery, or QoS routing mechanism using cognitive algorithms [50].

3 Research Methodology

3.1 Research Objective and Goals

The objective of this research project is development and implementation of heterogeneous IP based wireless access infrastructure upon the existing radio standards with the objective to provide higher capacity and QoS to the network users through the support of cooperative and cognitive functionalities.

In this context, this research project aims to:

2. explore the potential ways of implementing the future wireless infrastucture based on existing wireless networking standards;
2. study the main principles of cooperative and cognitive communication which lie in:
   (a) cooperation and information exchange between all member subnetworks;
   (b) support of reconfiguration capabilities of all nodes/user terminals within the network;
   (c) coexistence of the nodes/user terminals belonging to different RATs comprising the network;
   (d) intelligent resource planning involving cognitive reactive and proactive management of the network resources based on external (environmental) aspects, as well as on goals, capabilities, experience and knowledge.
3. develop the efficient radio resource management platform in order to provide increased spectrum utilization and enhanced end-to-end QoS for users of different RATs with and without fixed spectrum allocation.
4. investigate the problems of co-existance, intra- and cross-layer control between different RATs comprising the network, including:

(a) PHY layer channel modeling, including noise and interference models, log-distance path loss, shadow and multipath induced fading, physical layer transmission techniques (MCS, AMC);
(b) MAC/RLC layer design, including traffic generation models, packet scheduling, ARQ/HARQ, DCF/HCF, buffer status reporting, etc.;
(c) Cross-layer control: necessary parameters (such as packet arrival rate, buffer occupancy, SINR) are observed on MAC and PHY; control of available resources (such as bandwidth, data rate, buffer capacity) on PHY layer;
(d) Application layer QoS for users as a result of undertaken control on PHY/MAC layer.

3.2 Main Research Contributions

As has been indicated in Literature Review, a significant progress has been made in resource allocation and management in cognitive and cooperative radio networks. However, many challenges still remain. For instance, many of the theoretical works where the analysis of cooperative and cognitive networks is made using game-theoretic approach or queueing theory (e.g., [20, 21, 32, 40, 41]) are often very general and do not comply with specific features of different wireless networking standards, which means that the results of these works cannot be deployed in the "real-world" environment. Besides, in many of these works the performance of the wireless system is analyzed using parameters defined by authors themselves (see [40, 41]), whereas the end-to-end QoS of the users is not evaluated.

Improvement of the overall user-perceived QoS is the focus of the other works studying different resource allocation techniques and their application in specific RATs ([44-47, 50]). In these works resource allocation problem is formulated in terms of some optimization problem where some user utility is maximized subject to capacity constraints, interference requirements of PUs, etc. Most of these works maximize total network throughput, minimize transmission or propagation delay for the users ([44 - 47]). Although these optimization targets are easy to obtain analytically, they do not always imrove user-perceived QoS, since in many wireless networking standards (such as Wi-Fi or LTE) the total throughput and total packet end-to-end delay

are greatly influenced by many parameters. For instance, in LTE networks the packet end-to-end delay and loss can increase in times even if the transmission delay is small due to large amount of control signaling or noisy erroneous channels because of the scheduling and HARQ procedures [93], whereas in Wi-Fi networks packet end-to-end delay and loss greatly depend on contention in the wireless medium [69].

Another important factor that has not been considered in prior research is a heterogeneous nature of CRN. Most existing works (for example, [32]) on spectrum management and end-to-end QoS provisioning for CRN focus mainly on homogenous scenarios, ignoring the potential problems that could arise in a heterogeneous case (differences in QoS requirements, channel conditions, traffic priorities, etc.). Although such simplified approach may be allowed at the beginning of investigation, it does not give effective and practically realizable solutions.

This research project contains a collection of different resource allocation techniques designed to combat the challenges listed above, as well as to deal with some other aspects of the future wireless networks (such as heterogeneous users traffic, presence of multiple RATs, etc.). Although we consider the application of these techniques to specific wireless network interfaces (Wi-Fi and LTE), most of them can be deployed in any OFDMA-based network. In this thesis we consider three different network deployment scenarios, and offer some general network architecture and resource allocation policy which can be implemented using any of the proposed algorithms to improve the overall capacity and QoS in the network. For instance, in most of these algorithms we use buffer (queue) size in the bottleneck nodes of the network QoS indicator (or user utility), since buffer tends to grow when the network is congested (which might happen not only in case when the network is overloaded, but also if there are many errors or collisions in wireless channels).

We also consider that additional control signaling over the wireless medium can reduce service performance substantially, and therefore use IP-based infrastructure for control communication. Finally, most of the methods proposed in this project can be effectively used to deal with heterogeneous network environments: several attempts to solve this problem have been made previously, but most of the existing

techniques for heterogeneous traffic are either very complex (such as [76]) or lead to rather unfair resource allocation in the sense that applications with lower demand are allocated a higher transmission rate than applications with higher demand (e.g., [77, 78]).

Research Methodology

There are numerous ways for enforcing the optimal resource allocation in CRNs. One way is to build a stochastic model of the network (see, e.g., [40], [41]) characherized by one or mode stochastic parameters and/or one or mode deterministic parameters. After this, one can derive a stochastic optimization problem, according to which the network resources (service rate, bandwidth, power, etc.) will be allocate dynamically based on obtained optimal solutions of the problem. This way of resource allocation is very dependent on representation of probability density functions describing the stochastic external and/or iternal system parameters. In other words, if the stochastic parameter is not correctly modelled, then the consistent performance of the algorithm cannot be guaranteed. In CRNs, the stochastic parameters usually represent the external system parameters, such as user behaviour (duration of ON-OFF session periods, amount of generated traffic, etc.) or noise in the wireless channels. However, given the known difficulty of such parameters as traffic pattern of the users or channel noise, the accurate stochastic modelling is not always possible [180], [181].

An alternative research methodology for resource control in CRNs is based on game theory. In this case the resource allocation to the users is modelled as a game in which each user represents the player with the goal to maximize its utility. Many of such games have been formulated as non-cooperative (e.g., [182] – [184]). In these games, rational users (players) selfishly maximize their individual utilities without being concerned about the impact of their strategies on other users. A typical solution to a non-cooperative game is a Nash equilibrium solution (NES), in which each player has no chance to increase its utility by unilaterally deviating from this equilibrium. Unfortunately, the NES has been proven to be inefficient, meaning that the achievable total network utility can be low compared to the centralized optimization [185]. The non-cooperative game theoretic framework is very well

suited to network scenarios where infrastructure is sparse or completely absent, as in peer-to-peer and ad hoc networks. However, in infrastructure-based networks like cellular, broadband access, and to some extent wireless local access networks, where a centralized operator retains control over the common resource, the purely non-cooperative model is overly pessimistic, as it may not be able to fully capture the gain that could be obtained from coordination [186].

In order to find a Nash equilibrium that is more Pareto efficient, pricing mechanisms have been investigated ([187] – [189]). Pricing is typically referred to the penalty paid by each player. Pricing does try to move the users' behaviour to benefit the network, but this entails finding the right cost function, which sets up another optimization problem to solve.

While non-cooperative game theory studies competitive scenarios, cooperative game theory provides analytical tools to study the behaviour of rational players when they cooperate. The main branch of cooperative games describes the formation of cooperating groups of players, referred to as coalitions [190]. Coalitions can strengthen the players' positions in a game. Coalitional games have also been widely explored in application to distributed power control in CRNs for improving the network performance ([190] - [192]). However, implementing cooperation in large scale networks faces several challenges such as adequate modelling, network efficiency, complexity, and fairness. We further note that, for dynamic resource control, game theory has the disadvantage of fluctuation in external system parameters (user traffic, noise and interference) before the games converge [193].

The general research methodology used in this thesis for the development of resource allocation techniques (described in Chapters 4 – 8) is the following.

- First, we build analytical model describing the traffic behaviour and usage pattern in the considered network deployment scenario. This analytical model is fully deterministic, which means that each parameter in this model can be either observed (by the users or at the base station) or more or less accurately estimated using any of the prediction techniques provided in Chapter 1.

- Next, we formulate the objective of resource allocation (overall network utility), and state corresponding optimization problem.
- Based on this optimization problem, we derive a resource allocation algorithm, and implement it in the simulation model built upon the OPNET platform [112]. The simulation models of the users in all chapters (except Chapter 1) are the standard 3GPP2 traffic generation models described, for instance, in [194].
- Finally, we evaluate the service performance of the algorithm by comparing its performance with the performance of conventional and/or previously proposed resource allocation techniques. The validity of simulation results in this thesis is justified by high number of observations (10 ÷ 20 observations for each observation point in the graphs) which we collect with different random seed states (the seed value can be chosen in OPNET simulator). After collecting the results for each observation point, the average value is calculated, and is used to plot the final graph. Therefore, the accuracy of simulation results provided throughout this thesis is relatively high.

The main advantages of the proposed approach for resource allocation are rather straightforward.

- First of all, it achieves socially beneficial results by maximizing the overall network utility.
- Secondly, all of the proposed resource allocation algorithms yield optimal or near-optimal results, because here we rely on deterministic or predicted values of generated traffic and SINR, rather than on stochastic system parameters. Here we have to mention that the in case if we use traffic prediction in the algorithm, the resource allocation performance depends on the prediction accuracy. However, as we will show in Chapter 1, the prediction technique used in some of the algorithms in Chapters 4 – 8 is rather accurate (prediction error is less than 1%).
- Finally, almost all algorithms have low or moderate computational complexity, which provides the possibility to enforce resource control in large-scale networks.

Network Architecture

The IP based cognitive network architecture has been deployed upon the Wi-Fi (IEEE802.11a, g versions) and the 3d Generation Partnership Project (3GPP) long-term evolution (LTE) standard networks. These standards have been chosen because of the following reasons:

- both LTE and Wi-Fi are OFDMA-based standards, and therefore are suitable for the IP-based architecture;
- it is anticipated that LTE will get a widespread deployment in near future because of its high capacity, while Wi-Fi is already very popular;
- both Wi-Fi and LTE offer high service rates in the network, and therefore can support multimedia users;
- both Wi-Fi and LTE offer spectrum flexibility (scalable bandwidth for LTE and flexible data rate for Wi-Fi).

Network Deployment Scenarios

Most widely considered model of cognitive user behavior in the past has been based on non-cooperative network deployment scenario (shown on Figure 3). In this scenario, each secondary node senses the spectrum to find (on its own) the available unused bandwidth (spectrum hole) from primary node, and utilizes it according to the requirements of this primary node. At any time the connection of the secondary node can be blocked by the primary node (which usually happens in case if the bandwidth used by the primary node at the current state is not enough to satisfy the requirements of the primary node exceed the contemporary level) [33-34, 40-50]. Such network models have many disadvantages mainly because of the absence of cooperation and information exchange between all member nodes. As a result, secondary nodes will spend more time on unnecessary channel sensing and competing for access to the licensed spectrum bands, which eventually will lead to very poor service quality and increased power consumption of the user terminals.

In contrast to this approach, in this research project a different model of cognitive user behavior is proposed in which all member nodes cooperate by exchanging their network status information, and share the available spare capacity in an orderly manner. Information

exchange and coordination between the nodes will allow maximizing the overall capacity and QoS of the network. Secondary nodes will be able to (temporarily) borrow network resources in a more efficient manner to minimize blocking of connections, and also the need for continuous physical layer channel sensing will be eliminated, which will help to reduce the power consumptions of user terminals.

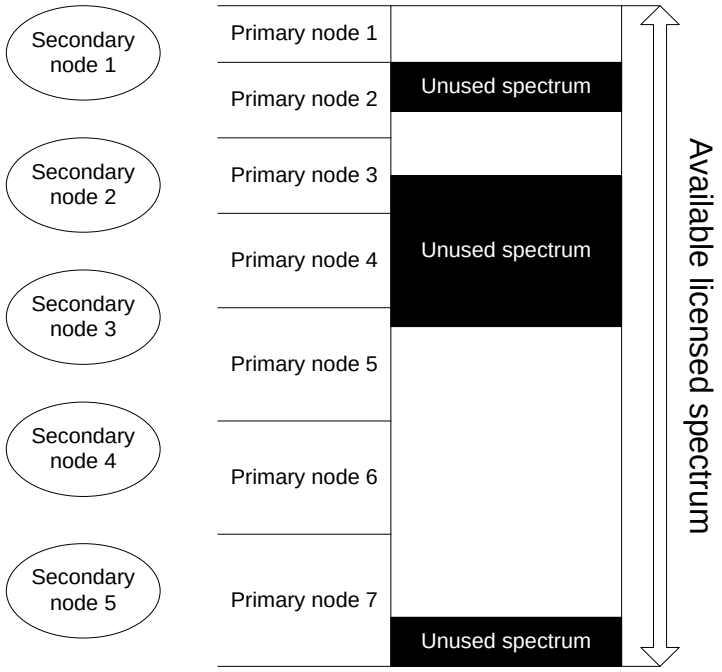


Fig. 3. A non-cooperative network deployment scenario

In this work integration of the cooperative and cognitive functionalities with the network infrastructure has been studied based on three most widely considered network deployment scenarios.

Scenario 1: A typical IEEE 802.22 standard Wireless Regional Area Network (WRAN) architecture comprising a number of service providers (SPs) with their base stations (BSs) as shown on Figure 4. Within the network, the SPs share the total available bandwidth among each other using the spectrum manager (SM)

according to some predefined flexible spectrum usage policy [51, 52]. All SPs/BSs have equal priorities in accessing the spectrum.

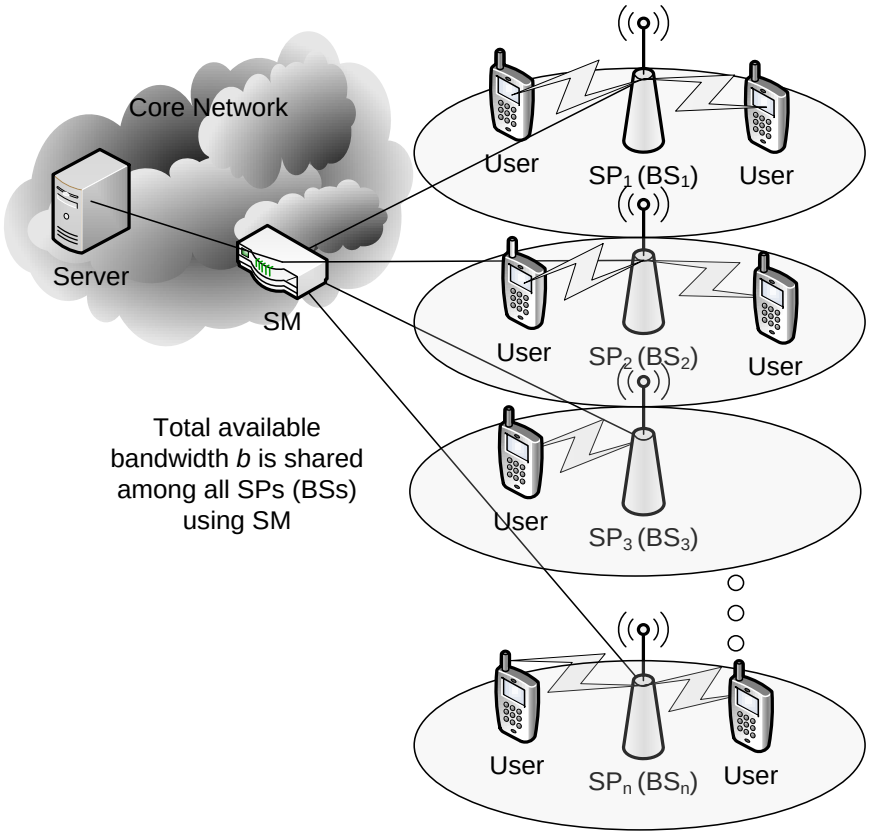


Fig. 4. IEEE802.22 standard network architecture (Scenario 1)

Scenario 2: A cognitive radio network architecture consisting of primary (licensed) BSs operating on their licensed spectrum bands (primary channels), secondary (unlicensed) BSs and SM (Figure 5). Each primary BS (PBs) can share its channel with one or more secondary BSs (SBs). In this case the primary station is given a prioritized access to its licensed spectrum band, whereas the secondary stations are served on the best-effort (non-prioritized) basis. Similar network deployment scenarios have been considered in [83 - 89].

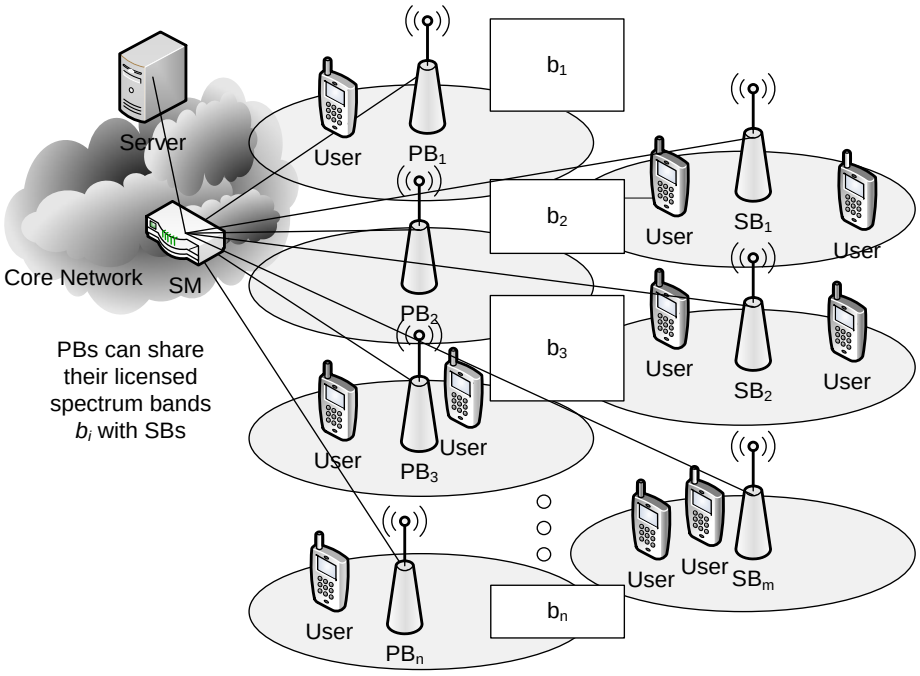


Fig. 5. Cognitive radio network architecture in Scenario 2

Scenario 3: A cognitive radio network architecture comprising a number of SPs with their BSs and SM as shown on Figure 6. Each BS operates on its licensed spectrum bands, and serves a number of primary and secondary users. Primary users (PUs) are the licensed network users who pay some price to their SPs for accessing the wireless services, and therefore have priority in accessing the spectrum. Secondary users (SUs) are unlicensed network users who can access the wireless services for free on best-effort basis. Similar scenarios of the network deployment have been considered in [91, 92].

In all scenarios above the IP links between SM and BSs are used to provide cooperation and information exchange within the network, and therefore the need for additional signaling over the wireless medium is eliminated. This issue is very important for the wireless network where the control signaling overheads introduce the additional delay, and have negative influence on user-perceived QoS.

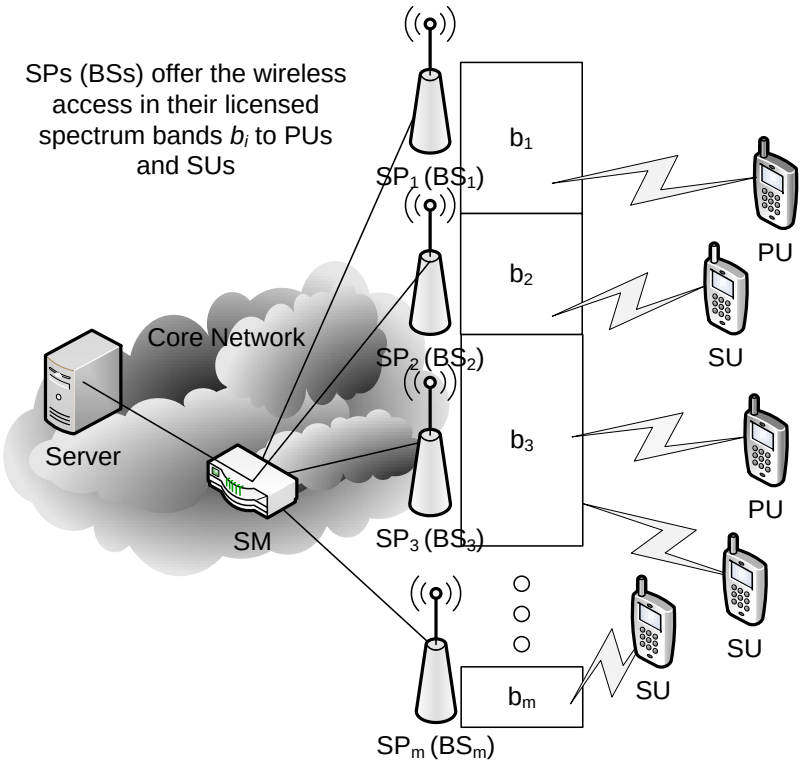


Fig. 6. Cognitive radio network architecture in Scenario 3

3.3 The Framework in Scenario 1

In this section a general framework on network modeling and resource allocation in Scenario 1 is briefly described. Mode detailed outline of the network model and spectrum access algorithm is provided in Scenario 1 is given in Chapters 5 – 7.

The IEEE 802.22 CR standard network architecture [51, 52] has been proposed in attempt to solve the problem of spectrum scarcity and reduced service quality created by the use of conventional fixed spectrum resource allocation policy. It has been proposed that the wireless access will be provided by a Wireless Regional Area Network (WRAN) comprising a number of SPs with their base stations. Within the network, the SPs share the total available bandwidth among each other using the spectrum manager (SM) according to some predefined flexible spectrum usage policy [51]. The standard describes the overall

network topology and a general dynamic spectrum access (DSA) framework on physical (PHY) and medium access control (MAC) layers, whereas the exact algorithm for spectrum allocation is not specified [52].

It is rather straightforward, that in order to realize great opportunities offered by IEEE 802.22 CR architecture, an appropriate choice of DSA policy is very important. However, development of the suitable spectrum allocation technique is a challenging task given the known difficulty of modeling and measuring the wireless medium [53]. Although a significant progress in diverse cognitive techniques during the last few years, many challenges still remain [54]. For instance, most research has focused on individual techniques for identifying and reducing the interference (by controlling transmit power, carrier sense, or scheduling) for the users of CR network (for instance, [55 - 58]). In general, however, the system performance depends on many external factors, including user behavior, traffic load, channel quality, etc. [54].

Some theoretical models of the user behavior and traffic load in CR network have been proposed in [59 - 62], but the assumptions made in theoretical research often fail under realistic operating conditions due to the fact that a system may operate in diverse environments (e.g., in different types of city, rural, campus, and indoor deployments) [54]. It is therefore very difficult to obtain some general theoretical model which can be applied for different network deployment scenarios. More rational would be to:

1. identify most critical parameters affecting the system performance;
2. investigate all available tools to analyze the service quality in the network based on the certain parametric observations collected in different locations at different time, and
3. apply these tools in spectrum allocation algorithm in order to improve the service performance of CR system [54].

Based on these considerations, we propose an alternative approach for resource allocation in IEEE 802.22 CRN by deploying reinforcement learning [63, 64] and using traffic prediction instead of complex analytical parameter modeling (description of the prediction technique applied in all scenarios is given in Chapter 1). This approach is further applied to three different resource allocation schemes derived for LTE, WLAN and combined LTE - WLAN networks which are briefly described below.

Resource allocation scheme for LTE based CRN

Proposed network model consists of a number of evolved NodeBs (eNBs) connected to a network resource manager (NRM) via IP based links. Assuming, that each eNB is characterized by a concave increasing utility function and positive weight, we conduct a weighted utility maximization framework, and develop a simple prediction-based resource allocation (PRA) algorithm.

Proposed way of resource allocation in the model has been induced by the concept used in traditional optimal flow and congestion control (OFC) where the resources are assigned based on speed of load increase in the bottleneck nodes [65]. This way of resource allocation prevents the growth of queues in eNBs (the growth of the queues in user terminals is neglectably small and usually do not lead to significant increase of delay or loss in the network). PRA firstly identifies eNBs with increasing (decreasing) load using appropriate load indicator, and then decrease (increase) the channel utilization of eNBs with increased (decreased) load using weighted proportional fairness criterion [66].

The values of load indicators in algorithm are obtained from the medium access control (MAC) and physical (PHY) layer information gathered from eNBs. Unlike most of the congestion and flow control algorithms where the nodes are described by some simple binary load indicators (congested or not congested node) [67, 68], PRA uses modified load indicators (MLI) described by more complex functions depending on queue size, loss and channel state information. To further increase the algorithm efficiency, resource allocation is performed based on predicted values of load in eNBs.

Resource allocation scheme for combined LTE - WLAN based CRN

Consider a combined LTE/WLAN network architecture comprising a number of service nodes (LTE eNBs and WLAN APs) connected to the System Architecture Evolution Gateway (SAE GW) via IP links. Control of the network resources is performed by the Network Resource Manager (NRM) located in SAE GW.

Resource allocation for heterogeneous networks consisting of multiple sub-networks based on different RATs (WLAN and LTE in our case) is a complex task due to diverse nature and the requirements of different networking standards. In such a network the impact of

physical layer characteristics (such as channel quality, spectrum efficiency, etc.) should not be underestimated. In other words, given the same bandwidth, the throughput in LTE eNBs and a WLAN APs will be different. Another important issue that should be considered is the packet mode channel access methods used in LTE/WLAN service nodes. For instance, IEEE 802.11 WLAN network uses contention based random multiple access technique for channel access over the wireless medium. This technique is usually characterized by numerous collisions, which can reduce the achievable throughput of WLAN users [69]. In LTE network the (potential) contention is resolved by using a Random Access Contention Resolution and Scheduling Request (SR) procedure. Hence, the collision probability in a LTE network is close to zero, thus, does not affect the service rates of LTE nodes [70].

Initial approach for resource allocation in combined LTE/WLAN CRN is very similar to the one deployed in LTE CRN. Each AP/eNB is assigned with appropriate bandwidth proportional to the value of its load control (LC) indicator, which measures the degree of load variation in service node. In this way a larger bandwidth is assigned to service nodes with increasing load, and smaller – to the nodes with decreasing load. The difference between the algorithm used for resource allocation in combined network and the algorithm used in LTE network is in the way the LC indicators are obtained. To be able to account for different spectral efficiencies, and channel access techniques deployed in LTE eNBs and WLAN APs, the spectrum efficiency and collision ratio metrics are measured discontinuously in each service node, and further used to calculate the values of LC indicators together with predicted traffic load in APs/eNBs.

Resource allocation scheme for WLAN based CRN

According to [54], the individual spectrum bands are used in a fairly homogeneous fashion. In contrast to them, the usage pattern in CRN is in general heterogeneous. Consider, for instance, the intra-campus network where some of the APs in this network can be located in academic schools, other APs can serve the staff buildings and the school libraries, whereas the rest can provide the wireless access in residential areas. It is reasonable to expect that the usage pattern in APs will be very different. For instance, the school APs might experience

heavy demand during the lecture hours and will not be used the rest of the time, the APs located in the offices and libraries will be loaded during the day-time and empty during the night, whereas the APs in residential buildings will be mostly used in the evening and night time. The web applications and traffic patterns of the individual users of these APs might also vary: the students and staff in the offices and the libraries might access the e-mail and perform the web-search, whereas in residential buildings the VoIP, video and on-line games might be used more frequently. Thus, to build a practically sustainable system it is important to keep in mind that different APs might operate in different conditions, i.e. the network usage is location and time dependent and the service demand in the network is heterogeneous.

Most of the resource allocation strategies for CRN have been deployed for homogeneous scenarios and not very efficient in case of heterogeneous network applications [59 – 61, 71 - 74]. This is due to the fact that all users in the network are characterized by similar utility functions. Existing approaches to deal with the problem of resource allocation in the network with heterogeneous user demands (for instance, [62, 75 - 78]) are either very complex (such as [76]) or lead to rather unfair resource allocation in the sense that applications with lower demand are allocated a higher transmission rate than applications with higher demand ([77, 78]). Therefore, in this work we suggest another approach for resource allocation in CRN, and propose to make a short-term resource allocation based on the long-term traffic prediction.

We consider the standard IEEE 802.22 CRN architecture comprising of Wi-Fi APs which share the total available bandwidth using the SM according to some predefined spectrum usage policy on a discrete-time basis. The system serves a number of wireless users connecting to the APs in their service area (cell) and generating a random traffic. Considered system can be well described using a model in which each AP is represented by a single infinite queue, whereas all users connected to the AP form the source served by this queue. The service rate of each queue depends on the portion of bandwidth assigned to respective AP and the spectrum efficiency of the wireless channel between the user and the AP. In this system we set the objective to allocate the service rates of APs in such way that the total system

bandwidth will not exceed the predefined limit based on some optimality criterion. The appropriate choice of the criterion is apparently one of the most critical factors affecting the performance of resource allocation for practical network implementations [75].

For most of the network applications (such as voice, video, data), the user-perceived QoS is determined in terms of the packet end-to-end delay and packet loss experienced by the user. For instance, for VoIP applications the satisfactory service is achieved when packet end-to-end delay does not exceed 300ms with packet loss less than 5%; for videoconference users the QoS requirements are the same as for VoIP applications; for streaming video the packet end-to-end delay should not exceed 4-5 sec with packet loss less than the QoS requirements for video applications are the satisfactory service network performance is achieved when packet end-to-end delay does not exceed 200ms with packet loss less than 1% [79]. Therefore, it would be reasonable to represent the optimization objective in terms of the packet delay or packet loss. However, in general it is very difficult to estimate the values of the packet delay or loss accurately, because they depend on many network parameters some of which might not be possible to observe directly. More convenient would be to use the queue size as an optimization objective because: i) it can be easily estimated using the Lindley's equation [80]; ii) it is the key parameter affecting both packet delay and loss.

Based on these considerations, we propose to represent the optimization objective in terms of the aggregate size of the queues over the long-term period in the future. In this way the resources will be allocated to APs to minimize the size of queues in the long-term future, which guarantees that bandwidth will be assigned in fair manner (the applications with lower average demand are allocated lower service rate than the applications with higher demand) and helps to overcome the negative impact of the bursty traffic on the overall QoS for the network users.

3.4 Framework in Scenario 2

In this scenario the problem of resource allocation for CRN based on the standard LTE network is considered. The advantages of LTE

system include increased peak data rate of up to 100Mbits/s for downlink and up to 50Mbits/s for uplink; improved spectral efficiency of up to 5bits/s/Hz for downlink and up to 2.5bits/s/Hz for uplink; improved cell edge performance (in terms of bit rate) and reduced latency [81]. Further, LTE can be deployed in different frequency bands of different sizes ranging from 1.4 MHz to 20 MHz and comes as both paired and unpaired bands. Paired frequency bands implies that uplink and downlink transmissions are assigned separate frequency bands, whereas in the case of unpaired frequency bands, uplink and downlink must share the same frequency band [82]. Such appealing characteristics make LTE to be the one of the most promising wireless standards for deployment in future CRNs.

Considered network model consists of a number of licensed (primary) eNBs sharing their licensed spectrum bands with unlicensed (secondary) eNBs using a central network manager (CNM) according to some predefined resource allocation policy. The eNBs are connected to the backbone server via a central network manager (CNM). Communication between eNBs, a backbone server and CNM is realized via the high-speed IP links to support fast transmission of data and control information. Each primary eNB operates on its fixed licensed spectrum band (primary channel) of a certain capacity. The primary eNB can share its channel with one or more secondary eNBs which don't have fixed licensed spectrum bands. In this case the primary eNB is given a priority in accessing the primary channel, whereas the secondary eNB(s) can access the primary channel on best-effort (non-prioritized) basis. The amount of capacity that primary eNBs share with secondary eNBs depends on the resource allocation policy used in CNM.

Different resource allocation algorithms have been proposed recently for resource allocation in LTE-based CRN architecture. Most research has focused on designing the lower layer techniques for spectrum sensing and spectrum mobility in CRN. The opportunistic spectrum access (OSA) for interference minimization in LTE-A has been investigated in [83]. It has been shown that implementation of the OSA in LTE-A enhances the overall system performance by intelligently aggregating otherwise unutilized spectrum. Relay selection and resource allocation in LTE-A cognitive relay network has been investigated in [84]. It has been assumed that primary stations

communicate via a relay assisted LTE-A network, some of the secondary stations play the role of the network relays, and the remainder nodes interact using the centralized network algorithm in the licensed spectrum. Simulation results conducted in the paper have shown that the proposed resource allocation algorithm shows increased throughput compared to the conventional random relay selection and uniform power allocation method. A cross-layer protocol of spectrum mobility and handover in cognitive LTE networks has been presented in [85]. The protocol has been developed based on the consideration of the Poisson distribution model of spectrum resources. Simulative performance study has illustrated that the proposed handoff protocol significantly reduces the expected transmission time and the spectrum mobility ratio within the network model.

The above techniques are very effective in identifying and reducing the interference in the physical channels, but do not improve the overall user-perceived quality of service (QoS) which is mainly expressed in terms of the packet end-to-end delay and loss for the network users. Theoretical studies of the user behavior and traffic load in CRN have been conducted in [86 - 89], but the assumptions made in these papers were very general in the sense that the specifics of LTE architecture have not been considered. Statistical traffic control scheme to ensure the QoS guarantees for all admitted traffic sources in cognitive LTE-A network has been proposed in [90]. However, the problem of user priority has not been addressed in the paper. In other words, it has been assumed that all traffic sources have the same priority.

We propose an alternative approach for resource allocation in LTE-based CRN. In order to provide the wireless access to secondary stations without compromising the QoS for the users of primary eNBs, we allocate the resources separately for primary and secondary eNBs using a two-stage procedure. During the first stage the resource are allocated for all primary eNBs to maximize the QoS for their users. During the second stage the rest of the service capacity of the primary channels is distributed among secondary eNBs.

Based on this approach, we derive two different algorithms for resource allocation in LTE-based CRN. Both algorithms do not involve additional network signaling over the wireless medium (the information exchange within CRN is performed over the high-speed IP links which enables the fast and reliable communication). First algorithm is simple,

has relatively short running time and is ideal for implementation in CRN with light and smooth traffic and/or when the processing capabilities of CRN are low and restrictive. Second algorithm uses future traffic predictions over the prediction window of the certain length for resource allocation. This algorithm is very effective in dealing with network congestions and therefore more suitable when the network traffic is heavy and/or bursty. However, because of the hire complexity and longer running time (than the first algorithm), it can be applied only if the processing capabilities of CRN are not restricted.

3.5 Framework in Scenario 3

In this work the network deployment Scenario 3 have been implemented based on the standard LTE architecture. Considered network model comprises a number of SPs offering the wireless services via a set of eNBs. Similar to the standard LTE system, considered network model operates on a slotted time basis with slot duration equal to 1 ms. Each eNB operates on a fixed licensed spectrum band and serves a number of primary users (PUs) and secondary users (SUs), randomly arriving to (and leaving) the network. PUs are the licensed network users who pay some prize for wireless services, and therefore get prioritized access to the spectrum bands within CRN. SUs are unlicensed network users who can access the wireless services for free, and therefore they are served on the best-effort (non-prioritized) basis. It is assumed that:

1. one SU can connect to at most one eNB;
2. the mean inter-arrival times of PUs and SUs (and the mean inter-departure times of PUs and SUs) are much greater than the slot duration, which is reasonable because in real network the mean inter-arrival times (and the mean inter-departure times) of the users are usually much greater than slot duration in LTE system;
3. the spectrum bands of eNBs are non-overlapping.

The goal of CRN is to serve the maximum number of SUs without violating the QoS of PUs. To achieve this goal, different algorithms have recently been proposed (see, for instance, [85, 89, 91, 92]). However, the conducted research has been very general, and did not

take into account the specifics of LTE radio interface (such as packet scheduling process and limited amount of control channels).

Based on the fact that for most of the network applications (such as voice or video) the QoS is determined in terms of the packet end-to-end delay, we propose to formulate the problem of spectrum assignment for SUs as an optimization problem with certain delay constrains of PUs. Hence, to solve the problem, it is necessary to find the relation between the packet end-to-end delay and the number of users in eNB.

According to [93], the packet end-to-end delay in LTE system comprises the following delay components:

- packet transmission and buffering delays in user equipment (UE) and the eNB;
- propagation delay between the UE and the eNB;
- packet delay due to hybrid automatic repeat request (HARQ) retransmissions;
- the uplink delay due to packet scheduling;
- processing delays of eNB and the UE;
- packet delay in core network.

Because of the small subframe size (the subframe duration in LTE is equal T_s = 1 ms), the transmission and the buffering delay components are very small in LTE system (2 and 1 ms, respectively) [93]. Propagation delay depends on the distance between the UE and eNB, whereas delay in core network depends on the distance between the eNB and a server (in orders of 1 ms for if the distance does not exceed 1000 km). Processing delays of eNB and the UE depend on processing capabilities of the equipment (typically around 5 ms) [93]. Delay due to HARQ retransmissions depends on the wireless channel quality (usually less than 4 ms). The largest delay component is delay due to uplink packet scheduling (in general, more than 8 ms) and constitutes the biggest part (≈36%) of the packet end-to-end delay. Unlike the other delay components, the scheduling delay depends on the number of users in eNB [93, 94].

Although many studies have been devoted to performance of different scheduling strategies (e.g. [95 - 99]), resulting packet end-to-end delay and loss for wireless users have been evaluated only by means of simulations, and no analytical verification of obtained results has been conducted. The average values of various delay components including delay due to packet scheduling have been given in [128].

However, no proper mathematical analysis confirming the delay values have been presented. To fill this void, we obtain the mathematical relation between the number of users in eNBs and the scheduling delay, and use this relation to formulate the optimization problem for spectrum assignment in cognitive LTE network. The corresponding resource allocation algorithm assigns the spectrum to SUs subject to the delay constraints of PUs. The algorithm description and performance analysis will be presented in Chapter 8 of this thesis.

3.6 Other Contributions

Other contributions of this thesis include different techniques that can be used in all considered network deployment scenarios to increase the efficiency of resource allocation, and performance evaluation of LTE network for its further deployment in future wireless network infrastructure.

Traffic Prediction Techniques for Resource Allocation

In order to increase the efficiency of resource allocation, prevent potential network congestions, decrease packet delay and connection loss for the wireless users, we deploy traffic prediction in all considered network deployment scenarios. Considering the known difficulty of parameter estimation for time-varying wireless channels and heterogeneous nature of the wireless traffic, comprising large number of different network applications (such as data, voice or video), we propose to use recursive estimation techniques applied with time-series models for traffic prediction. Unlike off-line estimation methods, these techniques do not require a long observation history, highly adaptive and have modest memory requirements [106]. The corresponding paper analyzes the performance of different on-line recursive identifications methods applied with various time-series models for real and theoretical traffic traces.

Priority Based Packet Transmission Technique

In future wireless network the coordination and information exchange between all member sub-networks is carried via wired links

connecting the SM to the APs/BSs (Figure 1). Hence, no additional control signaling over the wireless medium is required for control and resource allocation in B3G. In this way we avoid the potential loss of control information usual for the wireless channels (where it is related to the poor signal quality, errors, limited number of control channels, etc.).

To further increase the efficiency of resource allocation in all considered network deployment scenarios, we propose a priority based packet transmission technique which can be used to increase the capacity of the wired channels connecting SM to APs/BSs. In the corresponding paper we describe the technique, and show its implementation in infrastructure based network Wireless Local Area Network (WLAN) is described.

Performance Analysis of LTE Network

To implement existing wireless standards into the future wireless network infrastructure, it is essential to carry out a comprehensive performance analysis of these standard networks.

While the capacity and coverage of IEEE802.11g (Wi-Fi) network has been widely investigated (see, for instance, [173 - 176]), the service performance of LTE network is still not fully explored due to the numerous design characteristics which have direct impact on QoS for the LTE users. Therefore, as part of the framework on implementation of LTE in B3G infrastructure, the following features of LTE air interface have been studied:

- PHY layer channel modeling, including noise and interference models, log-distance path loss, shadow and multipath induced fading, physical layer transmission techniques (MCS, AMC);
- MAC/RLC layer design, packet scheduling, ARQ/HARQ, buffer status reporting, etc.;
- Application layer QoS for users.

The corresponding papers study these LTE design characteristics based on their impact on capacity and QoS for VoIP users. The VoIP applications have been used for the performance analysis due to the following reasons: (i) they are expected to form a significant part in future wireless traffic [117]; (ii) they have the strictest delay

requirements compared to other network applications (VoIP can only tolerate packet end-to-end delay of up to 100 ms and packet loss of up to 1%) [116]. Thus, the ability of LTE to achieve good performance for voice applications will automatically guarantee that QoS for other user applications will be satisfied.

CHAPTER 1: Traffic Predictions Techniques for Cognitive Wireless Networks

This chapter provides an overview and performance analysis of the various traffic prediction techniques for resource allocation in cognitive wireless network. The corresponding paper titled "A Predictive Network Resource Allocation Technique for Cognitive Wireless Networks" has been published in Proceedings of IEEE International Conference on Signal Processing and Communication Systems (ICSPCS), 2010.

1 Introduction

Accurate traffic prediction is crucial for resource allocation in future cognitive wireless networks. When radio resource occupancy is predicted accurately for each sub-network comprising the wireless network, the users can select an optimal channel, which will help to increase the QoS and minimize the connection loss probability [101].

In future, the wireless services will be provided through heterogeneous networks rather than using a single standard network [102, 103], assuming that the network will comprise a large number of sub-networks belonging to different radio access technologies (RATs). Traffic modeling and prediction in heterogeneous network is very complicated, and requires a long observation of history. Besides, due to slow variation of the stochastic network parameters, a short-term forecasting might not perform well. For more accurate prediction, the parameters should be estimated on-line to track time-varying traffic characteristics. These features are common for recursive identification methods which can be applied together with non-real-time identification methods to make more accurate parameter estimation. The advantages of recursive methods can be summarized as follows:

- they are central part of adaptive systems where the filtering action is based on the most recent model;

- they have relatively small (compare to off-line identification methods) requirements on primary memory;
- they can be modified into real-time algorithms to track time-varying parameters;
- they can be deployed for fault detection when the observed system has changed significantly.

Further in this Chapter we describe the recursive parameter estimation methods for resource allocation, evaluate the performance of these methods for short-term and long-term traffic prediction based on "real-world" and theoretical traffic traces, and discuss the implementation of these methods in cognitive wireless networks.

2 Recursive Techniques for Parameter Estimation

Let $y(t)$ be an observation of a random process Y at discrete time t. To generate prediction $y(t+1)$, information about past events, called time-series data $y(t-1), \ldots, y(1)$, is collected.

Many time-series models have been proposed for time-series analysis, such as autoregressive AR(n_a), moving average MA(n_c), autoregressive moving average ARMA(n_a, n_c) and autoregressive integrated moving average ARIMA (n_a, n_b, n_c), with n_a, n_b, n_c denoting the orders of the autoregressive, integrated and moving average parts, respectively [105].

Recursive parameter estimation methods use the following general model [106]:

$$y(t) = \varphi^{\mathbf{T}}(t)\boldsymbol{\theta} + e(t) \tag{1}$$

where $\boldsymbol{\theta}$ is a system parameter vector, and repressor $\varphi(t)$ depends on the past data and the model structure.

For the AR model [105]:

$$\varphi(t) = (-y(t-1) \ldots - y(t-n_a))^{\mathbf{T}}, \boldsymbol{\theta} = (a_1 \ldots a_{na})^{\mathbf{T}} \tag{2}$$

For the MA model [105]:

$$\varphi(t) = (e(t-1) \ldots e(t-n_c))^{\mathbf{T}}, \boldsymbol{\theta} = (c_1 \ldots c_{nc})^{\mathbf{T}} \tag{3}$$

For the ARMA model [105]:

$$\varphi(t) = (-y(t-1) \ldots - y(t-n_a)\, e(t-1) \ldots e(t-n_c))^{\mathbf{T}}, \boldsymbol{\theta} = (a_1 \ldots a_{n_a}\, c_1 \ldots c_{n_c})^{\mathbf{T}} \quad (4)$$

In the adaptive (real-time) identification methods, the parameter estimate $\hat{\boldsymbol{\theta}}(t)$ is computed in a recursive way by modification of a last obtained estimate $\hat{\boldsymbol{\theta}}(t-1)$ [106].

There are four recursive methods, recursive least squares (RLS), recursive instrumental variable (RIV), pseudo-linear regression (PLR), and recursive prediction error method (RPEM) used for parameter estimation of different models. Two of them are PLR and RPEM, which can be applied to track parameters of AR, MA, and ARMA models. The algorithm for parameter estimation in these methods is given by [106]:

$$P(t) = \frac{1}{\boldsymbol{\lambda}}[P(t-1) - \frac{P(t-1)\boldsymbol{\psi}(t)\boldsymbol{\psi}^{\mathbf{T}}(t)P(t-1)}{\boldsymbol{\lambda} + \boldsymbol{\psi}^{\mathbf{T}}(t)P(t-1)\boldsymbol{\psi}(t)}] \quad (5)$$

$$K(t) = P(t)\boldsymbol{\psi}(t) \quad (6)$$

$$\varepsilon(t) = y(t) - \varphi^{\mathbf{T}}(t)\hat{\boldsymbol{\theta}}(t-1) \quad (7)$$

$$\hat{\boldsymbol{\theta}}(t) = \hat{\boldsymbol{\theta}}(t-1) + K(t)\varepsilon(t) \quad (8)$$

where $\boldsymbol{\lambda}$ is a forgetting factor to discount the measurements obtained previously; the smaller is the value of $\boldsymbol{\lambda}$, the faster information is forgotten (usually $\boldsymbol{\lambda}$ is set in the range [0.95, 1]);
$P(t)$ can be found from Hessian approximation in Gauss-Newton algorithms $R(t)$ using:

$$P(t) = \overline{R}^{-1}(t), \overline{R}(t) = tR(t) \quad (9)$$

$K(t)$ is the gain vector showing how much the value $\varepsilon(t)$ will modify the different elements of $\boldsymbol{\theta}(t)$;
$\varepsilon(t)$ is the prediction error of estimation given by:

$$\varepsilon(t) = y(t) - \hat{y}(t) \quad (10)$$

$\boldsymbol{\psi}(t)$ - negative gradient of $\varepsilon(t)$ with respect to $\boldsymbol{\theta}(t)$ given by (5):

$$\boldsymbol{\psi}(t) = (-y^F(t-1) \ldots - y^F(t-n_a) \quad \varepsilon^F(t-1) \ldots \varepsilon^F(t-n_c))^{\mathbf{T}} \tag{11}$$

$y^F(t)$, $\varepsilon^F(t)$ - filtered data [106].

For the RPEM:

$$y^F(t) = y(t) - \hat{c}_1(t) y^F(t-1) - \ldots - \hat{c}_{n_a} y^F(t-n_a) \tag{12}$$

$$\varepsilon^F(t) = \varepsilon(t) - \hat{c}_1(t) \varepsilon^F(t-1) - \ldots - \hat{c}_{nc} \varepsilon^F(t-nc) \tag{13}$$

For the PLR:

$$y^F(t) = y(t) \tag{14}$$

$$\varepsilon^F(t) = \varepsilon(t) \tag{15}$$

i.e. filtering of RPEM is neglected [106].

The effect of initial values on performance of recursion has been widely discussed in literature (see, e.g. [106, 107]). Without any priori information it is common practice to set:

$$\hat{\boldsymbol{\theta}}(0) = 0, P(0) = \rho \mathbf{I} \tag{16}$$

where ρ is a "big" number. Usually P is set in such as way so that the following relation satisfies.

$$P^{-1}(0) << \sum_{s=1}^{t_0} \varphi(s) \varphi^{\mathbf{T}}(s) \tag{17}$$

where t_0 is in the range [106].

3 Traffic Prediction Performance

In this section the summary of traffic prediction performance is presented. More detailed performance evaluation of traffic prediction can be found in the corresponding paper.

One theoretical (Poisson packet arrival process with constant and varying mean) and two experimental traffic traces (data packages LBL-Conn-7 and DEC-Pkt1) have been used to observe the performance of different time-series models applied with recursive parameter estimation techniques in different network environments. The first

trace, LBL-Conn-7, contains the Transmission Control Protocol (TCP) traffic data between the Lawrence Berkeley Laboratory and the rest of the world in the format where timestamps have microsecond precision. After processing the trace for uplink connections only we get another trace where overall uplink data rate is calculated for each microsecond. Recursion starts immediately with the observation data. The second trace, DEC-Pkt1, contains the all wide-area traffic between Digital Equipment Corporation (DEC) and the rest of the world with cumulative traffic volume given for each microsecond. Last trace used in the performance evaluation is theoretical Poisson generated sequence. Even though the Poisson model has been reported to be unsuitable for Internet traffic modeling [108], Poisson generated traffic is still widely used in communication networks, and can be used as a good example of a process with non-zero mean and highly random (unpredictable) pattern (a Hurst parameter [109] of a Poisson process is equal $H = 0.5$).

Performance of the traffic prediction has been evaluated using two metrics: normalized mean squared error (NMSE) and prediction error ratio (PER) given by:

$$\mathrm{NMSE}=\frac{\sum_{t=1}^{N}(y(t)-\hat{y}(t))^2}{\sum_{t=1}^{N}(y(t)-\bar{y}(t))^2},\ \mathrm{PER}=\left|\frac{y(t)-\hat{y}(t)}{y(t)}\right|\cdot 100\% \tag{18}$$

where $y(t)$ is the actual value of data rate at time t; $\hat{y}(t)$ is predicted value of data rate at time t; $\bar{y}(t)$ is mean value of the data rate estimated at time t.

The NMSE values for short-term (1-step ahead) and long-term (10-step ahead) prediction obtained using PLR prediction techniques applied with AR, MA, ARMA and ARIMA time-series models after 3000 recursion for different traces are given in Table 2. Both short-term and long-term predictions results show that best prediction performance is obtained using the AR model. Results also show that the MA model is unstable and fail to predict traffic values accurately.

The appropriate order $p = 1$ from the family of AR(p) time-series models has been chosen by minimizing the Akaike Information Criterion (AIC) [110] given by:

$$AIC(N,p) = N \log V_N(\boldsymbol{\theta}_N) + 2p, V_N(\boldsymbol{\theta}_N) = \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\varepsilon}^2(t,\boldsymbol{\theta}) \quad (19)$$

where N – is the total number of recursions; V_N - loss function. Figure 7 shows AIC for trace LBL-Conn-7 after N = 3000 recursions.

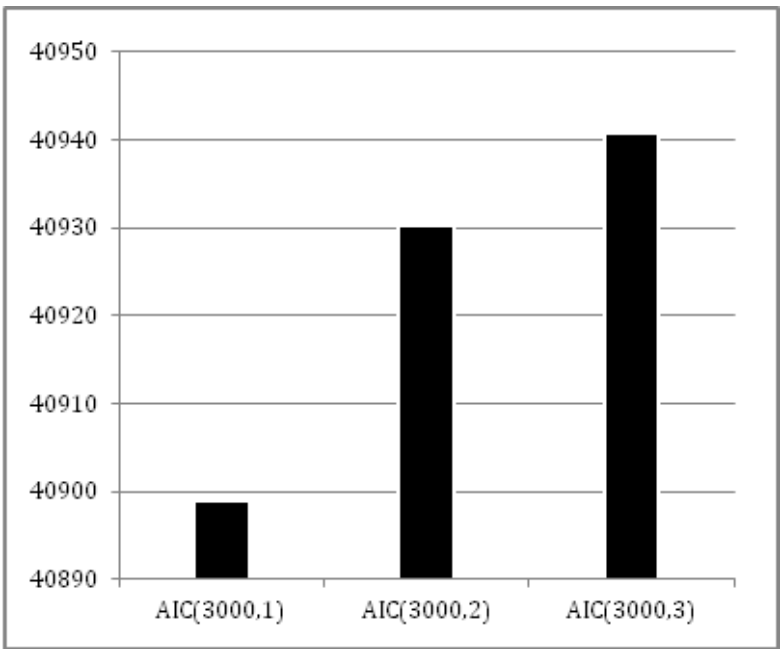


Fig. 7. *AIC*(*N*, *p*) after *N* = 3000 recursions for LBL-Conn-7

To compare performance of different recursive estimation methods, we tracked the traces using PLR and RPEM applied with AR(1) time-series model. Previous research reported that both techniques offer consistent performance, but behavior of the PLR in the transient phase might be better than that of the RPEM [106]. Our results confirm this observation – at the beginning of observations RPEM is less accurate than PLR (the accuracy of prediction estimated using PER metric during first 80 recursions and after 20000 recursions for one of the traces, DEC-Pkt1 are shown on Figures 8 and 9, respectively). Thus, RPEM takes more time than the PLR to estimate the parameters of the model. In our case the trace convergence delay for PLR method is approximately 10-15 recursions, whereas for RPEM is 50-60 recursions.

Table 2. Values of NMSE for different traces with PLR

Trace	*Model*	*NMSE*	
		1-step-ahead prediction	*10-step-ahead prediction*
LBL-Conn-7	AR(1)	0.071397	0.151657
	MA(1)	0.637899	1.225446
	ARMA(1,1)	0.072005	0.146376
	ARIMA(0,1,1)	0.0741	-
	ARIMA(1,1,0)	0.0742	-
	ARIMA(1,1,1)	0.2977	-
DEC-Pkt1	AR(1)	$4.81302*10^{-7}$	$4.60807*10^{-6}$
	MA(1)	1.742629	3.44529
	ARMA(1,1)	$4.89*10^{-7}$	$1.75*10^{-6}$
	ARIMA(0,1,1)	$4.81312*10^{-7}$	-
	ARIMA(1,1,0)	$6.3*10^{-7}$	-
	ARIMA(1,1,1)	$8.16*10^{-7}$	-
Poisson	AR(1)	1.578521	-
	MA(1)	3.576409	-
	ARMA(1,1)	1.035006	-
	ARIMA(0,1,1)	1.68656	-
	ARIMA(1,1,0)	1.471443	-
	ARIMA(1,1,1)	1.610663	-

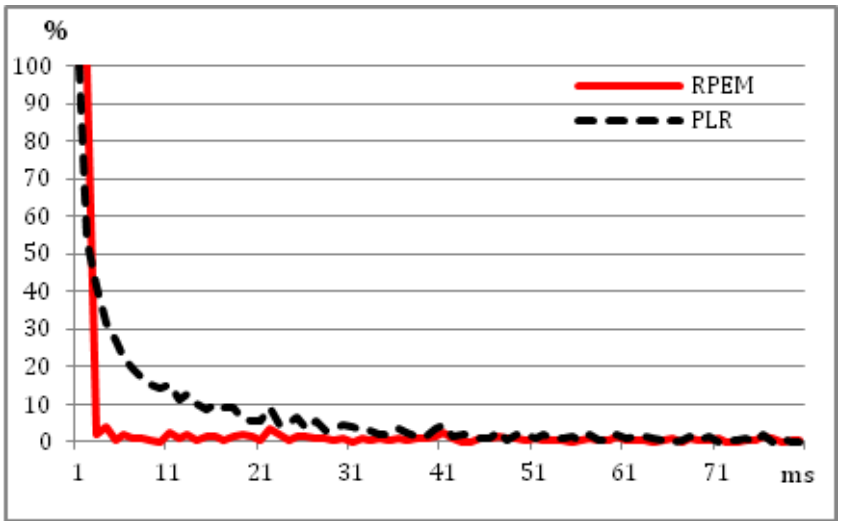


Fig. 8. PER at the beginning of observation using PLR and RPEM with AR(1) for DEC-Pkt1

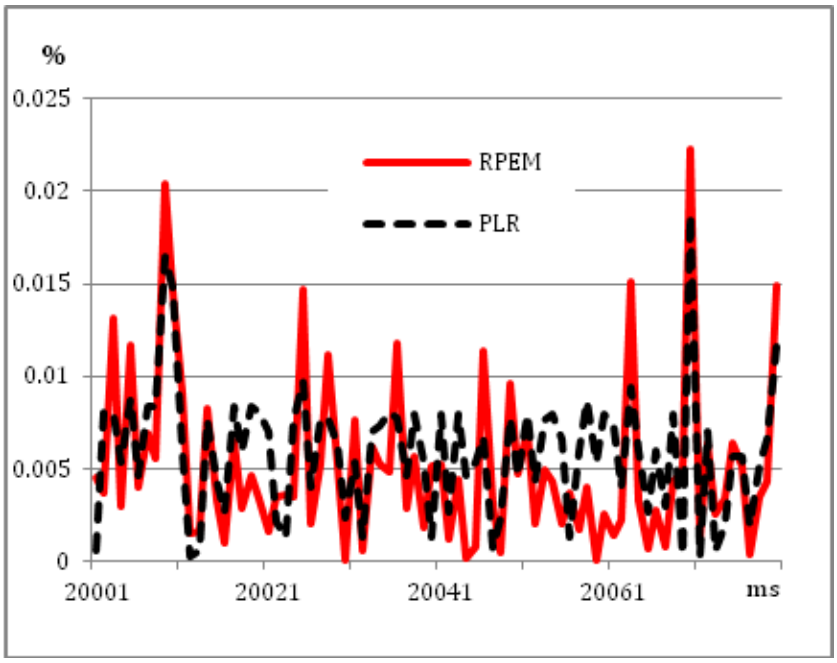


Fig. 9. PER after 20000 recursions using PLR and RPEM with AR(1) for DEC-Pkt1

Previous research has shown that drifting disturbances and non-zero means (such as in Poisson process) can well be treated by the family of

ARIMA time-series models [105]. ARIMA model is an ARMA model constrained to have the factor of $y(t) - y(t\text{-}1)$ [105]. However, our observations contradict these results: Table 2 shows that the lowest value of NMSE for Poisson traffic prediction is achieved by ARMA(1,1) model, whereas AR(1) model is more accurate in predicting the peaks of Poisson distributed traffic traces as illustrated on Figure 10. This Figure shows that the curve of traffic prediction with ARMA time-series model is smoother than that with AR time-series model.

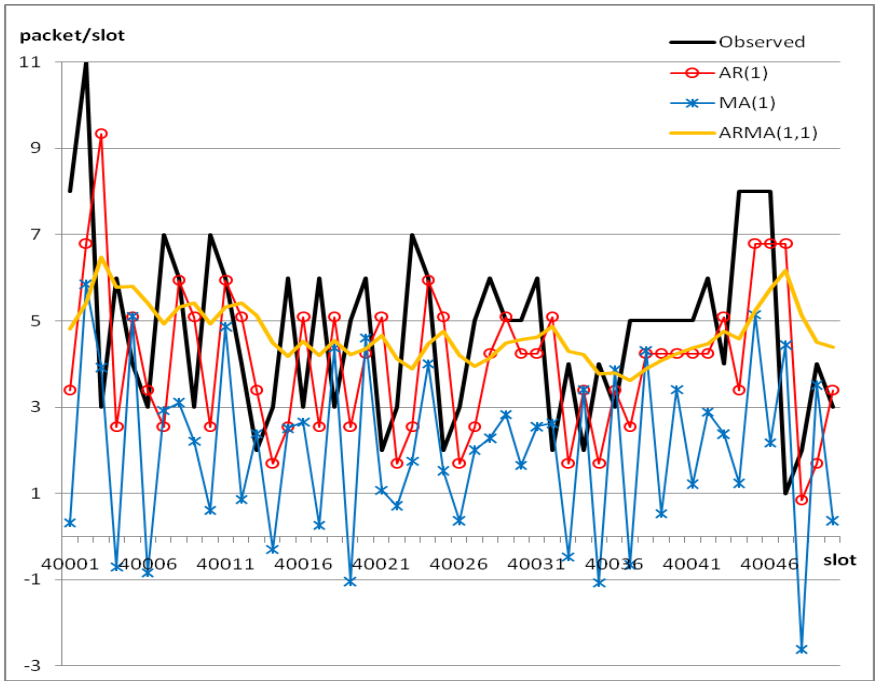


Fig. 10. 50ms-long observation and traffic prediction for Poisson generated traffic sequence

Results of this Chapter can be summarized as follows:

- Both PLR and RPEM techniques can be used for traffic prediction with time-series models, although PLR technique does not use filtering which is applied in RPEM. Compare to RPEM, PLR converges more quickly, less complex and less memory-demanding, and therefore has higher possibility of implementation in cognitive

wireless networks where prediction should be produced in very short time with high accuracy.

- Both AR and ARMA show relatively high (compared to other time-series models) accuracy for "real-world" and theoretical traffic traces. AR(1) achieves the lowest NMSE for LBL-Conn-7 and DEC-Pkt1 traces, ARMA(1,1) shows the lowest NMSE for Poisson generated traffic sequence.

Based on these observations, it can be recommended to use AR(1) for traffic prediction in cognitive wireless network. AR(1) shows very high accuracy for "real-world" traffic traces. Although AR(1) has lower (than ARMA(1,1)) NMSE for Poisson generated traffic sequence, it is still very accurate in predicting the peaks of this random sequence (Figure 10).

CHAPTER 2: Traffic Prediction Based Packet Transmission Priority Technique in an Infrastructure Wireless Network

In this chapter a priority based packet transmission techniques for an infrastructure based network Wireless Local Area Network (WLAN) is described. The corresponding algorithm derived in the paper can be used to support future wireless network infrastructure to improve the capacity and the QoS for the users under all network deployment Scenarios considered in this thesis. The corresponding paper has been published in Proceeding of IEEE Wireless Communications and Networking Conference (WCNC), 2011.

1 Introduction

With increasing demand for wireless data and multimedia services the role of infrastructure based WLAN is increasing. In an infrastructure based network the Quality of Service (QoS) in uplink and downlink channels is influenced by individual links which forms a multi-hop network to transmit traffic in both directions. Hence, the QoS for the network users is affected not only by the wireless link between respective Access Point (AP) and user terminal, but also by the radio access network (RAN) which connects the AP to other external networks (Figure 11) [111].

In this Chapter we introduce a novel traffic prediction based packet transmission priority technique which can be used (together with any of existing dynamic spectrum access (DSA) algorithms) to improve the capacity and the QoS in future wireless networks. In this technique the recursive parameter estimation is used to predict the size of next job at the downlink Ethernet channels connecting APs to the gateway. The corresponding algorithm allocates low (“0”) or high (“1”) priority to the queues of Ethernet channels using the Shortest-Job-First (SJF) approach. Thus, the queues will the smallest arriving jobs are assigned the higher priority “1”.

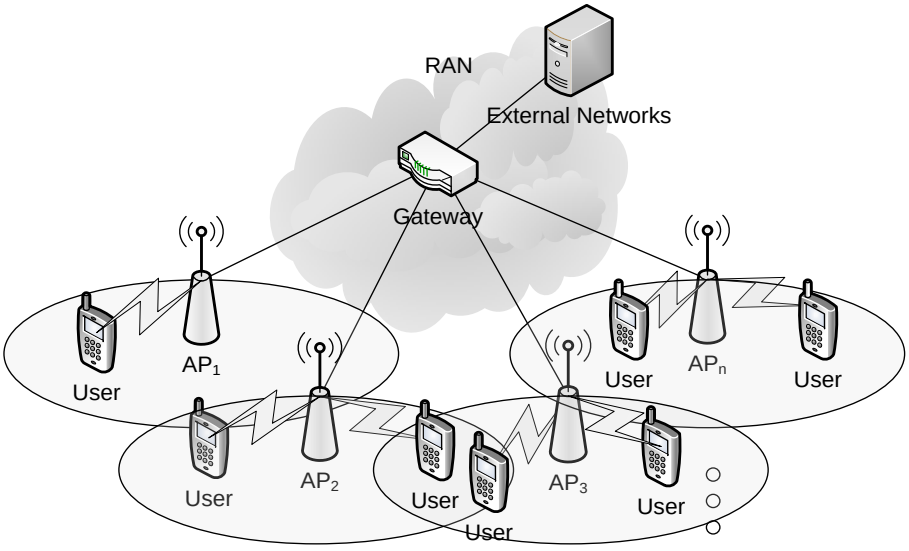


Fig. 11. A typical infrastructure based WLAN architecture

Two forms of the proposed algorithm have been considered. First form does not differentiate the jobs arriving to the same queue by their type of service (ToS). In the second form four priority levels are used to differentiate jobs arriving from different users: higher priorities are assigned to real-time applications, such as voice or video; lower priorities are assigned to data applications, such as http, ftp or email. Proposed algorithm has been simulated using OPNET platform [112], and compared with other commonly used priority allocation techniques. Further in the Chapter we present the packet transmission priority algorithm, show algorithm implementation in infrastructure based WLAN and provide simulative performance analysis of the algorithm.

2 Packet Tansmission Priority Algorithm

In this section we describe the traffic prediction based packet transmission priority algorithm for infrastructure based WLAN where the gateway connects multiple APs to external networks using Ethernet links (Figure 11). Here we present the example of algorithm implementation in the downlink direction. However, same approach can be applied in the uplink direction.

Proposed algorithm operates on a slotted-time basis, i.e. the time axis is partitioned into discrete mutually-disjoint time intervals, called time slots. The algorithm works by assigning different priorities to Ethernet channels in downlink direction based on predicted traffic load of respective APs. By default, all APs are assigned the lowest priority "0". Within each time slot, APs transmit their instantaneous traffic load information to the Gateway (using Ethernet links). The traffic load information is represented by the size of medium access control (MAC) queue of respective AP (in packets) in the downlink direction.

At the beginning of each time slot the gateway performs the following actions:

- receives traffic load information from APs;
- performs recursive prediction based on updated traffic load values (if for any reason the traffic load value hasn't been received, then the last available data is used);
- assigns priority to MAC queues of Ethernet channels connecting APs to the gateway.

The following approach is used to assign the priority to the downlink channels: if predicted traffic load of AP comprises less than half link data rate (in packets per slot), then a highest priority "1" is assigned to the respective channel. This approach is very similar to the approach used in Shortest-Job-First (SJF) scheduling algorithm used in computer operating systems. It has been shown that serving flows in order of job size using the approach "shorter flows served first" leads to significant reduction of response times for the flows of all lengths [113].

The SJF scheduling algorithm is probably optimal, because it yields the minimum average service time as well as high throughput. However, there is no way to know the size of the next job and hence, the SJF is not implemented at the level of short-term scheduling. For short term scheduling, it is necessary to arrange the flows according to the size of their jobs which is a challenging problem. In this paper we show a simple way to implement the SJF scheduling by predicting the size of next flow.

To analyze performance of the SJF algorithm, consider M/M/1 queue with unlimited storage representing the MAC queue of each Ethernet channel connecting APs to the gateway. Let λ be the job arrival rate, *B*

be the job service time, and $f_B(x)$ be the probability density function (p.d.f.) of service time distribution of the queue. Then, the mean waiting time with SJF scheduling algorithm for the jobs with a service time of $x \le E\{B\} \le x+dx$ can be calculated using the expression [114]:

$$E\{W(x)\} = \frac{\rho E\{R\}}{(1-\int_0^x \rho(y)dy)^2} = \frac{\rho E\{R\}}{(1-\lambda\int_0^x yf_B(y)dy)^2} \tag{20}$$

where $\rho = \lambda E\{B\}$ is the occupation rate (or utilization) of the queue; $\rho E\{R\}$ is the mean size of a job.

The overall mean waiting time for all jobs in M/M/1 queue with SJF job scheduling is given by [114]:

$$E\{W\} = \rho \int_{x=0}^{\infty} \frac{e^{-x}dx}{(1-\rho(1-e^{-x}-xe^{-x}))^2} \tag{21}$$

Now let us compare mean waiting time in SJF with that in commonly used First-In-First-Out (FIFO) job scheduling algorithm. The overall mean waiting time for all jobs in M/M/1 queue with FIFO job scheduling is given by [114]:

$$E\{W\} = \frac{\rho}{1-\rho} \tag{22}$$

Figure 12 shows the values of $E\{W\}$ in FIFO and SJF algorithms with different mean job size $E\{B\}$. Results demonstrate that SJF reduces waiting time for jobs of all sizes. More importantly, the difference between the waiting time in SJF and FIFO grows significantly when the channel utilization is high ($\rho > 0.8$).

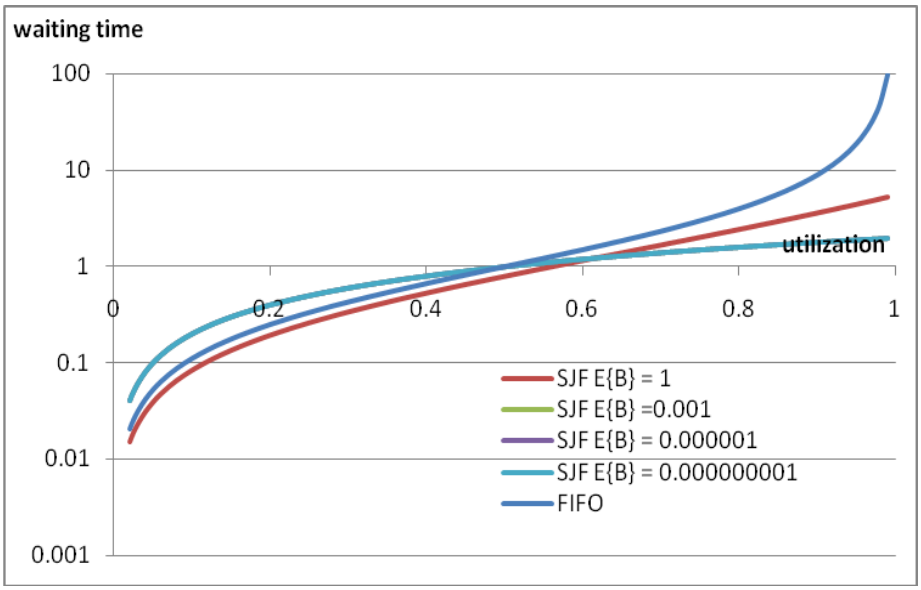


Fig. 12. Mean waiting time for jobs in FIFO and SJF

3 Algorithm Implementation

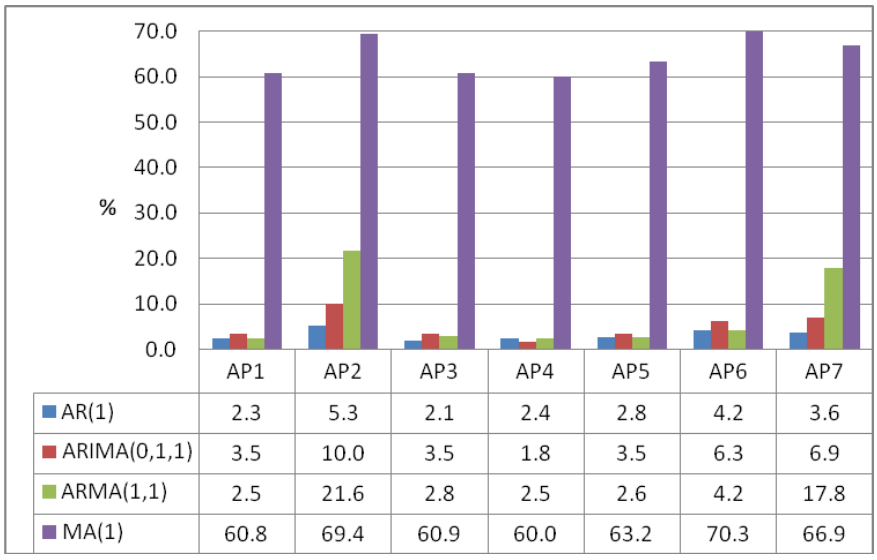
As it has been mentioned in prediction section, to implement the proposed algorithm in the network, it is important to predict the size of next job at each queue accurately. In this work we use pseudo-linear regression (PLR) recursive identification technique applied with autoregressive (AR) time-series model. The technique and the time-series model have been already described in Chapter 1. PLR has been chosen because of its relative simplicity, low memory requirements and shorter convergence delay compared to other recursive identification techniques (advantages of PLR for traffic predictions have been summarized in Chapter 1).

To choose best model for the parameter estimation, we applied different time-series models with PLR for traffic prediction in simulation model of the network developed in OPNET platform [112]. The network comprises seven IEEE 802.11g APs connected to external network server via gateway using Ethernet 1000BaseX duplex links. Each AP serves a number of wireless users generating random traffic. Traffic mixes used in each AP during simulation are listed in Table 3.

Table 3. List of traffic sources used in different access points

AP#	*Application Traffic*	*Number of traffic sources*	*Average data rate generated by traffic source (kbits/s)*
AP1	Voice	8	35
AP2	Voice, Video	15	500
AP3	Database, Ftp, Print, Remote Login	8	600
AP4	Http, E-mail, Database, Ftp, Print, Remote Login, Voice, Video	14	250
AP5	Http, E-mail, Database, Ftp, Print, Remote Login	5	1000
AP6	Http, E-mail	2	2000
AP7	Video	9	700

Figures 13, 14 show results of traffic predictions made by autoregressive AR(p), autoregressive integrated moving average ARIMA(p, d, q), autoregressive moving average ARMA(p, q) and moving average MA(q) models with p, d, q denoting the orders of the autoregressive, integrated and moving average parts, respectively. Prediction accuracy in simulation is measured using prediction error ratio PER (the expression for PER has already been provided in Chapter 1). Results show that MA time-series model is not suitable for traffic prediction in simulated network. ARMA model is not stable, and fails to predict the traffic in some APs (AP2 and AP7). AR and ARIMA models demonstrate very similar results with slight outperformance of AR model.



	AP1	AP2	AP3	AP4	AP5	AP6	AP7
AR(1)	2.3	5.3	2.1	2.4	2.8	4.2	3.6
ARIMA(0,1,1)	3.5	10.0	3.5	1.8	3.5	6.3	6.9
ARMA(1,1)	2.5	21.6	2.8	2.5	2.6	4.2	17.8
MA(1)	60.8	69.4	60.9	60.0	63.2	70.3	66.9

Fig. 13. Average PER in different APs

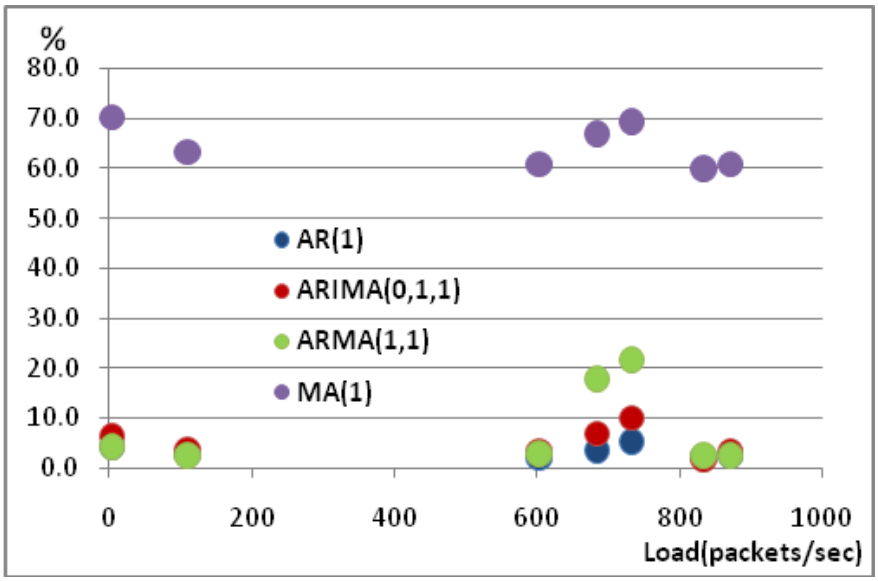


Fig. 14. Average PER for different traffic loads

4 Algorithm Performance

In this section the summary of algorithm performance is presented. More detailed performance evaluation of the proposed algorithm can be found in the corresponding paper.

Two different variations (forms) of the packet transmission priority algorithm have been implemented:

- First form of the algorithm, called MAC Priority Assignment scheme, uses only SJF approach for queue prioritization. Thus, depending on the size of the job waiting in MAC queues, different Ethernet channels are assigned low ("0") or high ("1") priority. The jobs are not prioritized by their type of service (ToS), i.e. within one queue the jobs initiated by voice, video or data users have same priority.
- Second form of the algorithm, called ToS & MAC Priority Assignment scheme, utilizes combined SJF-ToS strategy for queue prioritization. Thus, different Ethernet channels are assigned low ("0") or high ("1") priority depending on the size of the job waiting in respective queue. Within one queue the jobs are prioritized by their ToS (ToS priority levels for jobs arriving from different traffic sources are listed in Table 4).

Table 4. ToS priority for different traffic sources

ToS Priority	*Application Traffic*
0 (best-effort)	Http, Print, Remote Login
1 (low)	E-mail, Database, Ftp
2 (medium)	Video
3 (high)	Voice

We compare the performance of two proposed forms of packet transmission priority algorithm with performance of conventional WLAN where all Ethernet channels have low priority, and performance of WLAN with ToS prioritization. Figures 15 – 20 demonstrate

performance of different schemes in simulation models. Results of simulation show that:

- ToS Priority Assignment scheme improves QoS for certain types of users, but demontrate poor performance for non-prioritized users (with low and best-effort priority);
- MAC Priority Assignment scheme decreases Ethernet delay, but does not improve the QoS of the end-users;
- ToS & MAC Priority Assignment scheme very effective for all types of network applications: it decreases Ethernet delay and improves performance for prioritized and non-prioritized users.

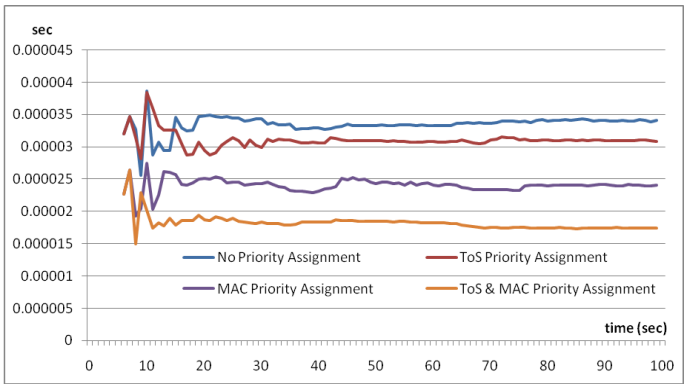


Fig. 15. Packet delay in Ethernet channels

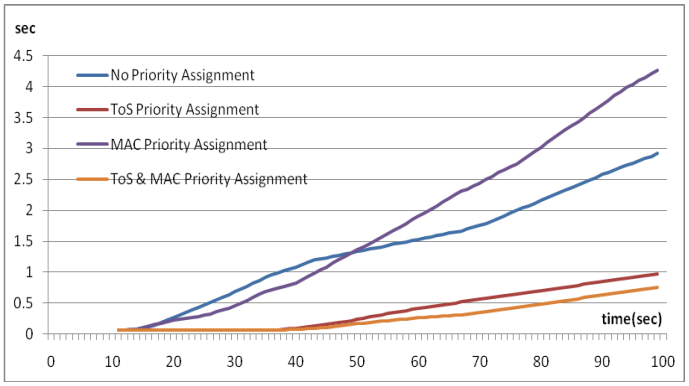


Fig. 16. Packet end-to-end delay for voice users

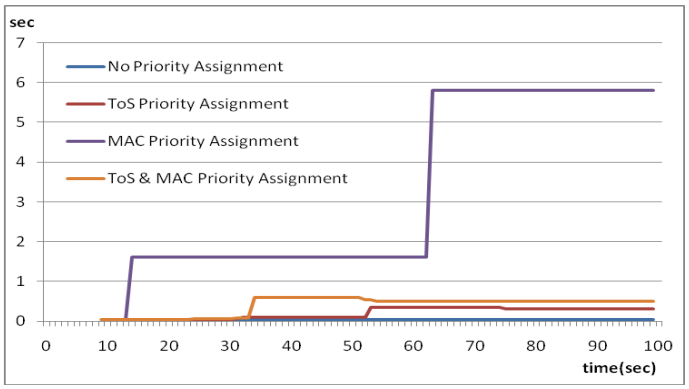


Fig. 17. Packet delay for email, ftp and database applications

CHAPTER 3: Performance Analysis of VoIP Services on the LTE Network

This Chapter is based on contributions of two corresponding papers devoted to the analysis of VoIP services performance in LTE network: one published in Proceeding of Australasian Telecommunication Networks and Applications Conference (ATNAC), and another published in International Journal of Internet Protocol Technology (IJIPT) in 2012.

Today a 3rd Generation Partnership Project Long Term Evolution (3GPP LTE) is considered to be the main standard for deployment in future wireless networks. Hence, performance analysis of LTE is a logical first step toward deployment of this network in B3G infrastructure. In this Chapter performance analysis of LTE is carried using VoIP user applications, which is mainly due to the fact that voice services are known to have the strictest (compared to other network applications) delay requirements: VoIP can only tolerate packet end-to-end delay of up to 100 ms and packet loss of up to 1% [116]. Thus, the ability of LTE to achieve good performance for voice applications will automatically guarantee that QoS for other user applications will be satisfied.

1 Introduction

IP based voice services are already supported by the 3GPP High Speed Packet Access (HSPA) standard, but the importance of VoIP support is even higher for LTE, considered to be the main standard for deployment in future wireless networks. Evolved Universal Terrestrial Radio Access Network (E-UTRAN) is targeted to support a high number of VoIP users. The maximum VoIP capacity of LTE network has been reported by the outage limit defined in TR 25.814 document and updated in report R1-070674 [115].

The uplink capacity of VoIP services in E-UTRAN have been investigated in [126, 127]. These studies took a closer look on the capacity and the coverage of LTE services depending on channel conditions based on physical (PHY) layer functionalities. A number of

studies have been devoted to the effect of semi-persistent packet (SMP) scheduling for voice users [116 - 120]. Various multiuser scheduling strategies (such as Fair Scheduling, Dynamic Subcarrier Assignment, and Adaptive Power Allocation) have been examined in the context of the Orthogonal Frequency Division Multiple Access (OFDMA) downlink in [121 - 125]. All mentioned works analysed the capacity and coverage of LTE only by means of physical layer simulations using physical layer QoS parameters (signal-to-interference-and-noice ratio (SINR), physical layer throughput, etc.). In other words, the effect of different scheduling strategies on the end-to-end QoS for VoIP users has not been considered.

This Chapter has been written based on contributions of two corresponding works devoted to the analysis of VoIP service quality in LTE Frequency Division Duplex (FDD) network. In this Chapter we provide a comprehensive analysis of the Medium Access Control (MAC) layer functionalities and investigate the VoIP service capacity of LTE system using combined PHY/MAC layer simulation model. In particular, we compare the service performance of LTE system with fully dynamic (FD) and semi-persistent packet (SMP) scheduling techniques. We also observe the VoIP capacity of LTE network depending on channel bandwidth, Modulation and Coding Scheme (MCS), link adaptation and Hybrid Automatic Repeat reQuest (HARQ). Unlike the other works where the performance of VoIP services in LTE system is evaluated only by means of PHY layer QoS characteristics, in this Chapter we show the impact of scheduling techniques and PHY/MAC layer design parameters on the end-to-end QoS for voice users.

The rest of the Chapter is organised as follows. First, we provide the background information on LTE radio interface and give a detailed description of LTE MAC layer. Then, we present the simulation model of the network and conclude the Chapter with some simulation results.

2 LTE Radio Interface

In this section some background information on the design, system architecture and radio interface of LTE FDD system is provided. This information will help to understand simulation results presented in this Chapter, and will be used as a reference further in the thesis. More

detailed description of LTE system radio interface can be found in [128].

The LTE based Evolved Packet System (EPS) is an evolution of the 3GPP system architecture where the vision of all-IP network is finally realized. EPS comprises the core network part, called Evolved Packet Core (EPC) and E-UTRAN radio access network, called LTE RAN or simply LTE (Figure 18). The functional split between EPC and LTE is illustrated on Figure 19. EPC provides access to external IP networks and performs a number of the core network related functions (QoS, security, mobility, etc) to terminals in active and idle state. The EPC can also be connected to other 3GPP (such as GERAN/UTRAN, GPRS and UMTS) and non-3GPP (such as WiMAX or cdma2000) networks. LTE performs all radio interface related functions to terminals in active state [129, 130].

As shown on Figure 18, EPC consists of one control-plane node, called a Mobility Management Entity (MME), and two user-plane nodes, called a serving gateway (S-GW) and a packet-data network gateway (P-GW). LTE comprises the base station, called enhanced NodeB (eNB) and mobile terminals, called user equipments (UEs). The eNBs are also connected to EPC by means of the S1 interface. The interface between eNBs is called the X2 interface. The eNBs are also connected to the EPC by means of the S1 interface [129, 130].

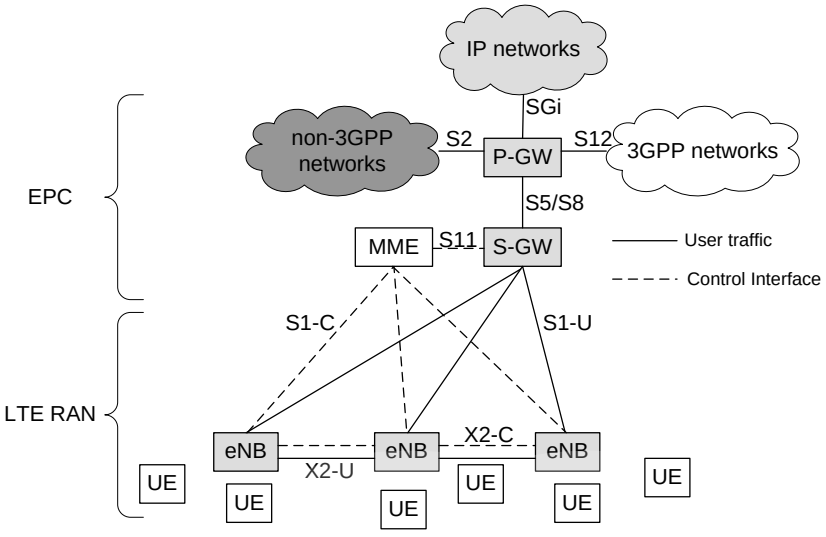


Fig. 18. EPS architecture [129, 130]

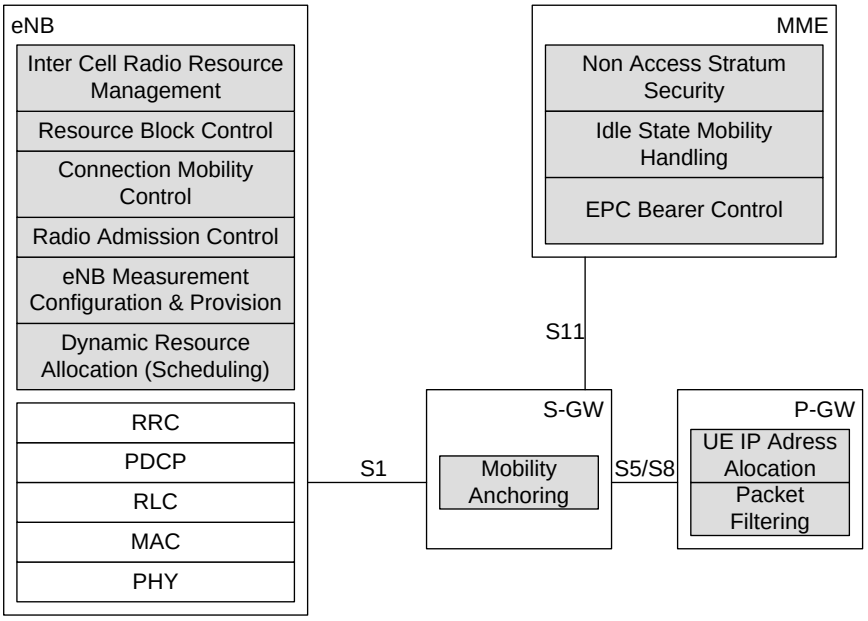


Fig. 19. Functional split between EPC and LTE [130]

The LTE user-plane protocol stack is shown on Figure 20. The physical layer or LTE Layer 1 (L1) is responsible mainly for coding, interleaving and modulation. The LTE Layer 2 (L2) is divided to three sublayers: the Packet Data Convergence Protocol (PDCP), the Radio Link Control (RLC) and the Medium Access Control (MAC) sublayers. The PDCP sublayer performs IP header compression and ciphering, supports lossless mobility in case of inter-eNB handovers, and provides integrity protection to higher layer control protocols. The RLC sublayer provides Automatic Repeat-reQuest (ARQ), data segmentation and concatenation (to minimize the protocol overheads). The MAC sublayer is responsible for Hybrid ARQ (HARQ), scheduling and random access (RA) [129, 130].

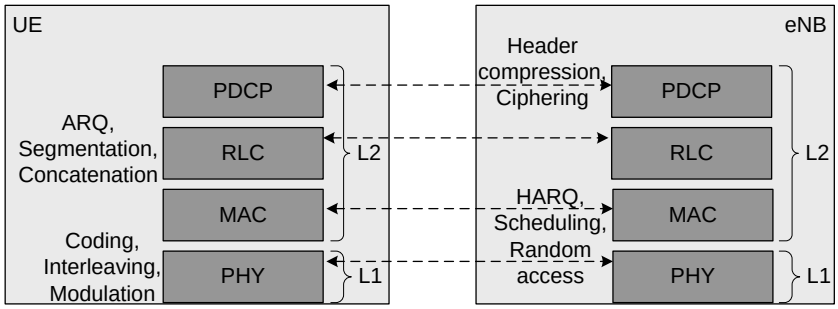


Fig. 20. LTE user-plane protocol stack [129]

The downlink transmission scheme of LTE system is based on conventional Orthogonal Frequency Division Mode (OFDM), where the available spectrum is divided into multiple subcarriers, which are modulated independently by a low rate date stream. The key features of OFDM are robustness against multipath fading and efficient receiver architecture. Besides, OFDMA supports multiple users on the available bandwidth, i.e. within one transmission time interval (TTI) subcarriers can be allocated to different users. The uplink transmission scheme of LTE system is based on Single-Carrier Frequency Division Multiple Access (SC-FDMA), which has better Peak-to-Average Power Ratio (PAPR) properties then OFDMA-based signals [129, 130].

The frame structure of the LTE FDD mode is shown on Figure 21. According to this structure, one radio frame with duration T_f = 10 ms is divided into 10 equal subframes with duration T_{sf} = 1 ms. Each subframe consists of 2 basic time units (slots) with duration T_s = 0.5 ms [131]. A basic radio resource unit in the LTE standard is called a resource block (RB). One RB consists of 12 subcarries with a constant subcarrier spacing Δf = 15 kHz, and has a duration of 1 slot. The number of RBs, N_{RB}, depends on the channel bandwidth. N_{RB} for different bandwidth values is given in Table 5. The capacity of one RB depends on the Modulation and Coding Scheme (MCS) which determines the bit rate. The possible MCS values and their code bit rates are given in Table 6 [132, 133].

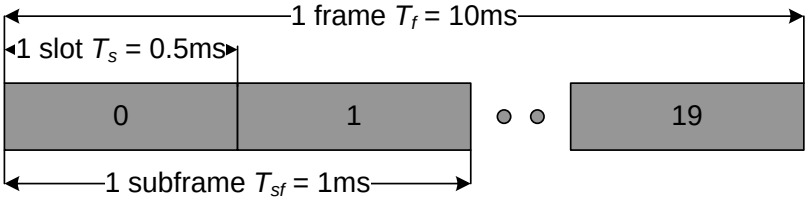


Fig. 21. Frame structure in LTE FDD [131]

Table 5. The number of RBs for different bandwidth [17]

Channel bandwidth, MHz	*1.4*	*3*	*5*	*10*	*15*	*20*
Number of resource blocks, N_{RB}	6	15	25	50	75	100

Table 6. MCS Description [131]

MCS index	*Modulation*	*Coding Rate*	*MCS index*	*Modulation*	*Coding Rate*
0	-	-	8	16QAM	0.478516
1	QPSK	0.076172	9	16QAM	0.601563
2	QPSK	0.117188	10	64QAM	0.455078
3	QPSK	0.188477	11	64QAM	0.553711
4	QPSK	0.300781	12	64QAM	0.650391
5	QPSK	0.438477	13	64QAM	0.753906
6	QPSK	0.587891	14	64QAM	0.852539
7	16QAM	0.369141	15	64QAM	0.925781

3 LTE MAC Layer Design

3.1 HARQ and Link Adaptation

As in any communication system, wireless LTE channels experience occasional transmission errors due to noise, interference, and/or fading. Since most of the RLC protocols are not prepared to deal with errors in packets, in LTE system the erroneous packets are dropped on MAC

sublayer without forwarding to higher layers. The MAC-based HARQ scheme is used for this purpose [130].

HARQ is a combination of Forward Error Correction (FEC) with ARQ. It enables to compensate for errors and to provide a better throughput performance. To detect the errors, the receiver node uses the Cyclic Redundancy Check (CRC). The standard generator polynomials for parity bits used in LTE system are given in [134]. If the receiver node detects erroneous packet, it will discard it, and will send Negative Acknowledgement (NACK) message to the packet sender node. The packet sender node will retransmit the packet in 8 ms after receiving the NACK message. This process will be repeated until the positive Acknowledgement (ACK) message is received or until the maximum retransmission limit will be reached. If the maximum retransmission limit for a packet is reached, and it still contains the errors, it will be dropped (packet loss due to channel errors) [129, 134].

Three types of HARQ can be deployed in LTE system: HARQ Type I, HARQ with Chase Combining (CC) and HARQ with Incremental Redundancy (IR). In HARQ type I erroneous packets with error are simply retransmitted until the ACK message received, or maximum retransmission limit is reached. In CC after receiving a NACK message, a node retransmits the packet with the same data and parity bit pattern as the original packet. The receiver node combines erroneous packet with its retransmission, and sends the combined signal to the decoder. CC increases the accumulated received signal to noise ratio for each retransmission, but does not give any additional coding gain. In IR after receiving a NACK message, a node retransmits the punctured data and parity bit pattern different from the original packet. To detect the error, the receiver node combines the original (erroneous) packet with the retransmitted. Thus, IR results in a higher coding gain when compared to CC [134].

For signaling, HARQ uses the following information: HARQ process number (currently, LTE supports up to 8 parallel HARQ processes are supported per UE); new data indicator (indicates whether the packet is a new transmission or a retransmission); the redundancy version (each redundancy version corresponds to a different set of parity bits); ACK/NACK [134].

In LTE system, HARQ is combined with Adaptive Modulation and Coding (AMC) to maximize the data rate by adjusting transmission

parameters to the current channel conditions. AMC is one of the realizations of dynamic link adaptation. In AMC algorithm the appropriate MCS for packet transmissions is assigned periodically (within short fixed time interval usually equal 1 TTI) by eNB based on instantaneous channel conditions reported to eNB by UEs. The higher MCS values are allocated to the channels with good channel quality to achieve higher transmission rate and throughput. The lower MCS are assigned to the channels with poor channel quality to decline the transmission rate and, consequently, to ensure the transmission quality [134].

The method for choosing MCS can be expressed as follows. Based on the instantaneous radio channel conditions the signal-to-interference-and-noise ratio (SINR) is calculated for each UE. Assume that entire SNR range is partitioned into is expressed as [134]:

$$MCS = \begin{cases} MCS_1, & SINR < \gamma_1 \\ MCS_2, & \gamma_1 \le SINR < \gamma_2 \\ MCS_3, & \gamma_2 \le SINR < \gamma_3 \\ \quad : & \quad : \\ MCS_m, & \gamma_{m-1} \le SINR < \gamma_m \end{cases} \tag{23}$$

where SINR is the signal-to-interference-and-noise ratio of the channel between UE and eNB; γ_i is the SINR threshold corresponding to -10dB bit error ratio (BER) given by the Additive White Gaussian Noise (AWGN) curves for each MCS (the standard AWGN curves can be found for instance in [135]). The LTE standard defines m = 29 MCS values [131]. However, since a UE only have 4 bits feedback to indicate its preferred MCS, the eNB uses only the first 15 MCS levels from the list provided in [131].

3.2 Random Access Channel Procedure

A Random Access Channel procedure (RACH) is used in LTE system for initial access, i.e. for originating, terminating or registration call in the network. The objective of RACH is to keep the transmissions from different UEs aligned with the frame timing at the eNB [129, 133].

Two types of a random access procedure are defined in LTE standard: contention-based RACH and non-contention-based RACH. The difference between these two types of RACH in that in contention-based RACH there is a possibility for failure in case if overlapping random access (RA) preambles, whereas in non-contention-based RACH RA preambles are unique for each UE [129, 133].

A contention-based RACH is illustrated on Figure 22. It is implemented in four steps:

1. a UE randomly selects a 5-bit long RA preamble sequence from the set of sequences available in the cell, and transmits it on an RA channels;
2. a eNB detects the preamble transmission, estimates the uplink transmission timing of the UE, and responds by providing the UE with the correct timing-advance value to be used for subsequent transmissions and with a first grant for an uplink transmission;
3. since it is possible that multiple UEs attempted RA with the same RA preamble sequence on the same RA channel, the UE provides its identity to the eNB with the first scheduled uplink transmission;
4. the eNB resolves the (potential) contention by echoing the received UE identity back. The UE, seeing it own identity echoed back, concludes that RA was successful and proceeds with the time-alignment [133].

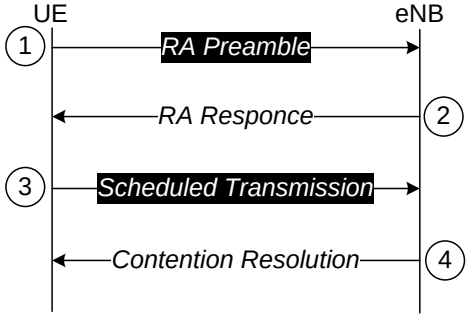


Fig. 22. Contention-based RACH [133]

A non-contention-based RACH is illustrated in Figure 23. It is implemented in three steps:

1. the eNB assigns the 5-bit long RA preamble to a UE, and transmits it on an RA channels;

2. the UE transmits the assigned preamble to eNB;
3. a eNB detects the preamble transmission, estimates the uplink transmission timing of the UE, and responds by providing the UE with the correct timing-advance value to be used for subsequent transmissions and with a first grant for an uplink transmission [133].

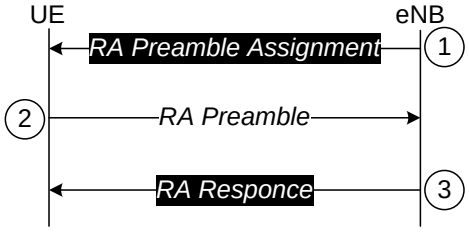


Fig. 23. Non-contention-based RACH [133]

3.3 Packet Scheduling Procedure

The LTE standard is based on the packet scheduling (PS) domain where the packets are normally scheduled using the FD packet scheduler, which allocates available resources to UEs separately for every packet transmission. Resources are allocated to UEs for uplink and downlink data transmission in terms of RBs. Thus, one UE can be allocated only the integer number of RBs in frequency domain, and these RBs do not have to be adjacent to each other. Resource allocation (scheduling) is usually performed periodically within a fixed time interval (scheduling period) with minimal duration 1 TTI. The scheduling is done by L2 packet scheduler in the eNB both for uplink and downlink transmissions. Depending on the implementation, the packet scheduling can be based on the quality of service (QoS) requirements, instantaneous channel conditions, fairness, etc. Besides, the scheduler has to ensure that HARQ retransmissions are performed on a timely basis (in LTE system a packet retransmission should be send in exactly 8 ms after receiving a NACK message) [129, 132, 133]. After resource allocation, the user data are carried by the PUSCH in uplink direction and Physical Downlink Shared Channel PDSCH in downlink direction. The scheduling decisions are carried by the PUCCH and PDCCH in uplink and downlink directions, respectively [131].

Being very flexible a fully dynamic scheduling is ideal for bursty, infrequent and bandwidth consuming data transmissions (e.g. web surfing, video streaming, emails), but less suited for real time streaming applications like voice. In FD, the average number of control channels per TTI (the number of control channels *#CCH*), can be estimated using [118]:

$$\#CCH = \lambda n(v / I_1 + (1-v) / I_2) \quad (24)$$

where v is the Voice Activity Factor (VAF); n is the total number of voice users; λ is the average number of transmissions; I_1 and I_2 are the inter-arrival times of voice packets and SID packets, respectively. For 5MHz bandwidth and 100 voice users using G.723.1(12.2 kbps) codec the number of control channels estimated according to (24) will be equal *#CCH* = 4.17 (a typical assumption is 6-10 downlink control channels per TTI for uplink traffic in the 5 MHz bandwidth [136]).

To reduce the amount of L1/L2 control signaling, a so-called Semi-Persistent Scheduling (SMP) has been proposed for the VoIP traffic [133]. The SMP persistent scheduling is used for initial transmissions, and dynamic scheduling for retransmissions. For initial transmissions, Radio Resource Control (RRC) signaling is used to allocate the PRBs and transmission parameters including the MCS to voice users at the beginning of an active or an inactive period (signaling for inactive period is necessary to notify the PS, that UE does not need any resources to keep effective channel utilization). The semi-persistent allocation technique is valid until the UE receives another control channel indication, which happens when the channel conditions have changed. The allocations for initial transmissions are sent on the PDCCH and PUCCH. Retransmissions are scheduled dynamically using the L1/L2 control channels.

The scheduling priority order for SMP PS in frequency domain can be described as follows:

1. reserve resources for HARQ retransmissions;
2. schedule semi-persistent UEs on pre-assigned resources;
3. schedule dynamic UEs;
4. schedule HARQ transmissions on the reserved resources [133].

Note, that since available retransmission resources in SMP are few compared to the FD case, the dynamic retransmission is very important

for the SMP, because it allows the scheduler to utilize any unused PRBs to increase the amount of retransmission opportunities [118, 113].

Assuming that initial transmissions are scheduled using SMP, the average number of control channels necessary per TTI can be estimated according to [118]:

$$\#CCH = (\lambda - 1)n(\nu / I_1 + (1-\nu) / I_2) \quad (25)$$

For 5MHz bandwidth and 100 voice users using G.723.1(12.2 kbps) codec the number of control channels estimated according to (25) will be equal , *#CCH* = 0.69, i.e. approximately 17% of that number in case of FD PS.

3.4 Buffer Status Report Procedure

A Buffer Status Report (BSR) is used to provide the network with information about the amount of data in the uplink buffer of a UE. For this, BSR generates a BSR MAC Control Element, which includes information about the amount of data available for transmission in the RLC and PDCP layers, when being triggered. The BSR shall be triggered when the uplink data becomes available for transmission and the data belongs to a logical channel with a higher priority than those for which data already existed in the UE transmission buffer, in this case the BSR is referred to as "Regular BSR." When the BSR is triggered but there is no allocated resource for a new transmission, the UE would then trigger SR procedure for requesting uplink resources, i.e. Uplink Shared Channel (UL-SCH) resources, to send the BSR MAC Control Element. Only the Regular BSR can trigger the SR procedure when the UE has no UL resources allocated for a new transmission for a current TTI [133].

When a BSR MAC Control Element is transmitted, the network may not be able to successfully receive the BSR MAC Control Element, and thus would not allocate any uplink transmission resources to the UE. In this case, if the reason for triggering the BSR is no longer satisfied, such as no higher priority data becomes available for transmission. For example, the UE would have no uplink transmission resources for use and enter into a "deadlock" situation. In this case the BSR retransmission mechanism will be applied, which utilizes the

Retransmission BSR Timer to enhance the reliability of BSR transmission. The UE starts the timer when a BSR MAC control element is generated, and restarts (if running) the timer when UL resources allocated for new transmission are received, e.g. on the PDCCH or in a Random Access Response, which indicates the BSR MAC Control Element is successfully received by the network. When the timer expires and the UE has data available for transmission in the buffer, the UE shall trigger a BSR, in which case the BSR is also referred to as "Regular BSR" [133].

4 Simulation Model of LTE Network

We consider a basic LTE FDD simulation model illustrated on Figure 24. The model consists of one eNB, one EPC and a communication server, connected to each other using IP-based links with 1Gbit/s data rate. The eNB serves a number of fixed VoIP users randomly positioned in the system area with a 1000m radius.

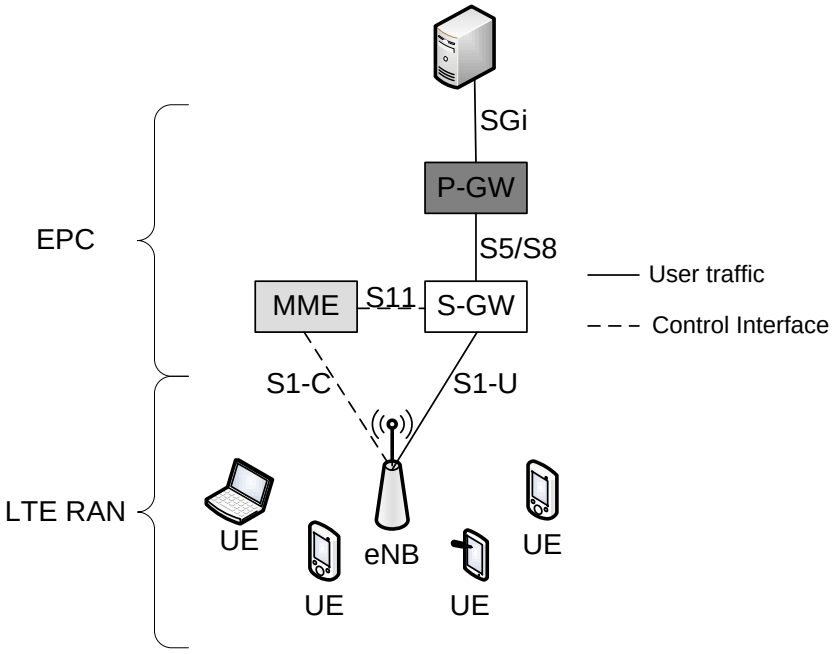


Fig. 24. Simulation model of the network

The radio channel between eNB and each UE is calculated according to the path loss model provided in the ITU-T Recommendation M.1225 [137] for the outdoor to indoor and pedestrian test environments. Log-normal shadow fading is assumed with a standard deviation of 10 dB for outdoor users and 12 dB for indoor users. The Spatial Channel Model (SCM) is used for multipath fading. Transmitter/Receiver antenna gains are equal 10dBi for the pedestrian environment and 2 dBi for the indoor environment. Receiver noise figure and thermal noise density are equal 5dB and -174 dBm/Hz, respectively. The average building penetration loss is 12 dB with a standard deviation of 8 dB. Other losses (in cables, connectors, combiners) are assumed to be equal 2dB.

In all simulations a de-coupled Round Robin (RR) packet scheduler in the time domain and Proportional Fair (PF) throughput packet scheduler in the frequency domain (described in details in [132]) is utilized. The structure of the scheduler is illustrated on Figure 25. In time domain RR scheduler assigns the priority metrics to all users connected to eNB based on their average and predicted throughput, and appends the users with the largest priority metric to a so-called Scheduling Candidate Set (SCS). SCS is then passed to the frequency domain PF packet scheduler. PF allocates RBs to the users in SCS (starting from the user with the highest priory metric) based on their average and predicted throughput. Described scheduling procedure provides flexibility (since both domains can be configured separately), reduces the overall complexity of scheduling algorithm, and insures that available resources are shared equally among the users when they have same load and channel conditions [132].

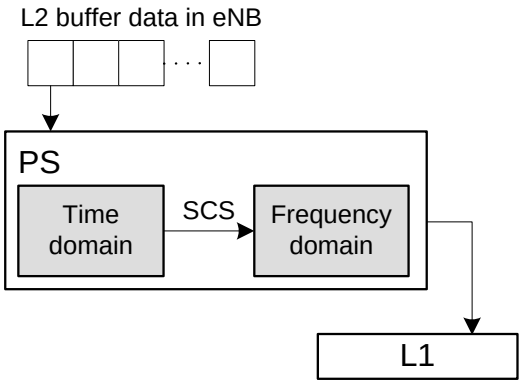


Fig. 25. The structure of the decoupled packet scheduler [132]

Other simulation parameters of the network model are summarized in Table 7. In all simulations, a non-contention-base RACH is deployed.

Table 7. Common Simulation Parameters

Parameter		*Value*
PHY profile:	Operation mode	FDD
	Cyclic Prefix Type	Normal (7 symb/slot)
	EPC Bearer Definitions	348kbit/s (Non-GBR)
	Carrier frequency	2GHz
	Subcarrier spacing	15kHz
BSR Parameters:	Periodic Timer	5 subframes
	Retransmission Timer	2560 subframes
L1/L2 Parameters:	Reserved Size	2 RBs
	Allocation Periodicity	1 TTI
HARQ Parameters:	Max No of Retransmissions	3
	HARQ Retransmission Timer	8 TTI
	Max No of HARQ processes	8 per UE

The VoIP services model deployed in simulation complies with the requirements of the real-time delivery session initiation and session description protocols (SIP/SDP) provided in [138, 139]. During each session, a VoIP user might be either in active (talk-spurts) or inactive (silent) state. The duration of each state is exponentially distributed with burst lengths of 0.65s and 0.352s, respectively.

Two voice coders are considered: G.711 (64 Kbps) and G.723.1 (12.2 Kbps). The codec payload size is equal 160 bytes for G.711 and 40 bytes for G.723.1 coder. The payload generation intervals of G.711 and G.723 coders are 20ms and 30ms, respectively. The Silence Insertion Descriptor (SID) packet inter-arrival time is 160ms for both coders. For bandwidth calculations it is considered that packet payload is typically adding 6 bytes for the L2 header, and 2 bytes for the RTP/UDP/IP compressed header [140]. Discontinuous Transmission, Voice Activity Detection and Comfort Noise Generation are also applied.

Simulation scenarios used for performance analysis of the network are listed in Table 8. In scenarios with ideal channel conditions, no link adaptation and no HARQ are applied (since they are unnecessary due to absence of errors). First, we compare the capacity of a LTE network using two voice coders. For all other simulations the G.723.1 coder has been deployed. We also simulate a scenario with semi-persistent scheduling in order to observe its influence on the VoIP capacity.

Table 8. Simulation Scenarios

No	*Simulation Description*			
	Channel	***Link Adaptation***	***HARQ***	***Packet Scheduling***
1.	5MHz ideal	MCS = 9	No HARQ	FD
2.	5MHz ideal	MCS = 15	No HARQ	FD
3.	10MHz ideal	MCS = 9	No HARQ	FD
4.	5MHz ideal	MCS = 9	No HARQ	SMP
5.	5MHz real	AMC	HARQ Type I	FD
6.	5MHz real	AMC	HARQ with CC	FD

In scenarios with real channel conditions, the link adaptation (AMC) and HARQ are applied. Based on the instantaneous radio channel conditions the SINR is calculated for each UE. The link adaptation algorithm maximizes the spectral efficiency (SE) by choosing the best MCS for a given SINR (detailed description of AMC algorithm was provided in III.A). Two types of HARQ are simulated: in HARQ type I packets with error are simply retransmitted until the errorless packets received, or retransmission limit (up to 3 retransmissions) is reached. In other scenario, HARQ with chase combining is applied (description of HARQ CC were given in III.A). In this mechanism we deploy 8 parallel stop-and-wait processes per UE in uplink and downlink directions.

5 Simulation Results

5.1 VoIP Service Performance in Ideal Channel Conditions

In this subsection performance of VoIP services in ideal channel conditions is briefly summarized. More comprehensive performance analysis can be found in the corresponding papers. In all scenarios with ideal channel conditions no AMC and no HARQ applied since they are unnecessary due to the absence of channel errors.

First, we compare the capacity of an LTE FDD network with 5MHz bandwidth and MCS = 9 for voice users using G.711 and G.723.1 coders. Figure 26 shows the mean MAC-layer uplink and downlink packet delay in LTE radio interface; Figure 27 shows the mean uplink and downlink packet loss; Figure 28 shows the mean application layer packet end-to-end delay (comprising the network, codec and play out delays) for VoIP users.

Results show that the lower bit rate codec rate (G.723.1) provides much higher capacity than the higher codec rate (G.711). According to [115], the satisfactory level of service for VoIP users is achieved when packet end-to-end delay is less than 100ms, and packet loss is less than 2%). With defined QoS levels the network can support not more than 30 users using G.711 coder, and up to 80-90 users using G723.1 coder. Results also show, that the downlink offers much higher capacity than uplink because of the following reasons:

1. the downlink spectral efficiency of Multiple-Input-Multiple-Output (MIMO) OFDMA is higher than the uplink spectral efficiency of SC-FDMA [14];
2. the downlink delay comprises only the buffering, transmission, queuing and processing delay components, while in the uplink the delay comprises not only buffering, transmission, queuing and processing delay components, but also the delay due to uplink packet scheduling [14].

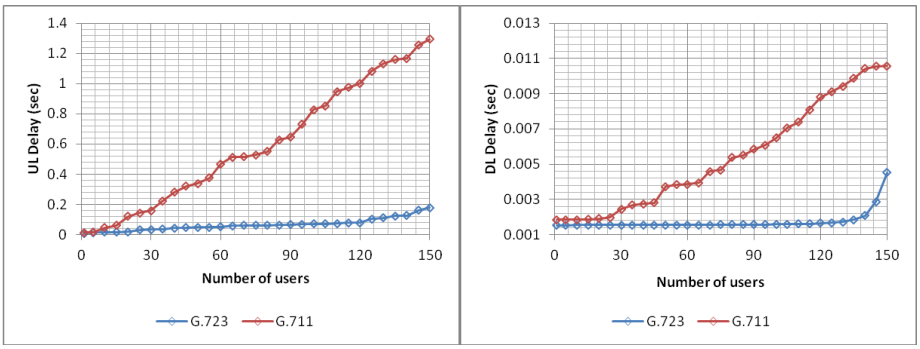


Fig. 26. Mean MAC-layer packet delay in UL and DL channels

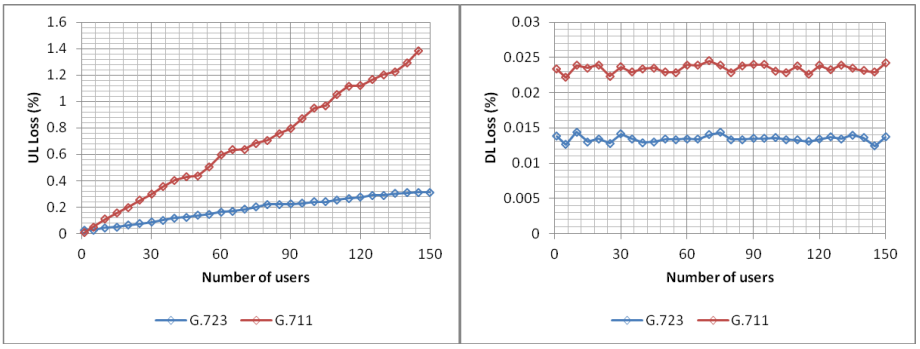


Fig. 27. Mean packet loss in UL and DL channels

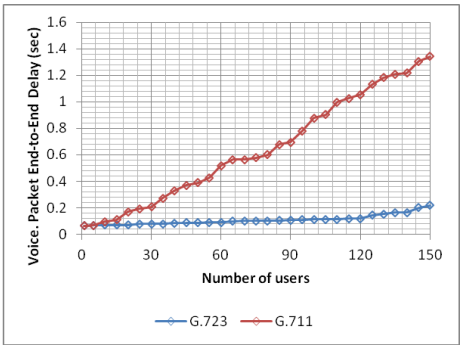


Fig. 28. Mean application-layer packet end-to-end delay

The maximal achievable bit rate in LTE network depends on the channel bandwidth and MCS. Figures 29 - 31 show the capacity of LTE network for VoIP users using G.723.1 coder in first three scenarios (i.e. for 5MHz&MCS9, 5MHz&MCS15 and 10MHz&MCS9). Results

show that the capacity of LTE network for VoIP users is the highest in scenario 3 (10MHz&MCS9), and the lower in scenario 1 (5MHz&MCS9). Such results correspond to the theoretical channel capacity which can be estimated using the modified Shannon Capacity expression given by [142]:

$$R = B\eta_{BW} \log_2(1+\frac{SNR}{\eta_{SE}}) \tag{26}$$

where R is the data rate, B is the bandwidth, SNR is a signal/noise ratio, η_{BW} and η_{SE} are the bandwidth efficiency and the spectral efficiency of the LTE network model respectively.

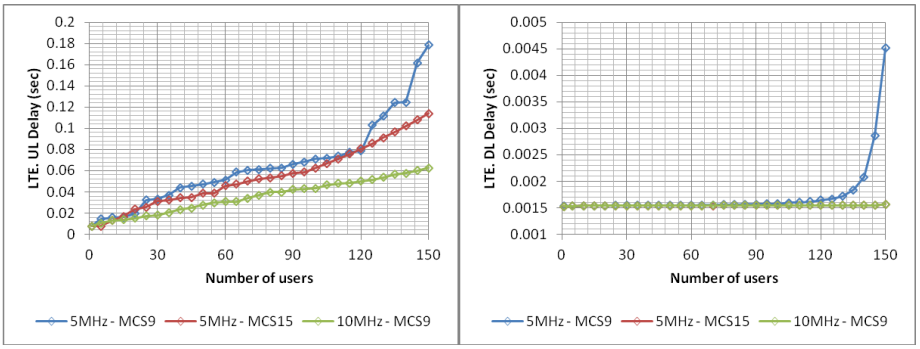


Fig. 29. Mean MAC-layer packet delay in UL and DL channels

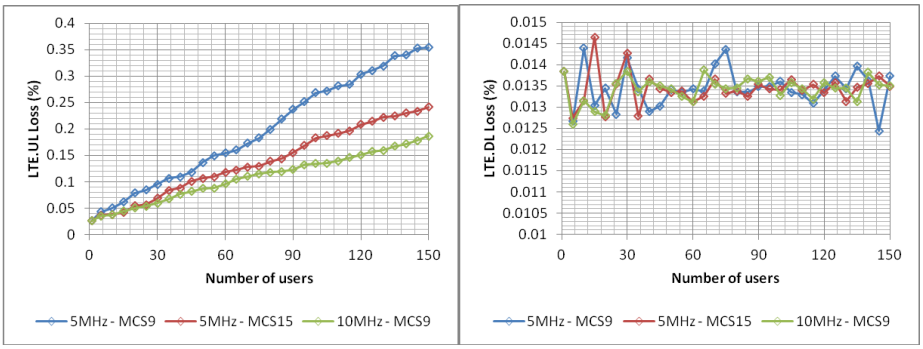


Fig. 30. Mean packet loss in UL and DL channels

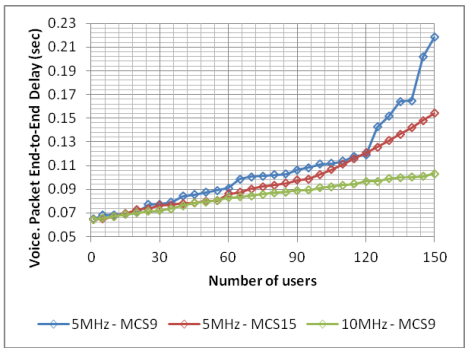


Fig. 31. Mean application-layer packet end-to-end delay

In (26) the bandwidth efficiency of a LTE network is reduced by several other issues including the Adjacent Channel Leakage Ratio (ACLR), Cyclic Prefix, Reference Signals, Synchronization Signal, Random Access Preamble and L1/L2 Control Channel overheads. Due to the requirements of the ACLR, the bandwidth efficiency of 0.9 is for a single antenna configuration considered in simulations. The other overheads reduce the downlink efficiency up to 0.62, and 0.78 for the uplink channel [143].

Full SNR efficiency is not possible in LTE due to limited code block length which depends on the link adaptation and scheduling. There are also restrictions to the maximum spectral efficiency from the supported values of MCS. In other words, the SNR efficiency is much more complicated to analytically compute than the bandwidth efficiency. Therefore, in this paper the value of η_{SE} is extracted by using curve fitting to link-level simulation results provided in [144], which gives us η_{SE} = 0.88 for uplink and downlink directions.

With defined QoS limits the experimental VoIP capacity for scenario 1 is 80 users, for scenario 2 - 100 users, and for scenario 3 - 140 users, which corresponds to the values of theoretical channel capacity in MHz obtained from expression (26).

The effect of packet scheduling procedure (FD or SMP) on QoS for VoIP users using G.723.1 coder is illustrated on Figures 32 - 34. Results show that SMP can increase the capacity of LTE network for voice users. However, this capacity improvement is rather small and incomparable with that when the bandwidth of the network is increased to 10 MHz.

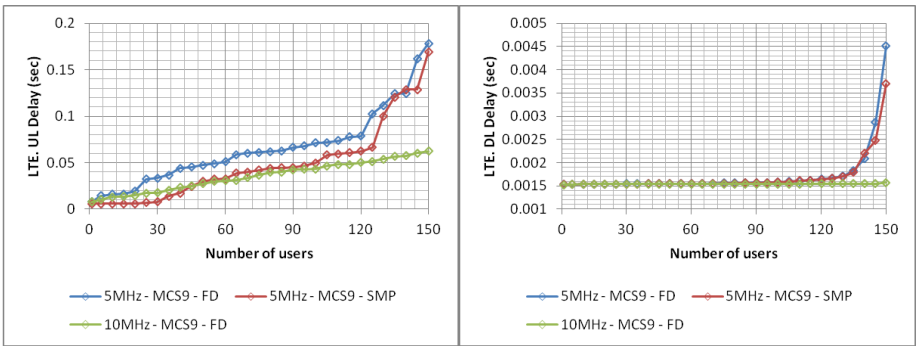


Fig. 32. Mean MAC-layer packet delay in UL and DL channels

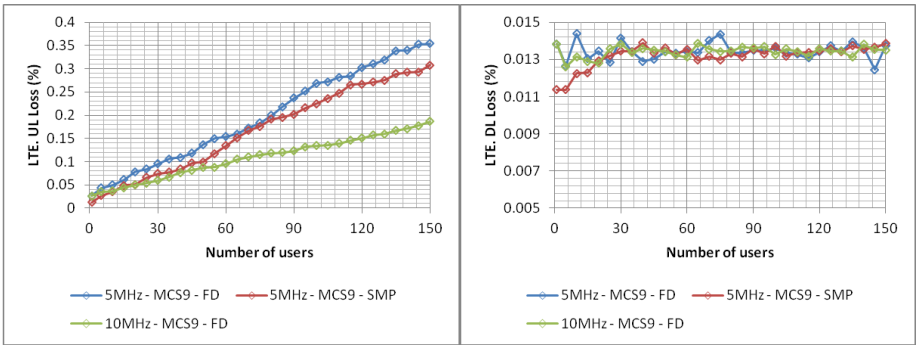


Fig. 33. Mean packet loss in UL and DL channels

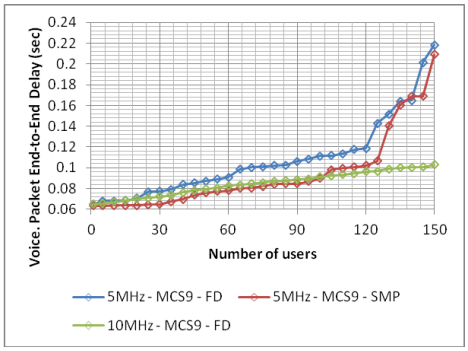


Fig. 34. Mean application-layer packet end-to-end delay

5.2 VoIP Service Performance in Real Channel Conditions

In this subsection performance of LTE network for VoIP users in real channel conditions is briefly described.

Figures 35 and 36 illustrate performance of the network in scenarios 5 and 6 (i.e. when different types of HARQ are applied). All results are given for voice users using G.723.1 coder. Graphs show that HARQ with CC decreases the packet error ratio (PER) by accumulating received signal (detailed description of different types of HARQ was provided in III.A). This means that HARQ with CC needs smaller number of retransmissions to successfully decode the channel errors than HARQ Type I. As a result, the mean packet delay and loss values achieved by HARQ with CC are smaller than those in HARQ Type I (since smaller amount of time is required for retransmissions, and less packets are lost due to channel errors).

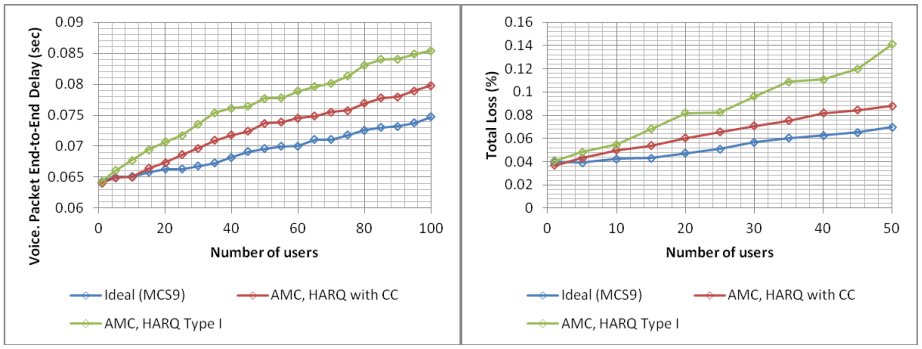


Fig. 35. Mean application-layer packet end-to-end delay and total (UL and DL) packet delay

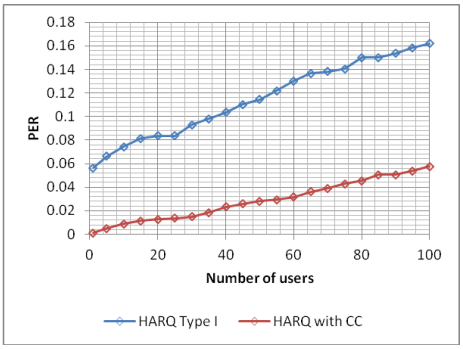


Fig. 36. Mean packet error ratio

To finalize, we provide the summary of LTE network performance for VoIP users below.

1. The VoIP capacity of LTE network can be increased by significant margin if the low bit rate codec, such as G.723.1, is used for VoIP services.
2. The SMP packet scheduling technique which has been proposed to reduce delay for voice services offers only a slight QoS advantages when compared to FD technique.
3. In contrast to the idea that ARQ techniques in general are not suitable for voice services, HARQ with CC gives substantial improvement in terms of packet error ratio, mean packet delay and loss when compared to simple HARQ Type I technique.

CHAPTER 4: Prediction Based Bandwidth Allocation for Cognitive LTE Network

In this chapter a resource allocation technique for LTE-based CRN in Scenario 1 is presented. Description of the network deployment scenario has already been provided in Overview of this thesis. Here we outline the algorithm for resource allocation, show its implementation in LTE-based network infrastructure and analyze the algorithm performance based on results of simulations in OPNET environment [112]. The corresponding paper is published in Proceeding of IEEE Wireless Communications and Networking Conference (WCNC), 2013.

1 Introduction

In 2004, a special IEEE 802.22 working group was set to develop a first worldwide wireless standard for cognitive radio (CR). It was proposed that the fixed wireless access will be provided by a Wireless Regional Area Network (WRAN) comprising a number of service providers (SPs) with their base stations. Unlike traditional wireless networks, where each SP operates on its fixed licensed bandwidth, in IEEE802.22 architecture SPs share the total available spectrum among each other according to some flexible spectrum usage policy to maximize the QoS for their users [51, 52].

According to the standard, the physical layer of IEEE802.22 network will use OFDMA technology, while the MAC payer will be based on cognitive radio. The exact algorithm for dynamic spectrum access (DSA) is not specified: the choice here is left for the network developers [52]. Consequently, a number of centralized and distributed resource allocation strategies have been proposed for IEEE802.22 network architecture (see, e.g. [145 - 153]). Most of these strategies are very effective in homogeneous traffic environment, but not very efficient in case of heterogeneous network applications, which is mainly due to the fact that all network users are characterized by similar utility functions (for instance, user throughput or service rate) [153].

To deal with heterogeneous network applications, we propose to apply the concept used in optimal flow and congestion control (OFC) to resource allocation in IEEE802.22 CRN. Recall, that in traditional OFC the network congestions are prevented by assigning the resources based on speed of load increase in the bottleneck nodes [65]. Considering, that the congestions in LTE network usually occur in eNBs (the growth of the queues in user terminals is neglectably small and usually do not lead to significant increase of delay or loss in the network), we suggest to allocate the network resources based on the speed on load increase in eNBs using the appropriate load control indicators. Unlike most of the OFC algorithms where the users are described by simple binary load indicators (for congested and not congested nodes) [65, 154], we use the modified load indicators (MLI) described by more complex functions depending on queue size and the loss in the nodes. To further increase the algorithm efficiency, we allocate the resources based on predicted load and channel state information, which may help to prevent the potential growth of delay and loss for the users.

The rest of the Chapter is organized as follows. In Section 2 we outline the algorithm for resource allocation. In Section 3 we show how the OFC approach can be applied for resource allocation in LTE-based CRN, and present the weighted utility maximization framework as a justification for used of the algorithm. In Section 4 we illustrate the performance of the proposed algorithm based on simulation model developed using OPNET platform [112].

2 Resource Allocation Algorithm

Consider the network model which comprises n eNBs sharing the total available bandwidth C using the network resource manager (NRM) located in LTE Evolved Packet Core (EPC). NRM is connected to eNBs via high speed Internet Protocol (IP) based links to enable fast transmission of data and control information.

The assumptions of the network model are summarized below:

1. uplink and downlink buffers of eNBs have known finite capacities denoted by $Q^{UL}{}_{max}$ and $Q^{DL}{}_{max}$;

2. parameter monitoring of parameters, bandwidth allocation and prediction can be performed only discontinuously within fixed time intervals (called monitoring intervals);
3. the length of the monitoring interval, the length of buffers and the amount of data arrived at the buffers during each monitoring interval are known;
4. data collected and predicted separately for uplink and downlink directions.

Proposed prediction based resource allocation (PRA) algorithm operates on a discrete-time basis, i.e. the time axis in the algorithm is partitioned into mutually disjoint intervals $\{[t\Delta t, (t+1)\Delta t]\}$, with t denoting the index of time interval. Prediction and resource allocation approach is synchronized with these intervals, i.e. Δt is equal to the duration of monitoring interval. At the beginning of each interval the eNBs are allocated some bandwidth based on prediction made from the data collected in the past.

We shall use the following notation:

n– the number of nodes (eNBs) in the network;

t – integer valued index of a monitoring interval;

x_i – bandwidth (in MHz) allocated to eNB i at $(t + 1)^{th}$ monitoring interval;

$Q_i^{UL}(t)$, $Q_i^{DL}(t)$, $Q_i(t)$ –length of the queue (in bits) at eNB i at t^{th} monitoring interval on the uplink, the downlink and the unspecified (general) channels, respectively;

$X_i^{UL}(t)$, $X_i^{DL}(t)$, $X_i(t)$ – the number of bits served at eNB i at t^{th} monitoring interval on the uplink, the downlink and the unspecified (general) channels, respectively;

$L_i^{UL}(t)$, $L_i^{DL}(t)$, $L_i(t)$ – the number of bits arrived to eNB i at t^{th} monitoring interval on the uplink, the downlink and the unspecified (general) channels, respectively;

$D_i^{UL}(t)$, $D_i^{DL}(t)$, $D_i(t)$ – the number of bits dropped at eNB i at t^{th} monitoring interval on the uplink, the downlink and the unspecified (general) channels, respectively.

Proposed PRA algorithm is summarized below:

1. **Input:** The eNBs monitor $\{Q_i^{UL}(t), Q_i^{DL}(t)\}_{i=1}^{n}$, $\{L_i^{UL}(t), L_i^{DL}(t)\}_{i=1}^{n}$ and send this information to the NRM using IP links.

2. **Prediction:** Using the input information collected up to t^{th} monitoring interval, NRM computes the predictions $\{\hat{L}_i^{UL}(t+1), \hat{L}_i^{DL}(t+1)\}_{i=1}^{n}$ using PLR parameter estimation applied with AR(1) time-series model (detailed description of PLR technique and AR model has already been provided in Chapter 1).
3. **Weight generation:** Based on the $\{\hat{L}_i^{UL}(t+1), \hat{L}_i^{DL}(t+1)\}_{i=1}^{n}$ and $\{Q_i^{UL}(t), Q_i^{DL}(t)\}_{i=1}^{n}$, NRM generates the weights

$$\omega_i = \frac{\left[Q_i^{UL}(t) + \hat{L}_i^{UL}(t+1) - \frac{x_i \Delta t}{SE^{UL}}\right]^+}{Q_i^{UL}(t)} + \frac{\left[Q_i^{DL}(t) + \hat{L}_i^{DL}(t+1) - \frac{x_i \Delta t}{SE^{DL}}\right]^+}{Q_i^{DL}(t)}, \quad 1 \le i \le n \tag{27}$$

where $[x]^+$ denotes $max(0, x)$. These weights are used later in weighted proportional fairness problem. The justification behind using the expression (27) will be given in the next section.

4. **Optimization:** Using the generated weights $\{\omega_i\}_{i=1}^{n}$, the NRM calculates the bandwidth $\{x_i\}_{i=1}^{n}$ that will be assigned to each eNB in the next $(t+1)^{th}$ monitoring interval based on the weighted proportional fairness criterion, and transmit these values to corresponding eNBs via IP links.
5. **Output:** Received bandwidth allocation values $\{x_i\}_{i=1}^{n}$ are assigned to the physical interfaces of the corresponding eNBs at the beginning of the $(t + 1)^{th}$ monitoring interval.

Since the sizes of the node buffers in the network model are limited, the length of node buffer $Q(t)$ is a nonlinear function of t, which can be readily verified from the Lindley's equation [80] for a finite buffer length given by

$$\begin{aligned} Q(t+1) &= \min\left\{Q_{\max}, \left[Q(t) + L(t+1) - X(t+1)\right]^+\right\} \\ &= \left[Q(t) + L(t+1) - X(t+1) - D(t+1)\right]^+ \end{aligned} \tag{28}$$

where Q_{max} is the maximal buffer size of the queue. This step makes $Q(t)$ inappropriate to use in the linear prediction. Therefore, instead of predicting the nonlinear parameter $Q(t)$, we predict the traffic arrived to each eNB, $L(t)$.

3 OFC for Resource Allocation in PRA Algorithm

In this section we show how OFC approach can be applied for resource allocation in LTE-based CRN. To prevent congestion in the bottleneck nodes, OFC algorithm: (i) identifies the nodes with increasing (decreasing) load (buffer size); (ii) increase (decrease) the channel utilization in the nodes with decreasing (increasing) load [65].

In LTE network, the congestions usually occur in the queues at eNBs, where the arrival traffic may be very high, whereas the size of the queues is limited. Hence, to avoid the congestions in CRN, PRA algorithm should: (i) identify eNBs with increasing (decreasing) queue size; (ii) increase (decrease) service rates of eNBs with increasing (decreasing) queue size. The corresponding bandwidth allocation problem can be formulated using, for instance, the weighted proportional fairness criterion [66], and applying it in the way described below.

Consider a network model comprising n eNBs, each characterized by some positive weight (load indicator) ω_i, and sharing the total available bandwidth C. Let x_i be the bandwidth allocated to eNB i at $(t + 1)^{th}$ monitoring interval. Then a weight proportionally fair bandwidth allocation should solve the optimization problem given by:

$$\begin{aligned} \text{maximize} \quad & f(x) = \sum_{i=1}^{n} \omega_i \log x_i \\ \text{subject to} \quad & g_i(x) := -x_i < 0, \quad 1 \le i \le n \\ & g_{n+1}(x) := \sum_{i=1}^{n} x_i - C \le 0 \end{aligned} \tag{29}$$

The problem (29) has a unique optimum, because its objective is represented by increasing, strictly concave, and continuously differentiable function over a convex feasibility region. The optimal

solution $x^* = \{x_1^*, ..., x_n^*\}^T$ can be found from the necessary and sufficient Karush–Kuhn–Tucker (KKT) conditions [155] given by:

$$\begin{aligned} &-\nabla f(x^*) + \sum_{i=1}^{n+1} \mu_i^* \nabla g_i(x^*) = 0 \\ &g_i(x^*) < 0, \;\; g_{n+1}(x^*) \le 0, \;\; 1 \le i \le n \\ &\mu_i^* g_i(x^*) = 0, \;\; \mu_i^* \ge 0, \;\; 1 \le i \le n+1 \end{aligned} \tag{30}$$

where: $\nabla f(x) = \left[\frac{\partial f}{\partial x_1} \;\; ... \;\; \frac{\partial f}{\partial x_n} \right]^T$, $\nabla g_i(x) = \left[\frac{\partial g_i}{\partial x_1} \;\; ... \;\; \frac{\partial g_i}{\partial x_n} \right]^T$, μ_i are non-negative Lagrange multipliers, associated with i^{th} constraint in (29); μ_i^* are the values of the Lagrange multipliers associated the optimal stationary point.

Using (29), conditions in (30) are equivalent to the following system of equations:

$$\begin{cases} -\frac{\omega_i}{x_i^*} - \mu_i^* + \mu_{n+1}^* = 0, \;\; 1 \le i \le n \\ x_i^* > 0, \;\; 1 \le i \le n \\ \sum_{i=1}^{n} x_i^* - C \le 0 \\ \mu_i^* x_i^* = 0, \;\; 1 \le i \le n \\ \mu_{n+1}^* (\sum_{i=1}^{n} x_i^* - C) = 0 \\ \mu_i^* \ge 0, \;\; 1 \le i \le n+1 \end{cases} \tag{31}$$

From the optimality conditions (31) we get:

$$x_i^* = \frac{\omega_i}{\mu_{n+1}^* - \mu_i^*}, \;\; 1 \le i \le n \tag{32}$$

Using (32) in complementary slackness conditions of (31) gives:

$$\frac{\mu_i^* \omega_i}{\mu_{n+1}^* - \mu_i^*} = 0, \;\; 1 \le i \le n, \;\; \mu_{n+1}^* \left(\sum_{i=1}^{n} \omega_i \Big/ (\mu_{n+1}^* - \mu_i^*) - C \right) = 0 \tag{33}$$

Now, if $\mu_{n+1}^{*} = 0$ then (32) implies: $x_i^{*} = \frac{\omega_i}{-\mu_i^{*}} \le 0, 1 \le i \le n.$ However, (31) means $\mu_i^{*} \ge 0$ and $\omega_i \ge 0$, $1 \le i \le n$ by construction. Hence, if $\mu_{n+1}^{*} = 0$, this contradicts (31). Thus, we conclude that $\mu_{n+1}^{*} > 0$, using which in (33) gives:

$$C = \sum_{i=1}^{n} \omega_i \Big/ (\mu_{n+1}^{*} - \mu_i^{*}) \tag{34}$$

Putting (34) into (32) gives us:

$$x_i^{*} = C\omega_i \Big/ \sum_{j=1}^{n} \omega_j , \ 1 \le i \le n \tag{35}$$

which clearly satisfies the feasibility constraints of (31) (because of the non-negativity of weight ω_i). The Lagrange multipliers corresponding to constraints in (29) are equal:

$$\mu_i^{*} = \begin{cases} 0, & 1 \le i \le n \\ \sum_{j=1}^{n} \omega_j \Big/ C \ge 0, & i = n+1 \end{cases} \tag{36}$$

After the solution of (29) has been found, we are ready to derive an appropriate load indicator, which can be used to detect the nodes with increasing (decreasing) load.

Recall, that in traditional OFC the load in different nodes is controlled based on assumption that in uncongested node the length of the buffer is non-increasing, i.e. $\gamma_i = \frac{Q_i(t+1)}{Q_i(t)}, \gamma_i \le 1$, whereas in congested node the length of the buffer experiences a multiplicative increase, $\gamma_i = \frac{Q_i(t+1)}{Q_i(t)}, \gamma_i > 1$ [65].

It is rather straightforward to verify that γ_i can only be applied for the nodes with infinite buffers. Applying the same algorithm in the nodes with finite buffer size might fail to detect the overload situation.

Indeed, if at some point the buffer of a eNB reaches its maximum value Q_{max}, and does not decrease over time, then γ_i will remain to be equal 1, indicating that the node is uncongested. Therefore, in order to

prevent the buffer overflows, we propose to modify the load control indicator γ_i used in [65] by adding the amount of data dropped to the numerator as shown in:

$$\gamma_i^{\mathrm{mod}} = \frac{Q_i(t+1) + D_i(t+1)}{Q_i(t)} \tag{37}$$

Combining (28) and (37), we obtain the following new expression for uplink and downlink modified load indicators γ_i^{ULmod} and γ_i^{DLmod} which can be used to detect the congestion in the nodes with limited buffer capacity:

$$\begin{aligned} \gamma_i^{UL\mathrm{mod}} &= \frac{\left[Q_i^{UL}(t) + L_i^{UL}(t+1) - X_i^{UL}(t+1)\right]^+}{Q_i^{UL}(t)} \\ \gamma_i^{DL\mathrm{mod}} &= \frac{\left[Q_i^{DL}(t) + L_i^{DL}(t+1) - X_i^{DL}(t+1)\right]^+}{Q_i^{DL}(t)} \end{aligned}, \quad 1 \le i \le n \tag{38}$$

Note, that in (38), the true parameters $\left\{L_i^{UL}(t+1), L_i^{DL}(t+1)\right\}_{i=1}^n$ and $\left\{X_i^{UL}(t+1), X_i^{DL}(t+1)\right\}_{i=1}^n$ are not available at the moment of allocation (i.e. at t^{th} monitoring interval). However, we can use the predicted values $\left\{\hat{L}_i^{UL}(t+1), \hat{L}_i^{DL}(t+1)\right\}_{i=1}^n$ instead of $\left\{L_i^{UL}(t+1), L_i^{DL}(t+1)\right\}_{i=1}^n$, and estimate $\left\{X_i^{UL}(t+1), X_i^{DL}(t+1)\right\}_{i=1}^n$ using well-known relation [156]:

$$X_i^{UL}(t+1) = \frac{x_i \Delta t}{SE^{UL}}, \quad X_i^{DL}(t+1) = \frac{x_i \Delta t}{SE^{DL}}, \quad 1 \le i \le n \tag{39}$$

where SE^{UL}, SE^{DL} are the spectral efficiencies of the uplink and downlink LTE channels, respectively (in bits per second (bps)/Hz).

We also consider that in LTE the bandwidth assigned to eNBs affects both the uplink and the downlink channels, and therefore both uplink and downlink load indicators[1] $\gamma_i^{UL\mathrm{mod}}$ and $\gamma_i^{DL\mathrm{mod}}$ should be taken into account for weight assignment, i.e.:

[1] In this work we assume that eNBs operate on unpaired frequency bands (i.e. uplink and downlink channels share the same frequency band). Therefore, a proper bandwidth allocation scheme should account both uplink and downlink load indicators. If the model comes on paired spectrum (uplink and downlink channels operate on different frequency

$$\omega_i = \gamma_i^{UL\mathrm{mod}} + \gamma_i^{DL\mathrm{mod}}, \quad 1 \le i \le n \tag{40}$$

Combining (38), (39) and (40) we get the expression (27) given in previous section.

4 Algorithm Performance

In this section a summary of the algorithm performance if presented. More detailed service performance evaluation can be found in the corresponding paper.

The simulation model of the network consists of seven eNBs connected to the server via NRM using 1Gbit/s data rate IP links. Each eNB serves a number of fixed user equipments (UEs), randomly positioned in the system area with 1km radius. The simulation model has been developed using the OPNET simulation and modeling package [112]. Parameters of the simulation model are listed in Table 9. The total available bandwidth in simulations is equal C = 35 MHz. The length of the monitoring time interval is set to be equal Δt = 10sec.

bands), the bandwidth would be allocated separately to the uplink (based on uplink load indicator) and downlink (based on downlink load indicator) transmissions.

Table 9. Simulation Parameters

Parameter		*Value*
PHY profile:	Operation mode	FDD
	Cyclic Prefix Type	Normal (7 Symbols per Slot)
	Carrier frequency	2GHz
	Subcarrier spacing	15kHz
	MCS	9 (16QAM with coding rate = 0.601563)
	EPC Bearer Definitions	348kbit/s (Non-GBR)
BSR parameters:	Periodic Timer	5 subframes
	Retransmission Timer	2560 subframes
PS scheme:	PF throughput	

In this work performance of the proposed PRA algorithm with modified load indicators (denoted via MLI) is compared with the performance of other commonly used resource allocation schemes: bandwidth allocation scheme with conventional load indicators (denoted via LI) and fixed bandwidth allocation scheme (denoted via FA). Description of the simulated schemes is provided in Table 10. Simulation parameters of LI and FA schemes are the same as in MLI.

Table 10. Simulated Bandwidth Allocation Schemes

Scheme	*Optimization Criteria*	*Weight Generation*
FA	Equal Fixed Bandwidth Allocation	$\omega_i = 1,\ 1 \le i \le n$
MLCI	Weighted Proportional Fairness	$\omega_i = \gamma_i^{UL\ mod} + \gamma_i^{DL\ mod},\ 1 \le i \le n$
LCI	Weighted Proportional Fairness	$\omega_i = \gamma_i^{UL} + \gamma_i^{DL},\ 1 \le i \le n$

Two scenarios have been used to observe the algorithm performance. In the first homogeneous scenario, referred as “voice”, each eNB served only VoIP users. In the second heterogeneous scenario, referred as “mix”, the traffic of eNBs is represented by a “mixed” traffic comprising VoIP, video and data users. The overall number of users of each type is in proportion to 2:2:3 for VoIP, video, and data, respectively. The VoIP traffic is generated by using the G.723.1 (12.2 Kbps) codec with a voice payload size 40 bytes and a voice payload interval 30 ms. Each VoIP user might be either in active (talk-spurts period) or inactive (silent period) state. The durations of the talk-spurts and silent periods are exponentially distributed with 0.65s and 0.352s means, respectively. Video services are simulated using a high resolution video model with a constant frame size equal 6250 bytes and exponentially distributed frame inter-arrival intervals (with mean equal 0.5s). Data users in simulations are HTTP1.1 users generating pages or images with exponential page inter-arrival intervals (mean equal 60sec). It is assumed that one page consists of one object, whereas one image consists of five objects. The object size is constant and equal 1000 bytes.

Figures presented below illustrate performance of different bandwidth allocation schemes for different traffic mixes with different traffic loads in the networks. Figure 37 shows the packet transmission and queueing delays calculated in the uplink and downlink directions in scenarios with voice users. Figure 38 shows the packet transmission and queueing delays in the uplink and downlink directions in scenarios with voice, video and data applications. From these graphs it follows that both MLI and LI reduce the transmission and queueing delays by a significantly margin for voice and mixed traffic scenarios. However, performance of MLI is better than performance of LI, especially in case of heterogeneous user traffic.

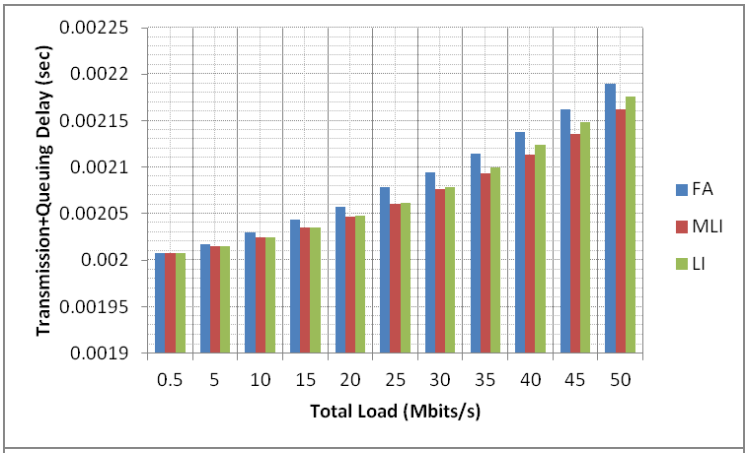


Fig. 37. Total MAC-layer packet delay in scenarios with voice traffic

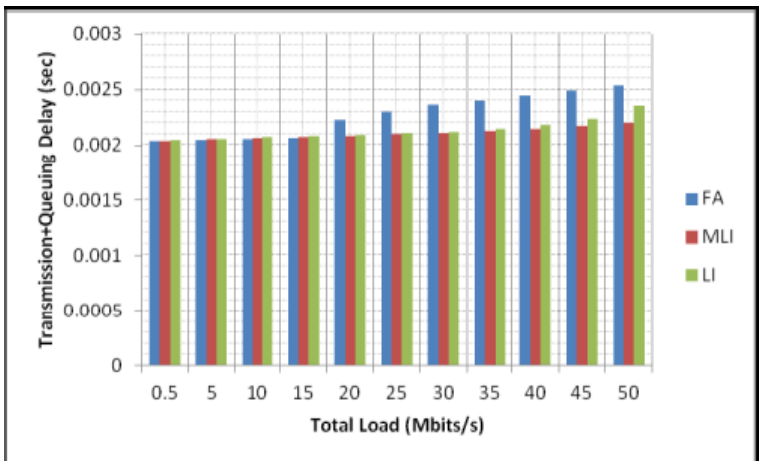


Fig. 38. Total MAC-layer packet delay in scenarios with mixed traffic

Significant performance improvement of MLI is achieved because it reduces not only the average packet queuing and transmission delays, but also the average delay due to uplink packet scheduling. Figures 39, 40 show the uplink packet scheduling delay with MLI, LI and FA in homogeneous and heterogeneous scenarios, respectively.

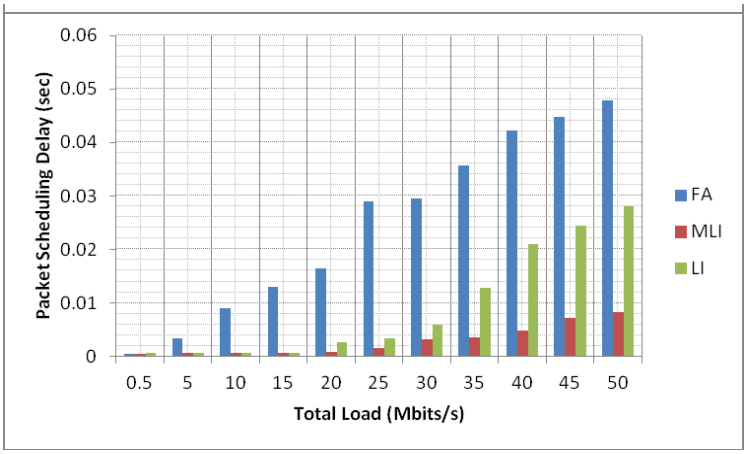


Fig. 39. Delay due to uplink packet scheduling in scenarios with voice traffic

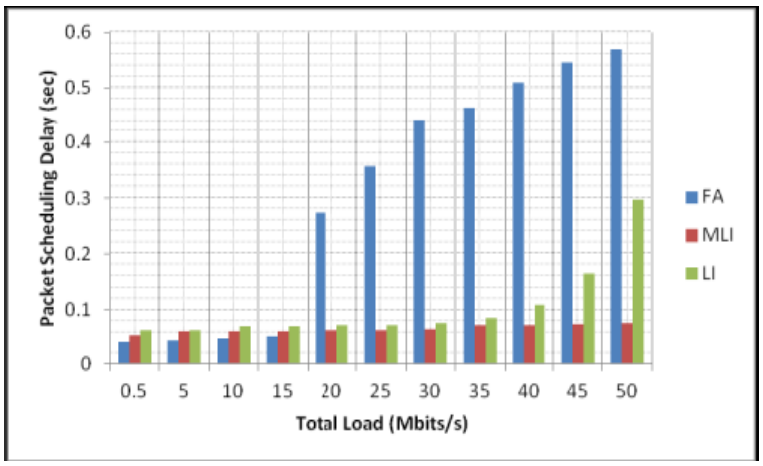


Fig. 40. Delay due to uplink packet scheduling in scenarios with mixed traffic

Delay due to uplink packet scheduling reduces with MLI because of the following reasons. In LTE, the network resources are allocated to users for uplink and downlink data transmissions in terms of resource blocks (RBs). Resource allocation (scheduling) is usually performed periodically within a fixed time interval, called scheduling period Δt_{sc} with minimal duration 1 transmission time interval (TTI) or 1 ms. The scheduling is done by the packet scheduler in the eNB both for uplink and downlink transmissions. The scheduling decisions are carried by the physical uplink control channel (PUCCH) and physical downlink

control channel (PDCCH) in uplink and downlink directions, respectively [94, 157, 158].

According to LTE specifications, PDCCH occupies first 1, 2, or 3 OFDM symbols (the number of OFDM symbols is indicated by control format indicator (CFI)) in a time slot extending over the entire system bandwidth. PDCCH is constructed from Control Channel Elements (CCEs). The number of CCEs indicates the capacity of PDCCH, i.e. the maximal number of users that can be scheduled during the considered TTI. For fixed CFI, the number of CCEs depends on the channel bandwidth (the number of CCEs for 5, 10 and 20 MHz bandwidth is given in Table 11) [159]. Based on these considerations, the number of users that can be scheduling during a considered TTI is equal to the number of CCEs.

Table 11. The Number of CCEs for Different Bandwidth [159]

The Number of CCEs per time slot, N_{CCE}	*Bandwidth (MHz)*				
	1.4	*3*	*5*	*10*	*20*
CFI = 1 OFDM symbol per time slot	1	2	3	8	17
CFI =2 OFDM symbols per time slot	2	7	12	25	50
CFI =3 OFDM symbols per time slot	4	12	20	41	84

If the traffic intensity of eNB is low, than the PUCCH/PDCCH are under-utilized; if the traffic intensity of eNB is high, than the users will be delayed not only because of the queuing, but also due to absence of free control channels. On the contrary, in MLI the number of users that can be scheduled during each TTI depends on the bandwidth assigned to eNB (N_{PS} = 2, 7, 12, 25 or 36 users for 1.4, 3, 5, 10 and 20 MHz bandwidth, respectively). Thus, if the traffic intensity in eNB is low, and eNB is assigned lower bandwidth (1.4 or 3 MHz), the control channel resources are not waisted. On the other hand, if the traffic intensity is high, than eNB is assigned higher bandwidth (10 or 20 MHz), and therefore the number of users that can be scheduling during

each TTI increases, which will decrease the delay due to uplink packet scheduling.

CHAPTER 5: Dynamic Resource Allocation in a LTE/WLAN Heterogeneous Network

In this chapter a resource allocation technique for a combined LTE/WLAN CRN in Scenario 1 is presented. Description of the network deployment scenario has already been provided in Overview of this thesis. Here we focus on specific challenges of resource allocation in the complex networks comprising more than one RAT, propose the algorithm for spectrum access in combined LTE/WLAN architecture, and evaluate its performance based on results of simulations in OPNET environment [112]. The corresponding paper has been published in Proceeding of IEEE International Congress on Ultra-Modern Telecommunications and Control Systems (ICUMT), 2012.

1 Introduction

With widespread use of wireless networks and the emergence of multiple deployed wireless standards the wireless network design paradigm is changing rapidly [1, 101]. In the future, the wireless services will be provided using heterogeneous network comprising multiple RATs rather than using a single standard network [102, 103]. Also, the emergence of software defined radio (SDR) will allow the users to connect to any RAT based on the offered QoS and capacity levels [102].

Radio resource management in heterogeneous network is a complex task due to diversified requirements of its member networks, including different physical layer characteristics, channel access mode, MAC-layer parameters, etc. For instance, IEEE 802.11 WLAN network uses contention based random multiple access technique for channel access over the wireless medium. This technique is usually characterized by numerous collisions, which can reduce the achievable throughput of WLAN users [69]. In LTE network the (potential) contention is resolved by using a Random Access Contention Resolution and Scheduling Request (SR) procedure. Hence, the collision probability in a LTE network is close to zero, thus not affecting the service rates of LTE nodes [70]. As a result, the throughput and QoS in LTE eNBs and WLAN APs operating on the same bandwidth will be different.

Initial approach for resource allocation in combined LTE/WLAN CRN is very similar to the one deployed in LTE-based CRN. Each AP/eNB is assigned with appropriate bandwidth proportional to the

value of its load control (LC) indicator, which measures the degree of load variation in service node. In this way a larger bandwidth is assigned to service nodes with increasing load, and smaller – to the nodes with decreasing load. The difference between the algorithm used for resource allocation in combined network and the algorithm used in LTE network is in the way the LC indicators are obtained. To be able to account for different spectral efficiencies, and channel access techniques deployed in LTE eNBs and WLAN APs, the spectrum efficiency and collision ratio metrics are measured discontinuously in each service node, and further used to calculate the values of LC indicators together with predicted traffic load in APs/eNBs. Further in this Chapter we briefly outline the algorithm for resource allocation, and show its performance based on simulation model developed using OPNET platform [112].

2 Resource Allocation Algorithm

In this section the proposed algorithm for resource allocation in combined LTE/WLAN CRN is briefly summarized. More detailed description of the algorithm can be found in the corresponding paper.

Considered cognitive network architecture is illustrated on Figure 41. It consists of *n* service nodes (LTE eNBs and WLAN APs) connected to the LTE System Architecture Evolution Gateway (SAE GW). The communication between each eNB/AP, SAE GW and external networks is realized using high-speed IP links. In LTE, the communication interface been LTE service nodes (eNBs) and SAE GW is called S1 interface, the interface between non-LTE service nodes (APs) is called S2 interface, and the interface between SAE GW and external networks is called SGi [129, 130].

In the model, resource allocation and control is carried by the Network Resource Manager (NRM) located in SAE GW. The network resources are represented by the total available bandwidth *C* and shared among service nodes according to a certain resource allocation policy.

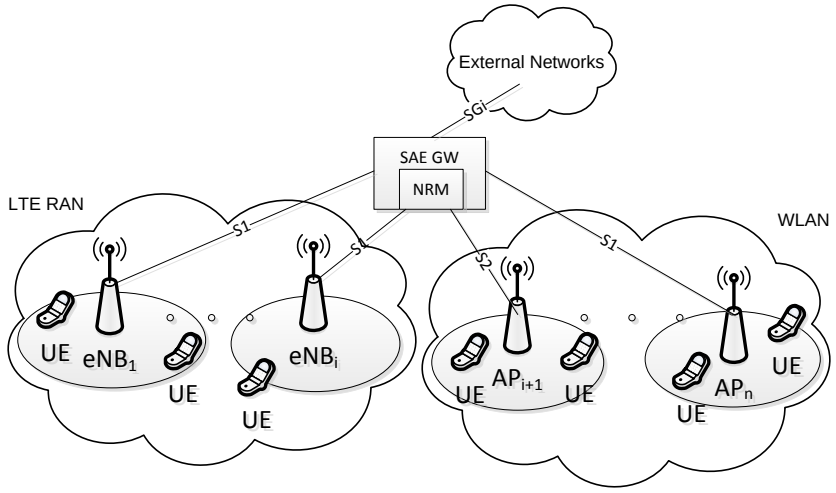


Fig. 41. Combined LTE/WLAN cognitive network architecture

The assumptions of the model are summarized below:

1. uplink (UL) and downlink (DL) buffers of service nodes have known finite capacities denoted Q^{UL}_{max} and Q^{DL}_{max};
2. parameter monitoring, bandwidth allocation and traffic prediction is performed within discrete fixed time intervals $\{[t\Delta t, (t+1)\Delta t]\}$, with t denoting the index of time interval.
3. the buffer length, the amount of data arrived to the buffers of service nodes and the number of collisions occurred during each time interval are known;
4. data collected and predicted separately for each UL and DL direction.

The following notations are used in this Chapter:

x_i – bandwidth (in Hz) allocated to service node i at the $(t+1)^{th}$ time interval.

$Q_i^{UL}(t)$, $Q_i^{DL}(t)$, $Q_i(t)$ –length of the buffers (in bits) of service node i at t^{th} time interval on the uplink, the downlink and the unspecified (general) channels, respectively;

$X_i^{UL}(t)$, $X_i^{DL}(t)$, $X_i(t)$ – the amount of data served (in bits) at the buffer of service node i at t^{th} time interval on the uplink, the downlink and the unspecified (general) channels, respectively;

$L_i^{UL}(t)$, $L_i^{DL}(t)$, $L_i(t)$ – the amount of data arrived (in bits) to the buffer of service node i at t^{th} time interval on the uplink, the downlink and the unspecified (general) channels, respectively;

$D_i^{UL}(t)$, $D_i^{DL}(t)$, $D_i(t)$ – the amount of data dropped (in bits) from the buffer of service node i at t^{th} time interval on the uplink, the downlink and the unspecified (general) channels, respectively.

$Nc_i^{UL}(t)$, $Nc_i^{DL}(t)$, $Nc_i(t)$ – the number of collisions occurred at the service node i at t^{th} time interval on the uplink, the downlink and the unspecified (general) channels, respectively.

The algorithm proposed for resource allocation in LTE/WLAN CRN is very similar to the one proposed for LTE network. To deal with heterogeneous network applications, we propose to apply the concept used in optimal flow and congestion control (OFC) [65], and assign the network resources to the service nodes based on the values of load control (LC) indicators. However, in combined network architecture it is necessary to take into account that different wireless networking standards have different spectral efficiency, which means that for the same allocated bandwidth the throughput and the QoS of the users of LTE eNBs and WLAN APs will be different. A contention based random multiple access technique is deployed in WLANs that is usually characterized by numerous collisions, which can reduce the achievable throughput of WLAN users [69]. In LTE, the potential contention is resolved by using a Random Access Contention Resolution and Scheduling Request (SR) procedure. Therefore the collision probability in a LTE network is close to zero, thus not affecting the service rates of LTE nodes [70].

To be able to account for different spectral efficiencies, and channel access techniques in a LTE/WLAN network, it is necessary to include the spectral efficiency and the collision ratio metrics in bandwidth allocation algorithm. In our model these metrics are used to calculate the amount of data served during each time interval Δt as shown below

$$X_{iLTE}(t+1) = \frac{x_i \Delta t}{SE_{LTE}}, \; X_{iWLAN}(t+1) = (1 - Pc_i)\frac{x_i \Delta t}{SE_{WLAN}} \tag{41}$$

where SE_{WLAN}, SE_{LTE} are the spectral efficiency of IEEE802.11g and LTE standards in bits/s/Hz, respectively (the SE value depend on MCS used in respective service node); Pc_i is the collision ratio measured in

service node *i* which is the ratio of number of collisions to data volume at each service node given by

$$Pc_i^{UL} = \frac{Nc_i^{UL}(t)}{L_i^{UL}(t)}, Pc_i^{DL} = \frac{Nc_i^{DL}(t)}{L_i^{DL}(t)} \tag{42}$$

The bandwidth allocation algorithm consists of a number of steps described below. At time interval *t*:

1. **Input:** The service nodes monitor $\{Q_i^{UL}(t), Q_i^{DL}(t)\}_{i=1}^n$, $\{L_i^{UL}(t), L_i^{DL}(t)\}_{i=1}^n$, $\{Nc_i^{UL}(t), Nc_i^{DL}(t)\}_{i=1}^n$ and send this information to the NRM using IP links.
2. **Prediction:** Using the input information collected up to t^{th} monitoring interval, NRM computes the predictions $\{\hat{L}_i^{UL}(t+1), \hat{L}_i^{DL}(t+1)\}_{i=1}^n$ using PLR parameter estimation applied with AR(1) time-series model (detailed description of PLR technique and AR model has already been provided in Chapter 1).
3. **Load Control:** Based on the values $\{\hat{L}_i^{UL}(t+1), \hat{L}_i^{DL}(t+1)\}_{i=1}^n$, $\{Nc_i^{UL}(t), Nc_i^{DL}(t)\}_{i=1}^n$ and $\{Q_i^{UL}(t), Q_i^{DL}(t)\}_{i=1}^n$, NRM assigns the uplink and downlink collision-based load control (CLC) indicators for WLAN APs:

$$\gamma_i^{UL} = \frac{\left[Q_i^{UL}(t) + \hat{L}_i^{UL}(t+1) - (1 - \frac{Nc_i^{UL}(t)}{L_i^{UL}(t)}) \frac{x_i \Delta t}{SE_{WLAN}^{UL}} \right]^+}{Q_i^{UL}(t)}$$

$$\gamma_i^{DL} = \frac{\left[Q_i^{DL}(t) + \hat{L}_i^{DL}(t+1) - (1 - \frac{Nc_i^{DL}(t)}{L_i^{DL}(t)}) \frac{x_i \Delta t}{SE_{WLAN}^{DL}} \right]^+}{Q_i^{DL}(t)} \tag{43}$$

and LTE eNBs:

$$\gamma_i^{UL} = \frac{\left[Q_i^{UL}(t) + \hat{L}_i^{UL}(t+1) - \frac{x_i \Delta t}{SE_{LTE}^{UL}} \right]^+}{Q_i^{UL}(t)}$$
$$\gamma_i^{DL} = \frac{\left[Q_i^{DL}(t) + \hat{L}_i^{DL}(t+1) - \frac{x_i \Delta t}{SE_{LTE}^{DL}} \right]^+}{Q_i^{DL}(t)} \quad (44)$$

where $[x]^+$ denotes $max(0, x)$. These weights are used later in weighted proportional fairness problem. The justifications behind using the expression (43), (44) have already been provided in Chapter 4.

4. **Optimization:** Using generated LC indicators, NRM generates the weights $\{\omega_i = \gamma_i^{UL} + \gamma_i^{DL}\}_{i=1}^n$, and calculates the bandwidth $\{x_i\}_{i=1}^n$ that will be assigned to each service node in the next $(t+1)^{th}$ monitoring interval based on the weighted proportional fairness criterion, and transmit these values to corresponding eNBs via IP links.
5. **Output:** Received bandwidth allocation values $\{x_i\}_{i=1}^n$ are assigned to the physical interfaces of the corresponding service nodes at the beginning of the $(t + 1)^{th}$ monitoring interval.

3 Algorithm Performance

In this section a summary of the proposed algorithm performance is provided. More detailed performance analysis can be found in the corresponding paper.

An OPNET based simulation model has been developed to observe the performance of the proposed resource allocation algorithm using collision-based load control (LC) indicators. It comprises seven service nodes: four LTE eNBs and three Wi-Fi (IEEE802.11g) APs, communicating with SAE GW using 1Gbit/s IP links. Each AP/eNodeB serves a number of fixed user terminals through the radio interface randomly positioned in a total coverage area with a 1000m radius.

In this work performance of the proposed bandwidth allocation scheme (denoted CLC) is compared with the performance of two other bandwidth allocation schemes. First scheme (denoted LC) is similar to the one used for bandwidth allocation in LTE network, i.e. when the collision ratio metric is not considered in expression for load control indicators. Another scheme (denoted FA) is fixed bandwidth allocation scheme in which all eNBs are assigned fixed 5 MHz bandwidth, all APs are operate with 11 Mbits/s data rate (corresponding to 7MHz bandwidth) [69, 70].

In all simulations total available bandwidth is equal C = 40 MHz, the time interval for resource allocation is equal Δt = 10s. Both LTE and WLAN service nodes can operate only with discrete values of transmission rates as listed in Table 12. Most important simulation parameters of LTE and WLAN air interfaces are provided in Table 13. Performance of the bandwidth allocation schemes were analyzed for high priority delay-sensitive VoIP services. VoIP traffic was simulated by using G.723.1 (12.2 Kbps) codec with 40 bytes codec sampling size, and 30 ms codec sample interval. Discontinuous Transmission, Voice Activity Detection and Comfort Noise Generation are also applied.

Table 12. Transmission Parameters [69, 70]

Interface	*The set of scalable bandwidth/service rate values*		
	Scalable Parameter	*Range*	*Buffer Capacity in kbits/s*
IEEE 802.11g	Service Rate in Mbits/s	1, 2, 5.5, 11, 6, 9, 12, 18, 24, 36, 48, 55	1024
LTE	Bandwidth in MHz	1.4, 3, 5, 10, 15, 20	1620, 3250, 6500,14000, 19500, 26000

Table 13. Key Simulation Parameters

Interface	*Parameter*		*Value*
LTE	PHY Profile:	Operation mode	FDD
		Cyclic Prefix Type	Normal (7 Symbols per Slot)
		Carrier frequency	2GHz
		Subcarrier spacing	15kHz
		MCS	16QAM with coding rate 0.601563
		EPC Bearer Definitions	348kbit/s (Non-GBR)
	BSR Parameters:	Periodic Timer	5 subframes
		Retransmission Timer	2560 subframes
	PS Scheme	PF Throughput	
IEEE 802.11g	PHY Profile:	Multiple Access Method	CSMA/CA
		Carrier frequency	2GHz
		Subcarrier spacing	312.5kHz
		MCS	BPSK, QPSK, 6QAM, 64QAM
	Retransmission Parameters:	Long Retry Limit	4
		Short Retry Limit	7

Figures below demonstrate the application-layer performance (packet end-to-end delay and packet loss) of the considered bandwidth allocation schemes for VoIP users. These figures show that both LC and CLC reduce the delay in LTE and WLAN interfaces. However, CLC show better performance than LC in terms of the packet delay,

especially for WLAN network under high load conditions (0.5% packet loss in LC versa 1.5% and 3% loss in CLC and FA, respectively).

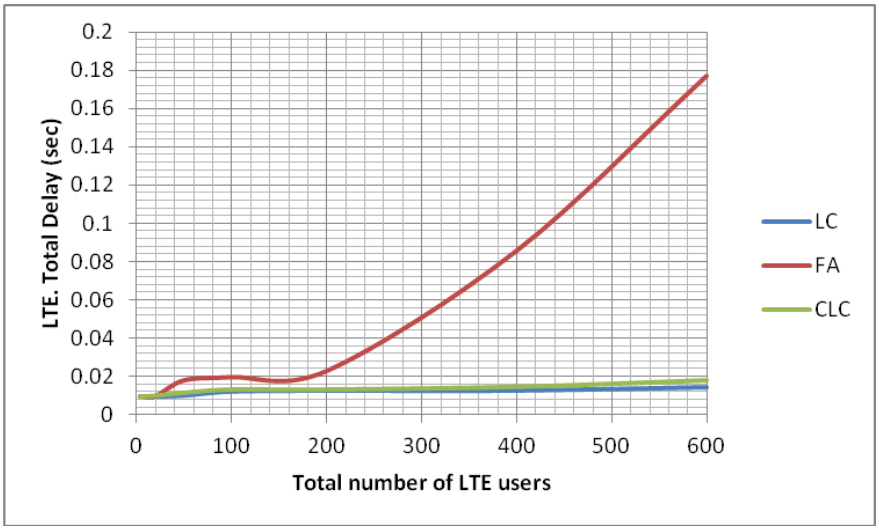


Fig. 42. Packet end-to-end delay in LTE network

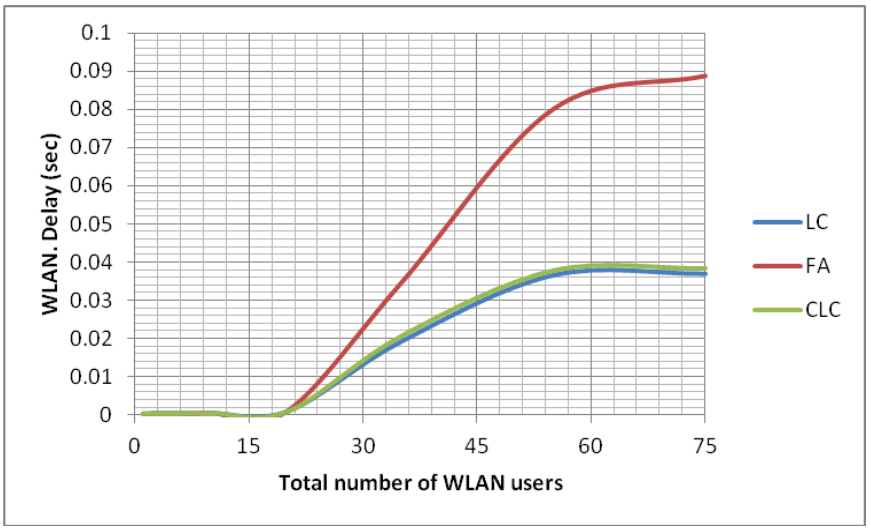


Fig. 43. Packet end-to-end delay in WLAN network

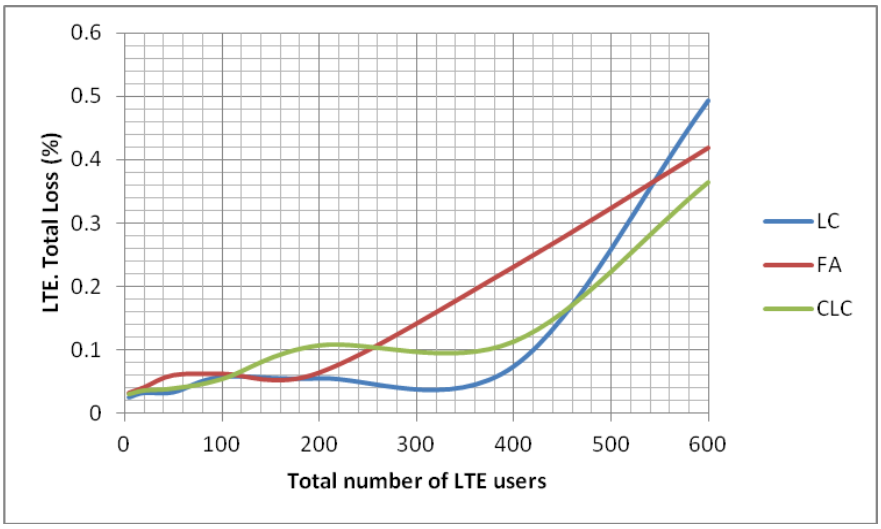


Fig. 44. Packet loss in LTE network

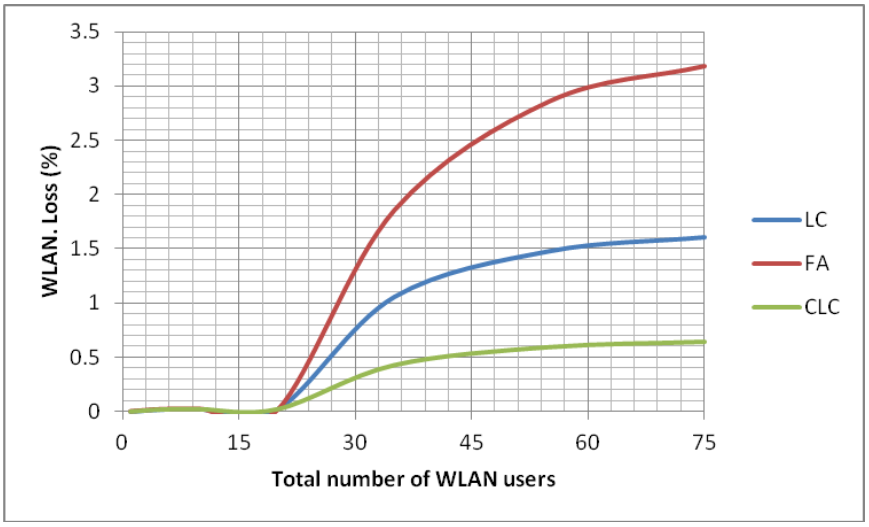


Fig. 45. Packet loss in WLAN network

Such results can be explained using the Figure 46 showing the collision ratio in the WLAN network. Collisions in WLAN reduce the service rate of APs resulting in higher packet loss of WLAN users. Since the LC algorithm does not account the collision ratio for the bandwidth allocation, it performs purely in terms of the WLAN loss ratio. On the other hand, the CLC algorithm outperforms the LC

algorithm by allocating more bandwidth to WLAN users with non-zero collision ratio (ref. to (41)).

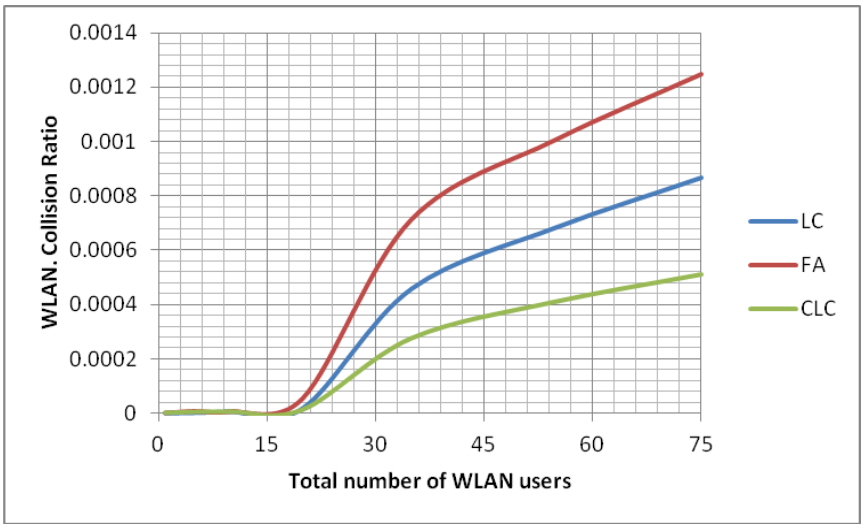


Fig. 46. Packet collision ratio in WLAN network

CHAPTER 6: Resource Allocation Algorithm in Cognitive LTE Network with Heterogeneous User Traffic

In this chapter a resource allocation technique for a cognitive LTE in Scenario 1 is presented. Description of the network deployment scenario has already been provided in Overview of this thesis. Here we present the proposed approach for resource allocation, derive resource allocation algorithm, and present the results of algorithm performance based on simulation model developed in OPNET environment [112]. The corresponding paper is published in proceedings of IEEE GLOBal COMmunications (GLOBECOM) Conference in December 2013.

1 Introduction

A significant progress has been recently made in resource allocation for IEEE802.22 CRN architecture. Nevertheless, many challenges still remain [54]. For instance, most research has focused on individual techniques for identifying and reducing the interference (by controlling transmit power, carrier sense, or scheduling) for the users of CRN (see, for instance, [55 - 58]). In general, however, the system performance depends on many external factors, including user behavior, traffic load, channel quality, etc. [54].

Some theoretical models of the user behavior and traffic load in CR network have been proposed in [59 - 62], but the assumptions made in theoretical research often fail under realistic operating conditions due to the fact that a system may operate in diverse environments (e.g., in different types of city, rural, campus, and indoor deployments) [54]. It is therefore very difficult to obtain some general theoretical model which can be applied for different network deployment scenarios. More rational would be to: i) identify most critical parameters affecting the system performance; ii) investigate all available tools to analyze the service quality in the network based on the certain parametric observations collected in different locations at different time, and iii) apply these tools in spectrum allocation algorithm in order to improve the service performance of CR system.

Based on these considerations, we propose an alternative way for resource allocation in the standard IEEE 802.22 CRN architecture where each SP is represented by its LTE evolved NodeBs (eNBs). Using the fact that the overall system performance depends on many factors (including user behavior, traffic load, channel quality, etc.), which are difficult to model analytically, we propose to apply some form of reinforcement learning [160, 161] in spectrum allocation algorithm and propose to make a short-term resource allocation based on the long-term traffic predictions. In this way we enforce the network to learn from the environment, and adapt according to the current network conditions.

The rest of the Chapter is organized as follows. In section 2 we describe the main idea behind the proposed resource allocation approach and how it can improve the network performance in heterogeneous network environment. In section 3 we formulate the optimization problem, specify the solution techniques and summarize the proposed algorithm for resource allocation in CR system. In section 4 we give the example of the algorithm implementation in LTE-based CRN and provide the detailed performance analysis of the algorithm.

2 Proposed Approach for Resource Allocation

According to [54], the individual spectrum bands are used in a fairly homogeneous fashion. In contrast to them, the usage pattern in CRN is in general heterogeneous. Consider, for instance, the intra-campus network where some of the eNBs are located in academic schools, other eNBs serve the staff buildings and the school libraries, whereas the rest provide the wireless access in residential areas. It is reasonable to expect that the usage pattern in eNBs will be very different. For instance, the school eNBs might experience heavy demand during the lecture hours and will not be used the rest of the time, the eNBs located in the offices and libraries will be loaded during the day-time and empty during the night, whereas the eNBs in residential buildings will be mostly used in the evening and night time. The web applications and traffic patterns of the individual users of these eNBs might also vary: the students and staff in the offices and the libraries might access the e-mail and perform the web-search, whereas in residential buildings the

VoIP, video and on-line games might be used more frequently. Thus, to build a practically sustainable system it is important to keep in mind that different eNBs might operate in different conditions, i.e. the network usage is location and time dependent and the service demand in the network is heterogeneous.

Most of the resource allocation strategies for CRN have been deployed for homogeneous scenarios and not very efficient in case of heterogeneous network applications [59 - 62, 71 - 74]. This is due to the fact that all users in the network are characterized by similar utility functions. Existing approaches to deal with the problem of resource allocation in the network with heterogeneous user demands (for instance, [76 - 78]) are either very complex (such as [76]) or lead to rather unfair resource allocation in the sense that applications with lower demand are allocated a higher transmission rate than applications with higher demand ([77, 78]).

In this work we suggest to deploy an alternative approach for resource allocation in cognitive LTE network, and propose to make a short-term resource allocation based on the long-term traffic prediction (here and after we call this approach a "multi-step allocation"). The idea behind this approach can be formulated as follows. Consider a system illustrated on Figure 47. The system operates on a slotted time basis, e.g. the time axis of the system is partitioned into discrete mutually disjoint intervals (time slots) $\{[t\Delta, (t+1)\Delta]\}$, $t = 0, 1, 2, \ldots$, of the length Δ with t denoting the integer valued index of a time slot.

In the system n sources, denoted via $s_1, s_2, \ldots, s_n$, randomly generate the packets and send them to the dedicated queues, denoted via $q_1, q_2, \ldots, q_n$, respectively. Let $A_i(t)$ be the (random) number of packets arrived to the queue q_i at time slot t. We assume that at any time slot t the random parameter $A_i(t)$ can be observed and collected from any queue in the system.

The queues receive the packets, and serve them with some deterministic packet service rate. Let $X_i(t)$ be the (deterministic) number of packets served by the queue q_i at time slot t. We assume that the parameter $X_i(t)$ is adjustable (i.e. can be changed), positive and at any time slot t it can be observed and collected from any queue in the system. Clearly, the size of the queue q_i at time slot t, denoted via $Q_i(t)$, depends on the number of packets arrived and served by the queues,

$A_i(t)$ and $X_i(t)$. The relation between $Q_i(t)$, $A_i(t)$ and $X_i(t)$ is established in the well-known Lindley's equation [23] given by:

$$Q_i(t+k) = \left\lceil Q_i(t+k-1) + A_i(t+k) - X_i(t+k) \right\rceil^+, \quad i = 1,..., n \tag{45}$$

where $[x]^+$ denotes $max(0, x)$.

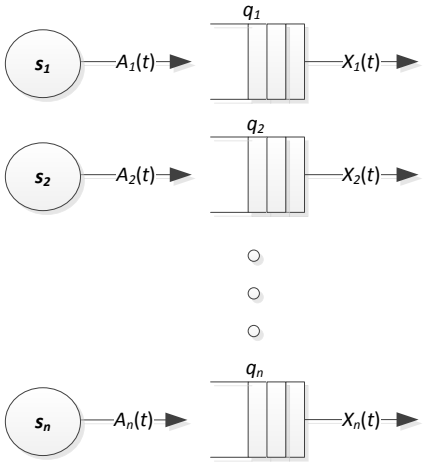


Fig. 47. The discrete-time system with n queues

We assume that at any time slot t the parameter $Q_i(t)$ can be observed and collected from any queue in the system.

Suppose, that at any time slot t we can make the predictions $\hat{A}_1(t+k)$, …, $\hat{A}_n(t+k)$ and adjust the service rates $X_1(t+k)$, …, $X_n(t+k)$ for a short-term (k = 1) and a long-term (k = 2, 3, …) period in the future. Then the size of the queues can be estimated recursively using (1).

An illustrative example of such estimation for the system comprising two sources, with C = 6, $Q_1(t) = Q_2(t) = 0$ and $X_1(t+k) = X_2(t+k) = 3$ for k = 1, 2, 3, … 10 is shown on Figures 48 - 51.

The sources generate the packets in rather different manner (Figure 48 shows the packet arrivals from the sources during the observation period t, …, t+10). The average number of packets generated by the first source during the observation period is much smaller than the number of packets generated by the second source (1 packet per time for the first source slot versa 4.1 packets per time slot for the second source). However, the peak number of packets generated by the first source is greater than the peak number of packets generated by the

second source (10 packets per time for the first source slot versa 5 packets per time slot for the second source).

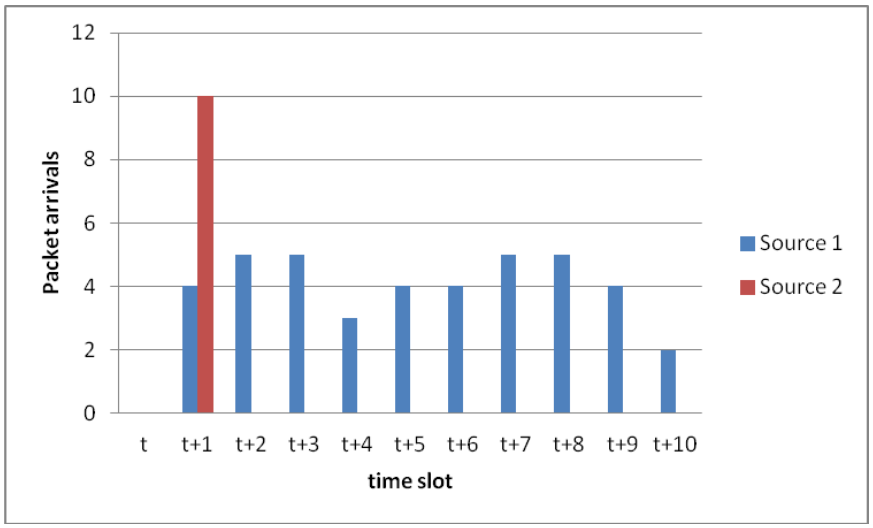


Fig. 48. Packet arrivals generates by source 1 and source 2 during 10 time slots

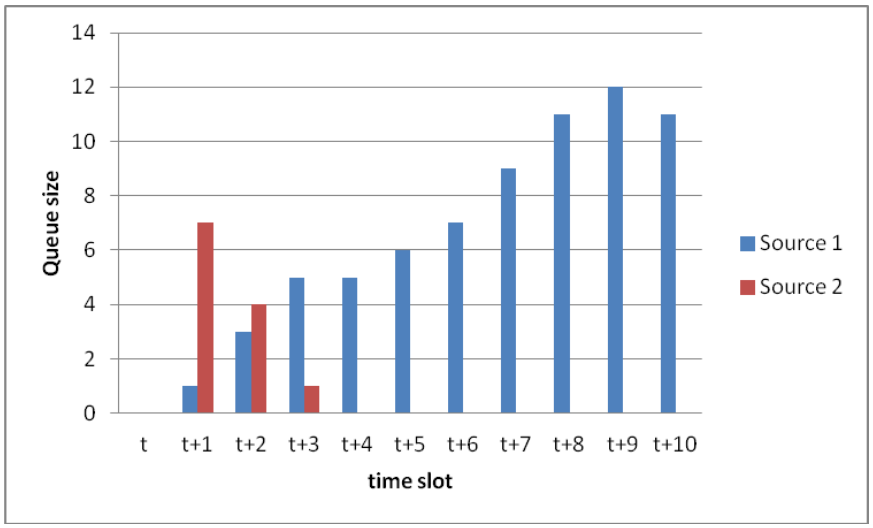


Fig. 49. Queue size with constant service rate (3 packet/slot) of the queue 1 and queue 2

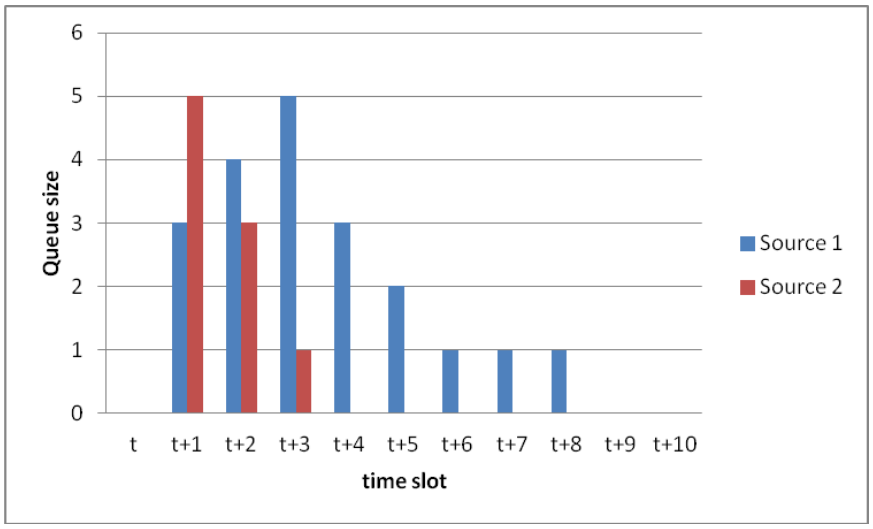


Fig. 50. Queue size with "single-step allocation"

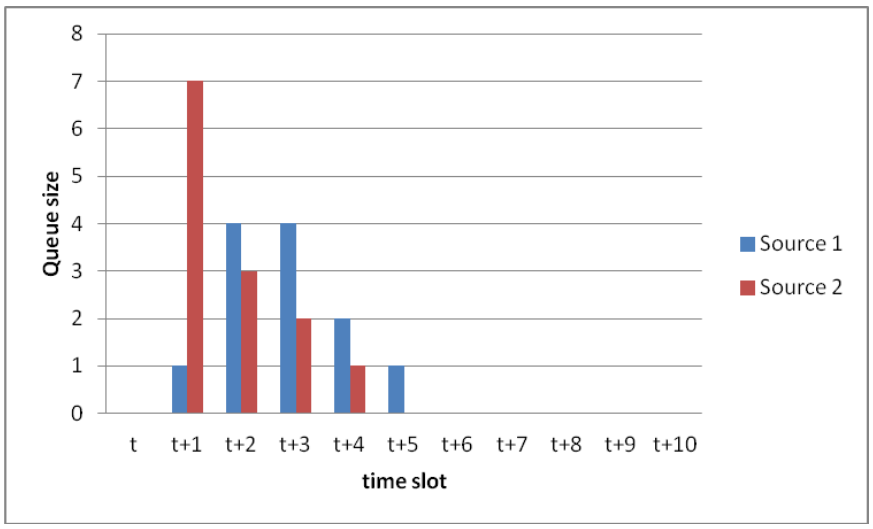


Fig. 51. Queue size with "multi-step allocation"

Consequently, in the short-term (k = 1) future the size of the queue q_1 will be much bigger than the queue q_2, because the batch of packets generated by the first source at time slot t is much longer that the batch of packets generated by the second source (Figure 49 shows the size of the queues (in packets) during this time interval). However, in the long-term future all packets generated by the first source will be served

within 2 time slots, whereas the size of the queue q_2 will increase, because the second source will keep generating the packets all the time within the considered period.

Now assume that at any time slot t the number of packet served by the queues within the next time slot, $X_1(t+1)$ and $X_2(t+1)$ can be adjusted based on some optimality criteria given that $X_1(t+1) + X_2(t+1) \leq C$. Suppose, we decide to make a short-term allocation to minimize the aggregated size of the queues in the short-term future, i.e. at any time slot t during the observation period we allocate $X_1(t+1)$, $X_2(t+1)$ to minimize $Q_1(t+1) + Q_2(t+1)$. Here and after in the paper we call this approach for resource allocation a "single-step allocation" to emphasis the fact that the resources are allocated based on the short-term ($k = 1$) prediction. In this case the queue of the first source will be cleared within 3 time slots, whereas the queue of the second source will be cleared within 8 time slots (the size of the queues in the system with single-step resource allocation is shown on Figure 50). Apparently, this approach for resource allocation is rather unfair because the first source with lower average demand is allocated higher service rate than the second source with higher demand.

On the other hand, we can decide to make a short-term allocation to minimize the aggregated size of the queues in the long-term future, i.e. at any time slot t during the observation period we allocate $X_1(t+1)$, $X_2(t+1)$ to minimize $Q_1(t+k) + Q_2(t+k)$ for $k = 1, 2, 3, \ldots 10$. We call this approach for resource allocation a "multi-step allocation" to emphasis that the resources are allocated based on the long-term ($k > 1$) prediction. In this case the queues of the second source will be cleared more quickly, and as a result, the delay experienced by the second source will be less than that in single-step allocation (the size of the queues in the system with multi-step resource allocation is shown on Figure 51). Besides, by applying the multi-step allocation we also decrease the average size of the queues during the observation period (2.5 packets versa 2.9 packets in single-step allocation).

This example shows that in heterogeneous network environment the traditional approach for resource allocation when a short-term resource allocation is made based on the short-term traffic prediction (see, for instance [63, 64]) leads to rather unfair resource allocation when the applications with lower average demand are allocated higher service rate than the applications with higher demand. On the contrary, by

applying a multi-step allocation when the resources are allocated based on the long-term traffic prediction we increase the fairness of resource allocation and decrease the average size of the queues (and therefore the average packet delay) in the system.

3 Resource Allocation Algorithm

3.1 Optimization Problem

Consider the standard IEEE 802.22 CR network architecture where n LTE eNBs numbered eNB_1, …, eNB_n share the total available bandwidth b using the SM according to some predefined spectrum usage policy. The system operates on a discrete-time basis, e.g. the time in the system is partitioned into discrete mutually disjoint intervals, called time slots, with t denoting the integer valued index of a time slot.

The system serves a number of wireless users connecting to the eNBs in their service area (cell) and generating a random traffic expressed in bits per time slot (bps). We use the notation $A_i(t)$ to denote the aggregated user traffic in eNB_i at time slot t.

The service rate of the eNB depends on the portion of bandwidth assigned to the eNB and the spectrum efficiency of the wireless channel between the user and the eNB. The relation between the service rate, the bandwidth and the spectrum efficiency of the wireless channel is given by well-known expression [156]:

$$b_i(t) = SE_i \cdot X_i(t), \quad i = 1,..., n \tag{46}$$

where $X_i(t)$ is the service rate (in bits per time slot or bps) of eNB_i at time slot t; $b_i(t)$ is the bandwidth (in Hz) assigned to eNB_i at time slot t; SE_i is the spectrum efficiency (in bits per time slot per Hz or bps/Hz) of the wireless channel, which depends on the physical channel characteristics (such as modulation and coding rate) and the channel quality (signal-to-noise ratio).

Clearly, the system described above can be described using the model shown on Figure 47. Here each eNB will be represented by a single infinite queue, whereas all users connected to the eNB will form the source served by this queue.

We set the objective to allocate the service rates of eNBs in such way that the total system bandwidth will not exceed the predefined limit *b* based on some optimality criterion denoted via *f*. The appropriate choice of the criterion *f* is apparently one of the most critical factors affecting the performance of resource allocation for practical network implementations [75].

For most of the network applications (such as voice, video, data), the user-perceived QoS is determined in terms of the packet end-to-end delay and packet loss experienced by the user. For instance, for VoIP applications the satisfactory service is achieved when packet end-to-end delay does not exceed 300ms with packet loss less than 5%; for videoconference users the QoS requirements are the same as for VoIP applications; for streaming video the packet end-to-end delay should not exceed 4-5 sec with packet loss less than the QoS requirements for video applications are the satisfactory service network performance is achieved when packet end-to-end delay does not exceed 200ms with packet loss less than 1% [79]. Therefore, it would be reasonable to represent the optimization objective in terms of the packet delay or packet loss. However, in general it is very difficult to estimate the values of the packet delay or loss accurately, because they depend on many network parameters some of which might not be possible to observe directly. More convenient would be to use the queue size as an optimization objective because: 1) it can be easily estimated using the Lindley's equation [80]; 2) it is the key parameter affecting both packet delay and loss.

Based on such considerations, we propose to represent the optimization objective *f* in terms of the aggregate size of the queues over the long-term period $t+1, \ldots, t+k$ in the future as:

$$f = \sum_{i=1}^{n} \sum_{j=1}^{k} Q_i(t+j) \tag{47}$$

Then, the optimization problem for resource allocation is formulated as follows. We assume that:

1. at any time slot t the traffic generated by the users and the size of the queue, $A_i(t)$ and $Q_i(t)$, can be observed and collected from all eNBs in the network;
2. at any time slot t the service rates of eNBs, $X_i(t)$, can be adjusted in such way that the aggregated bandwidth of the eNBs does not exceed the total available bandwidth b, i.e.:

$$\sum_{i=1}^{n} b_i(t) = \sum_{i=1}^{n} SE_i \cdot X_i(t) \le b, \quad 0 \le \omega_i \le 1 \tag{48}$$

3. at any time slot t we can make short-term and long-term predictions of the traffic generated by the users, $\hat{A}_i(t+1)$, …, $\hat{A}_i(t+k)$, in all eNBs.[2]

With these assumptions the size of the queues over the long-term period $t+1$, …., $t+k$ in the future can be estimated recursively using the Lindley's equation [80]:

$$\hat{Q}_i(t+j) = \begin{cases} \left[Q_i(t+j-1) + \hat{A}_i(t+j) - X_i(t+j) \right]^+, & j = 1 \\ \left[\hat{Q}_i(t+j-1) + \hat{A}_i(t+j) - X_i(t+j) \right]^+, & j = 2,..., k \end{cases} \tag{49}$$

If we denote via $x = [X^T(t+1), \ldots, X^T(t+k)]^T$ the non-negative vector of service rate allocations at eNBs, then the optimization problem will be to find the optimal service rate allocation $x^* = [X^{*T}(t+1), \ldots, X^{*T}(t+k)]^T$ that will minimize the aggregated size of the queues during the long-term period in the future:

$$\begin{aligned} \min \quad & f(x) = \sum_{i=1}^{n} \sum_{j=1}^{k} \hat{Q}_i(t+j) \\ \text{subject to} \quad & g_{jn-n+i}(x) = -X_i(t+j) \le 0, \ 1 \le j \le k, \ 1 \le i \le n \\ & g_{kn+j}(x) = \sum_{i=1}^{n} SE_i \cdot X_i(t+j) - b \le 0, \ 1 \le j \le k \end{aligned} \tag{50}$$

[2] In this work we used PLR method applied with AR(1) time-series model for traffic prediction. Detailed description of this technique and the model has been already provided in Chapter 1.

3.2 Smooth Approximation of the Optimization Objective

In (50) the inequality constraints are linear, but the objective function is a non-smooth convex function. In this paper we will follow the approach presented in [162] for non-smooth convex optimization problem. The main idea is to construct a sequence of convex functions $\hat{f}_0(x),\ \hat{f}_1(x),\ \hat{f}_2(x),...$ such that [162]:

$$\lim_{n\to\infty} \hat{f}_n(x) = f(x) \tag{51}$$

Subsequently, solve (50) via a sequential convex optimization approach as follows:

$$\begin{aligned} & x_l = \arg\min_x \hat{f}_l(x) \\ & \text{subject to} \quad g_{jn-n+i}(x) = -X_i(t+j) \le 0,\ 1 \le j \le k,\ 1 \le i \le n \\ & \qquad\qquad g_{kn+j}(x) = \sum_{i=1}^{n} SE_i X_i(t+j) - b \le 0,\ 1 \le j \le k \end{aligned} \tag{52}$$

Then we solve $\hat{x}_0, \hat{x}_1, \hat{x}_2, ...$ until the sequence converges. To simplify the notation let us define:

$$h(\alpha,\beta) = \lceil \alpha - \beta \rceil^+ = \begin{cases} \alpha - \beta, & \text{if } \alpha \ge \beta \ge 0 \\ 0, & \text{otherwise} \end{cases} \tag{53}$$

Clearly, (49) can be written as:

$$\hat{Q}_i(t+j) = \begin{cases} h\{Q_i(t+j-1) + \hat{A}_i(t+j), X_i(t+j)\}, & j = 1 \\ h\{\hat{Q}_i(t+j-1) + \hat{A}_i(t+j), X_i(t+j)\}, & j = 2,...,k \end{cases} \tag{54}$$

Now consider the functions:

$$\hat{h}_l(\alpha,\beta) = \frac{1}{l}\ln[1 + \exp\{l(\alpha-\beta)\}],\quad l = 1, 2, 3, ...,\quad \alpha \ge 0,\ \beta \ge 0 \tag{55}$$

It is rather straightforward to verify that:

$$\lim_{n\to\infty} \left| \hat{h}_n(\alpha,\beta) - h(\alpha,\beta) \right| = 0 \tag{56}$$

uniformly for $\alpha \geq 0$, $\beta \geq 0$. In fact, the convergence is quite fast as illustrated on Figure 52.

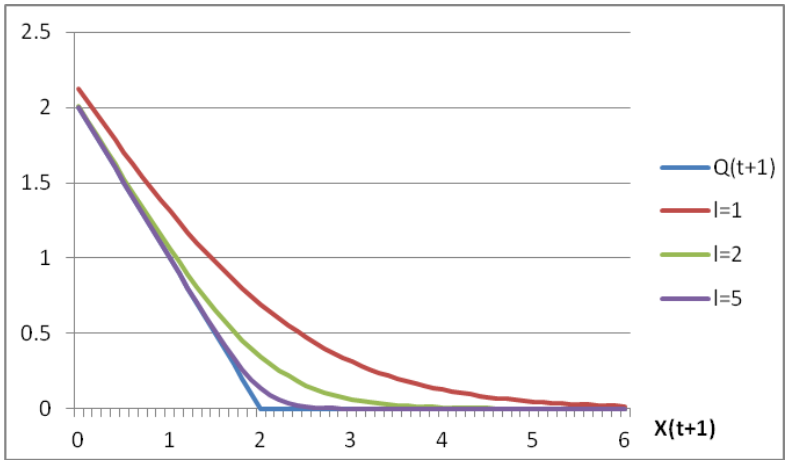


Fig. 52. Real and approximate $Q(t+1)$ with different values of l

With the approximation (55) we can approximate (54) smoothly as:

$$\hat{Q}_i^{(l)}(t+j) = \begin{cases} \hat{h}_l\left\{Q_i(t+j-1) + \hat{A}_i(t+j), X_i(t+j)\right\}, & j=1 \\ \hat{h}_l\left\{\hat{Q}_i^{(l)}(t+j-1) + \hat{A}_i(t+j), X_i(t+j)\right\}, & j=2,...,k \end{cases} \tag{57}$$

Using (57) we then use a smoothed approximation of $f(x)$ as, see (50):

$$\hat{f}_l(x) = \sum_{i=1}^{n}\sum_{j=1}^{k}\hat{Q}_i^{(l)}(t+j) \tag{58}$$

Note, that (58) implicitly depends only on $\{Q_i(t)\}_{i=1}^n$. This is because we can use (54) recursively to express $Q_i^{(l)}(t+k), Q_i^{(l)}(t+k-1),..., Q_i^{(l)}(t+1)$ in terms of $\{Q_i(t)\}_{i=1}^n$. For instance,

$$\hat{Q}_i^{(l)}(t+j) = \hat{h}_l\left[\hat{h}_l\left\{\hat{Q}_i^{(l)}(t+j-2) + \hat{A}_i(t+j-1), X_i(t+j-1)\right\} + \hat{A}_i(t+j), X_i(t+j)\right] \tag{59}$$

and the recursive procedure can be repeated until the expression is given only in terms of $Q_i(t)$. For the purpose of numerical algorithm development this underlying recursion can be utilized efficiently.

The recursive procedures to compute the first and second derivatives of $\hat{Q}_i^{(l)}(t+j)$ with respect to x are derived as follows. The first-order derivatives of $\hat{h}_l(\alpha,\beta)$ with respect to α and β are given by:

$$\frac{\partial \hat{h}_l(\alpha,\beta)}{\partial \alpha} = \frac{\exp\{l(\alpha-\beta)\}}{1+\exp\{l(\alpha-\beta)\}}, \frac{\partial \hat{h}_l(\alpha,\beta)}{\partial \beta} = -\frac{\exp\{l(\alpha-\beta)\}}{1+\exp\{l(\alpha-\beta)\}} \tag{60}$$

The second-order derivatives of $\hat{h}_l(\alpha,\beta)$ with respect to α and β are given by:

$$\begin{aligned} &\frac{\partial^2 \hat{h}_l(\alpha,\beta)}{\partial \alpha^2} = \frac{\partial^2 \hat{h}_l(\alpha,\beta)}{\partial \beta^2} = \frac{l\exp\{l(\alpha-\beta)\}}{(1+\exp\{l(\alpha-\beta)\})^2}, \\ &\frac{\partial^2 \hat{h}_l(\alpha,\beta)}{\partial \alpha \partial \beta} = -\frac{l\exp\{l(\alpha-\beta)\}}{(1+\exp\{l(\alpha-\beta)\})^2} \end{aligned} \tag{61}$$

Let:

$$\eta_l(\alpha,\beta) = \frac{\exp\{l(\alpha-\beta)\}}{1+\exp\{l(\alpha-\beta)\}}, \xi_l(\alpha,\beta) = \frac{l\exp\{l(\alpha-\beta)\}}{(1+\exp\{l(\alpha-\beta)\})^2} \tag{62}$$

then:

$$\begin{aligned} &\frac{\partial \hat{Q}_i^{(l)}(t)}{\partial x} = \eta_l\left[\hat{Q}_i^{(l)}(t-1) + \hat{A}_i(t), X_i(t)\right]\cdot\left(\frac{\partial \hat{Q}_i^{(l)}(t-1)}{\partial x} - \frac{\partial X_i(t)}{\partial x}\right), \\ &\frac{\partial^2 \hat{Q}_i^{(l)}(t)}{\partial x^2} = \xi_l\left[\hat{Q}_i^{(l)}(t-1) + \hat{A}_i(t), X_i(t)\right]\cdot\left(\frac{\partial^2 \hat{Q}_i^{(l)}(t-1)}{\partial x^2} - \frac{\partial^2 X_i(t)}{\partial x^2}\right) \end{aligned} \tag{63}$$

Similar to (54), expressions (55) and (56) can be used recursively to express $\frac{\partial Q_i^{(l)}(t+k)}{\partial x}$, $\frac{\partial Q_i^{(l)}(t+k-1)}{\partial x}$,..., $\frac{\partial Q_i^{(l)}(t+1)}{\partial x}$ and $\frac{\partial^2 Q_i^{(l)}(t+k)}{\partial x^2}$, $\frac{\partial^2 Q_i^{(l)}(t+k-1)}{\partial x^2}$,..., $\frac{\partial^2 Q_i^{(l)}(t+1)}{\partial x^2}$ in terms of

$\left\{\frac{\partial Q_i(t)}{\partial x}\right\}_{i=1}^{n}$ and $\left\{\frac{\partial^2 Q_i(t)}{\partial x^2}\right\}_{i=1}^{n}$, respectively. For this, the recursive procedures are given by:

$$\frac{\partial \hat{Q}_i^{(l)}(t)}{\partial x} = \left(\frac{\partial \hat{Q}_i^{(l)}(t-1)}{\partial x} - \frac{\partial X_i(t)}{\partial x}\right) \times$$

$$\eta_l\left\{\eta_l\left\{\hat{Q}_i^{(l)}(t-2) + \hat{A}_i(t-1), X_i(t-1)\right\}\left(\frac{\partial \hat{Q}_i^{(l)}(t-2)}{\partial x} - \frac{\partial X_i(t-1)}{\partial x}\right) + \hat{A}_i(t), X_i(t)\right\},$$

$$\frac{\partial^2 \hat{Q}_i^{(l)}(t)}{\partial x^2} = \left(\frac{\partial^2 \hat{Q}_i^{(l)}(t-1)}{\partial x^2} - \frac{\partial^2 X_i(t)}{\partial x^2}\right) \times$$

$$\xi_l\left\{\xi_l\left\{\hat{Q}_i^{(l)}(t-2) + \hat{A}_i(t-1), X_i(t-1)\right\}\left(\frac{\partial^2 \hat{Q}_i^{(l)}(t-2)}{\partial x^2} - \frac{\partial^2 X_i(t-1)}{\partial x^2}\right) + \hat{A}_i(t), X_i(t)\right\} \tag{64}$$

should be repeated until the expression is given only in terms of $\left\{\frac{\partial Q_i(t)}{\partial x}\right\}_{i=1}^{n}$ and $\left\{\frac{\partial^2 Q_i(t)}{\partial x^2}\right\}_{i=1}^{n}$, respectively.

Now we are ready to redefine the primary optimization problem (50) in terms of approximation $\hat{f}_l(x)$ given by (58) as follows:

$$\begin{aligned} &\min && \hat{f}_l(x) \\ &\text{subject to:} && g_{jn-n+i}(x) = -X_i(t+j) \le 0, \;\; 1 \le j \le k, \;\; 1 \le i \le n \\ & && g_{kn+j}(x) = \sum_{i=1}^{n} SE_i \cdot X_i(t+j) - b \le 0, \;\; 1 \le j \le k \end{aligned} \tag{65}$$

Note, that (65) is a smooth convex optimization problem which can be solved by using any of the standard methods of convex minimization, such as subgradient, subgradient projection, "Bundle methods" [163], interior-point [164], cutting-plane [165], etc. In this work we used the primal-dual interior point algorithm because of its high efficiency and better than linear convergence, especially for the cases when the high accuracy is required [165]. Description of the

primal-dual interior point method can be found in [165]; solution of the problem (65) using this method can be found in the corresponding paper.

3.3 Resource Allocation Algorithm

Proposed resource allocation algorithm is summarized on Figure 53. At time slot t:

1. The algorithm updates the values of $A_i(t)$ and $Q_i(t)$ for AP_1, …, AP_n.
2. Based on the updated values $A_i(t)$, the algorithm make the predictions $\hat{A}_i(t+1)$, …, $\hat{A}_n(t+k)$ for AP_1, …, AP_n using PLR technique applied with AR(1) time-series model (description of the technique and the model has already been provided in Chapter 1).
3. Based on updated values $Q_i(t)$ and predicted values $\hat{A}_i(t+1)$, …, $\hat{A}_n(t+k)$ for AP_1, …, AP_n, the algorithm finds the optimal solution of the problem (16) using primal-dual interior point algorithm [165].
4. At time slot $t+1$ the users will be assigned the optimal service rate $X_i^*(t+1)$.

Algorithm 1. Generic multi-step resource allocation algorithm

Given C, n, k, γ
At time slot $t = 0, 1, 2, ...$
    For each i^{th} user queue
        1. **If** $t > 1$
            set $X_i(t) := X_i^*(t)$
        else
            set $X_i(t) := C/n$
        2. **Update** $A_i(t)$, $Q_i(t)$
        3. **Predict** $\hat{A}_i(t+1)$, …, $\hat{A}_n(t+k)$
    For each i^{th} user queue **find optimal** $X_i^*(t+1)$,..., $X_i^*(t+k)$

Fig. 53. Proposed algorithm for resource allocation

4 Algorithm Performance

4.1 Algorithm Implementation

We now present the example of the algorithm implementation in the standard IEEE 802.22 CR network architecture [51, 52]. Considered network model consists of n LTE eNBs sharing the total available bandwidth b via the SM connected to eNBs via IP links. The SM operates both as a central gateway connecting the wireless network to external networks and as a central processor responsible for resource allocation in CR system.

The following assumptions are made in this particular algorithm implementation:

1. The slot duration in the algorithm is equal $\Delta = 1$ s.
2. In general, the traffic pattern generated by the users in uplink and downlink direction is very different. Therefore, we collect the traffic generated by the users and the size of the queues in all eNBs separately for uplink and downlink direction. We denote the uplink and downlink traffic generated by the users in eNB_i at time slot t by $A_i^{UL}(t)$ and $A_i^{DL}(t)$, respectively; we denote the uplink and downlink size of the queue at eNB_i at time slot t by $Q_i^{UL}(t)$ and $Q_i^{DL}(t)$, respectively.
3. We predict the traffic generated by the users in all eNBs separately for uplink and downlink direction. We denote the uplink and downlink predictions of the traffic generated by the users of eNB_i during time period $t+1, \ldots, t+k$ via $\hat{A}_i^{UL}(t+1), \ldots, \hat{A}_i^{UL}(t+k)$ and $\hat{A}_i^{DL}(t+1), \ldots, \hat{A}_i^{DL}(t+k)$, respectively.
4. We find the optimal service rate allocation separately for uplink and downlink direction, and denote via $X_i^{UL}(t)$ and $X_i^{DL}(t)$ the service rate assigned to eNB_i at time slot t in uplink and downlink directions, respectively.
5. The bandwidth is assigned to eNB_i based on the obtained uplink and downlink service rate allocation $X_i^{UL}(t)$ and $X_i^{DL}(t)$ using the expression:

$$b_i(t) = \frac{X_i^{UL}(t) + X_i^{DL}(t)}{SE_i}, \; i = 1, ..., n \tag{66}$$

where the values of SE_i are collected on the physical layer (PHY) of the corresponding eNB (based on the physical channel characteristics and the channel quality).

Based on these assumptions, the proposed scheme for resource allocation can be described as follows. At time slot t:

1. Each eNB collects the values of $A_i^{UL}(t)$ and $A_i^{DL}(t)$, $Q_i^{UL}(t)$ and $Q_i^{DL}(t)$ and SE_i and sends them to SM via IP links.
2. Based on updated values $A_i^{UL}(t)$ and $A_i^{DL}(t)$, SM makes the predictions $\hat{A}_i^{UL}(t+1)$, …, $\hat{A}_i^{UL}(t+k)$ and $\hat{A}_i^{DL}(t+1)$, …, $\hat{A}_i^{DL}(t+k)$ for each eNB in the network.
3. Based on updated values $Q_i^{UL}(t)$ and $Q_i^{DL}(t)$ and predicted values $\hat{A}_i^{UL}(t+1)$, …, $\hat{A}_i^{UL}(t+k)$ and $\hat{A}_i^{DL}(t+1)$, …, $\hat{A}_i^{DL}(t+k)$ SM finds the optimal service rate allocation at the next time slot $X_i^{UL}(t+1)$ and $X_i^{DL}(t+1)$ using the primal-dual interior point algorithm [165], calculates the bandwidth to be assigned to eNBs at the next time slot $b_i(t)$ using (66), and sends the values $b_i(t)$ to respective eNBs.

4.2 Simulation Model

A simulation model of the network has been developed upon the OPNET platform [112]. The model consists of $n = 7$ eNBs sharing the total available bandwidth $b = 35$ MHz, EPC with advanced SM functionalities, and a number of fixed user equipments (UEs). EPC is connected to the eNBs via 1Gbit/s data rate IP links. The radio network model has been developed according to the requirements of ITU-T Recommendation M.1225. Other simulation parameters have been set in accordance with LTE specifications [129, 133, 134] (the simulation parameters of the network model are listed in Table 14).

Heterogeneous user traffic in simulations comprises voice, video and data applications. The following models have been used to simulate voice, video and data services.

- The voice over IP (VoIP) service uses ON-OFF model with exponentially distributed ON-OFF periods. The mean duration of ON and OFF periods are 0.65s and 0.352s, respectively. The VoIP traffic is generated using G.723.1 codec with 12.2 Kbps rate, payload size equal 40 bytes and a payload interval equal 30 ms [167].

- Video services are simulated using a high resolution video model with a constant frame size equal 6250 bytes and exponentially distributed frame inter-arrival intervals (with mean equal 0.5s) [167].
- Data applications are represented by the HTTP1.1 model with exponential page/image inter-arrival intervals (mean equal 60sec). It is assumed that one page consists of one object, whereas one image consists of five objects. The object size is constant and equal 1000 bytes [167].

Table 14. Simulation Parameters of the Model

Parameter		*Value*
Radio Network Model:	Pass loss	$L=40\log_{10}R+30\log_{10}f+49$, R – distance (km), f – carrier frequency (Hz)
	Shadow fading	Log-normal shadow fading with a standard deviation of 10/12 dB for outdoor/indoor users
	Penetration loss	The average building penetration loss is 12 dB with a standard deviation of 8 dB
	Multipath fading	Spatial Channel Model (SCM), Suburban macro
	Cell radius	1 km
	UE velocity	0 km/s
	Transmitter/Receiver antenna gain	10 dBi (pedestrian), 2 dBi (indoor)
	Receiver antenna gain	10 dBi (pedestrian), 2 dBi (indoor)
	Receiver noise figure	5 dB
	Thermal noise density	–174 dBm/Hz
	Cable, connector, and combiner losses	2 dB
PHY profile:	Operation mode	Frequency Division Duplex (FDD)
	Cyclic Prefix Type	Normal (7 Symbols per Slot)
	EPC Bearer Definitions	348kbit/s
	Carrier frequency	2GHz
	Subcarrier spacing	15kHz

Admission Control Parameters:	Physical Downlink Control Channel (PDCCH) symbols per subframe	3
	UL Loading Factor	1
	DL Loading Factor	1
	Inactive Bearer Timeout	20 sec
Buffer Status Report Parameters:	Periodic Timer	5 subframes
	Retransmission Timer	2560 subframes
Layer1/Layer2 Control Parameters:	Reserved Size	2 RBs
	Cyclic Shifts	6
	Starting PRB for Format 1 messages	0
	Allocation Periodicity	5 subframes
Random Access (RA) Parameters:	Number of Preambles	64
	Preamble Format	Format 0 (1-subframe long)
	Number of RA Resources per Frame	4
	Preamble Retransmission Limit	5 subframes
	RA Response Timer	5 subframes
	Contention Resolution Timer	40 subframes
Hybrid Automatic Repeat ReQuest (HARQ) Parameters:	Maximal Number of Retransmissions	3 (UL and DL)
	HARQ Retransmission Timer	8 subframes (UL and DL)
	Maximal Number of HARQ processes	8 per UE (UL and DL)

In this work, performance of the proposed algorithm is compared with the performance of two most recent techniques designed to deal with heterogeneous user traffic: max-utility bandwidth allocation described in [77] and bandwidth allocation for the users with heterogeneous utilities described in [78]. In max-utility resource allocation the spectrum is assumed to be discrete: the total available bandwidth b is divided into N resource blocks, and within one time slot one resource block can be used only by one base station. Other settings of the algorithm are adopted from [77]. In the scheme with heterogeneous user utilities the spectrum is continuous. The bandwidth is assigned according to the distributed algorithm based on user demands of eNBs subject to the power and capacity constraints. Detailed description of the algorithm is given in [78].

We use the following abbreviations to differentiate the bandwidth allocation techniques: “Max Utility” for the algorithm described in [77], “Het Utility” for the scheme proposed in [78], “Singlestep” for the proposed algorithm with $k = 1$ and “Multistep” for the proposed algorithm with $k = 3$. We also benchmark the performance of these schemes with the performance of the simple fixed equal bandwidth allocation (denoted via “FA”) when all eNBs are assigned fixed bandwidth $b/n = 5$ MHz.

4.3 Simulation Results

Performance of different bandwidth allocation schemes in scenarios with low (< 18 Mbits/s), medium (18 ÷ 36 Mbits/s) and high (> 36 Mbits/s) load is summarized on Figures 54 - 57. Figures 54 and 55 show the total (UL and DL) MAC layer packet delay and loss for all user applications. Figures 56, 57 illustrate the application layer packet end-to-end delay (which consists of total MAC layer delay, coder/decoder delay, packetization and serialization delay, and de-jitter buffer delay) for voice and video users.

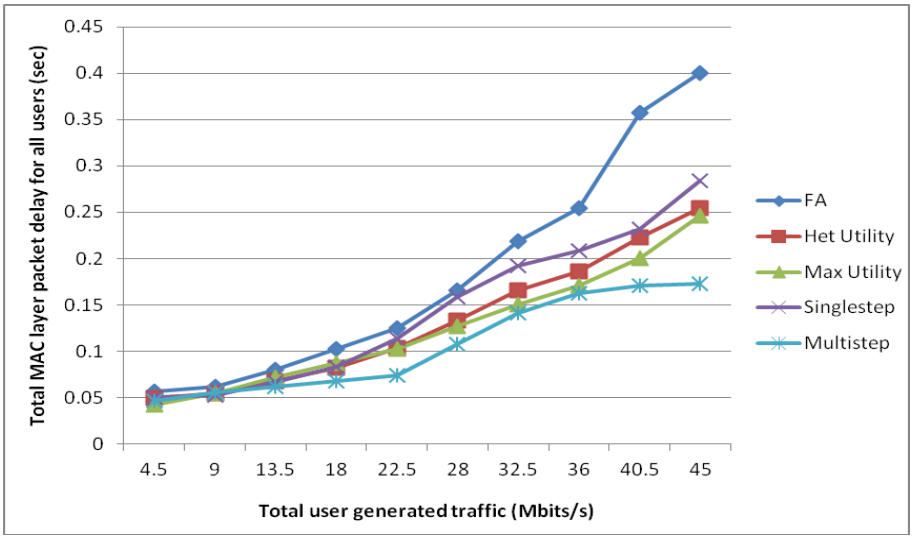


Fig. 54. Total MAC layer packet delay for all types of applications

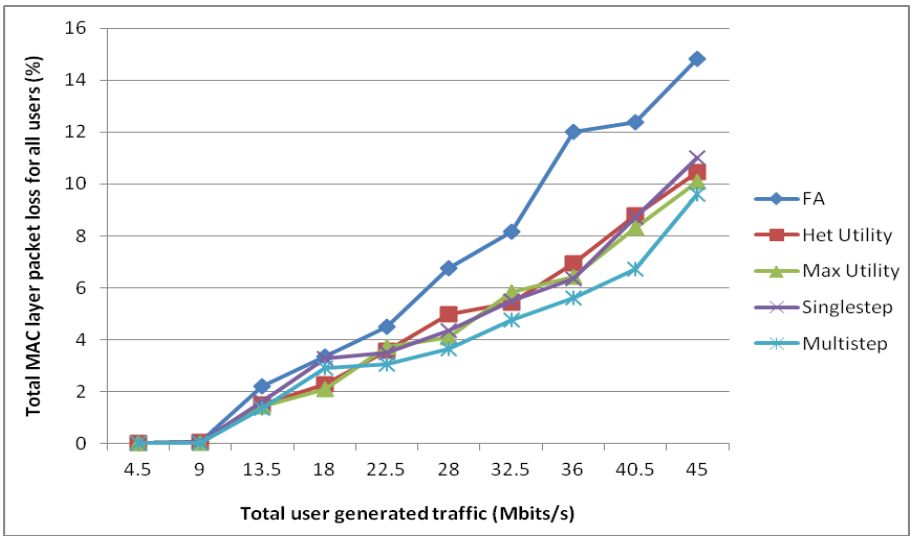


Fig. 55. Total MAC layer packet loss for all types of applications

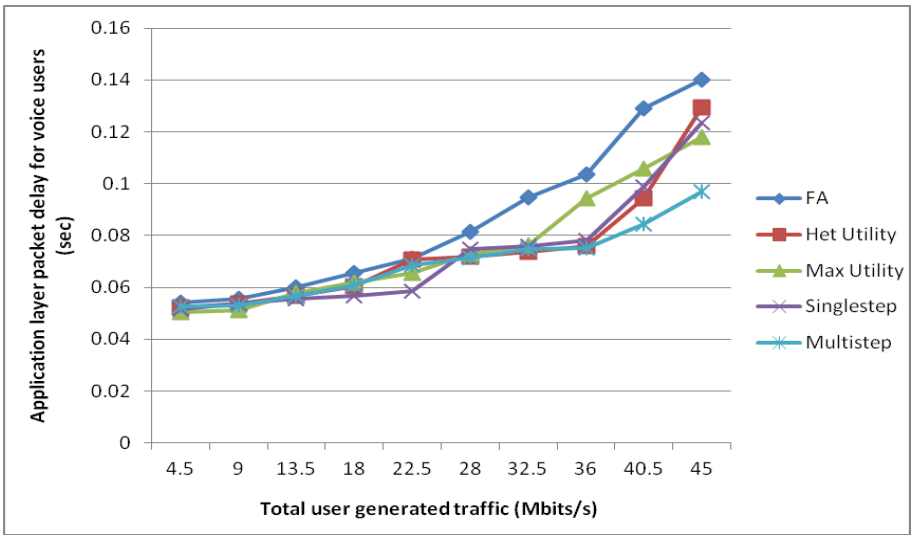


Fig. 56. Application layer packet delay for voice applications

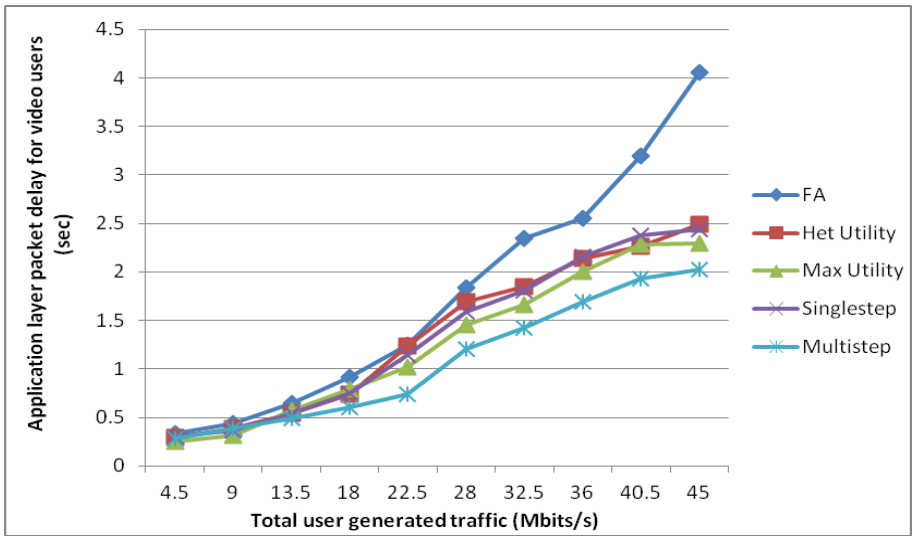


Fig. 57. Application layer packet delay for video applications

It follows from these graphs that:

- FA demonstrates the worst (highest delay and loss) performance compared to all other schemes in medium and high load scenarios;
- Het Utility, Max Utility and Singlestep bandwidth allocation show almost similar performance in all scenarios;
- Multistep bandwidth allocation outperforms all other schemes in scenarios with medium and high load.

These results show that the proposed "multi-step approach" can be effectively used for resource allocation in CRNs with heterogeneous user traffic

CHAPTER 7: A Two-Stage Resource Allocation Procedure for Cognitive LTE Network

In this chapter a resource allocation technique for a cognitive LTE network in Scenario 2 is presented. Description of the network deployment scenario, as well as previous research on resource allocation for LTE-based CRN in Scenario 2 had already been summarized in Overview of this thesis. Here we present the proposed approach for resource allocation, derive two different algorithms for resource allocation, and present the results of algorithm performance based on simulation model developed in OPNET environment [112]. The corresponding paper has been published in Computer Networks, 2014.

1 Introduction

In this work we consider a problem of resource allocation for the Third Generation Partnership Project (3GPP) long-term evolution (LTE) cognitive radio network (CRN). The CRN is made up of the licensed (primary) service stations which can share their spectrum resources with unlicensed (secondary) stations. The goal of resource allocation is to provide the wireless access to secondary stations without compromising the quality of service (QoS) for primary stations. To accomplish this goal, we utilize a simple procedure consisting of two stages. During the first stage the spectrum resources are allocated to primary stations to maximize the QoS for primary users. During the second stage the rest of the service capacity of the primary channels is distributed among secondary stations.

Theoretical framework conducted in the paper is closely tightened with the specifics of LTE design. In particular, the problem of resource allocation is formulated as an integer-programming optimization problem based on assumption that the spectrum in LTE system is discrete (with a resource allocation granularity of 180 kHz in frequency domain and 1 ms in time domain).

Two algorithms of different complexities are derived in the paper based on the proposed two-stage resource allocation procedure. Both algorithms do not involve additional network signaling over the wireless medium. First algorithm is more suitable for implementation in CRN with light and smooth traffic and/or when the processing capabilities of CRN are low and restrictive (because of its simplicity and short running time). Second algorithm can be used in the network with heavy and/or bursty traffic. However, the implementation of this algorithm requires high processing capabilities.

The rest of the paper is organized as follows. In section 2 we formulate the main idea behind the proposed resource allocation approach. Om section 3 we derive two algorithms for resource allocation in LTE-based CRN architecture. In section 4 we conduct the comparative performance analysis of the algorithm.

2 Resource Allocation Procedure

Considered network model consists of n primary (licensed) eNBs (PBs) numbered PB_1, PB_2, ..., PB_n, and m secondary (unlicensed) eNBs (SBs) numbered SB_1, SB_2, ..., SB_m. The eNBs are connected to the backbone server via a central network manager (CNM). The communication between the eNBs, CNM and the server is realized using high-speed IP links to facilitate fast data transmission (Figure 5).

Each primary eNB operates on its fixed licensed spectrum band (primary channel) with some certain serving capacity. The primary eNB can share the primary channel with one or more secondary eNBs which don't have fixed licensed spectrum bands. The PBs have prioritized access to the primary channel. The capacity (in bps) that a primary eNB station shares with a certain secondary base station depends on the channel allocation policy used by CNM.

An LTE network operates on a slotted-time basis, i.e. the time axis is partitioned into mutually disjoint time intervals (slots) $\{tT_s, (t + 1)T_s\}$, $t = 0, 1, 2, ...$, with T_s denoting the slot length and t denoting the slot index. Each eNB serves a number of wireless users located within its service area (cell). The user-generated traffic (in bps) is enqueued in the user equipments (UEs), and then transmitted to respective eNBs using the packet scheduling procedure described in LTE standard [81]. In this

procedure, the information about the amount of data (in bps) enqueued in the buffers of UEs are constantly transmitted to the eNB, so that the base station "knows" the exact amount of data generated by the users at any time slot *t*. This information is used by the eNB to allocate the uplink transmission resources to UEs using a certain scheduling algorithm.

In the downlink, the transmission resources are allocated based on the amount of data arriving from CNM via IP link. Depending on the algorithm implementation, the packet scheduling can be based on the quality of service (QoS) requirements, instantaneous channel conditions, fairness, etc. [81]. The CRN architecture described above can be well represented by the system model comprising the set of *n* primary channels belonging to PB_1, PB_2, ..., PB_n and *m* secondary channels belonging to SB_1, SB_2, ..., SB_m. Each eNB (PB or SB) in the model is represented by the uplink and downlink infinite queue (primary or secondary).

The general system model of CRN is illustrated on Figure 58. Note, that this model can be applied to both uplink and downlink directions. Similarly, the discussion in the rest of the paper is applicable to both directions.

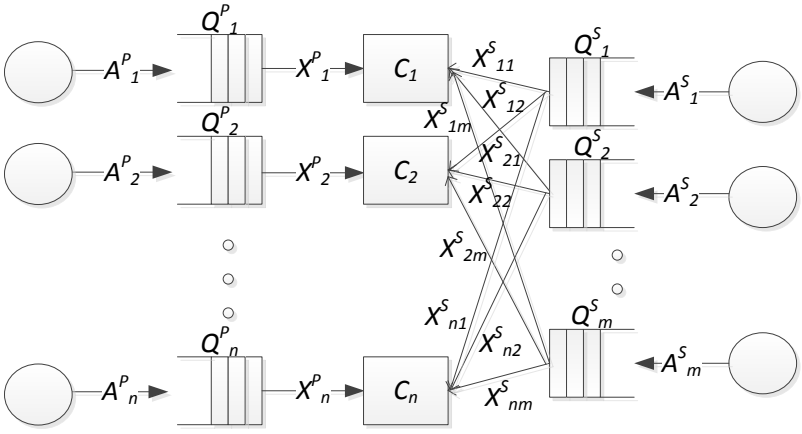


Fig. 58. System model of CRN

Here and after we use the notation:

- $Q^P_i(t)$ - the size of the queue (in bits) at PB_i at the beginning of t^{th} time slot;
- $Q^S_j(t)$ - the size of the queue (in bits) at SB_i at the beginning of t^{th} time slot;
- $A^P_i(t)$ - the total amount of data (in bits/slot or bps) generated during t^{th} time slot by the wireless users of PB_i;
- $A^S_j(t)$ - the total amount of data (in bps) generated during t^{th} time slot by the wireless users of SB_i;
- $X^P_i(t)$ - the service rate (in bps) of PB_i during t^{th} time slot;
- $X^S_{ij}(t)$ - the service capacity (in bps) that PB_i shares with SB_j during t^{th} time slot;
- C_i – the service capacity (in bps) of primary channel belonging to PB_i.

Note, that for any $i \in \{1, 2,..., n\}$, $j \in \{1, 2,..., m\}$ at any time slot t the values $Q^P_i(t)$, $Q^S_j(t)$, $A^P_i(t)$, $A^S_j(t)$ can be observed in respective PB_i/SB_j. Also, the total number of bits served by a primary channel cannot exceed it maximum service capacity, i.e.:

$$X^P_i(t) + \sum_{j=1}^{m} X^S_{ij}(t) \le C_i,\ i = 1,..., n \tag{67}$$

Ideally, CRN should allocate the service rates $X^P_i(t)$ and $X^S_{ij}(t)$ in order to maximize the quality of service (QoS) for the users of PBs and SBs, and preserve the service priority of PBs in their channels. One way to achieve this goal is to perform the service rate allocation separately for primary and secondary base stations using a two-stage procedure. In the first stage we allocate the service rate for all PNs within CRN to maximize the QoS for primary network users. Then, the unused service capacity of the primary channels can be distributed among SBs.

In this work we propose to allocate the service rates $X^P_i(t)$, $X^S_{ij}(t)$ based on the size of the queues $Q^P_i(t + 1)$ and $Q^S_j(t + 1)$. Thus, to maximize the QoS for the users of SBs and PBs and maintain the service priority of PBs, we should first minimize $Q^P_i(t + 1)$ and $Q^S_j(t + 1)$ for all PBs. Then utilize the unused service capacity such that $Q^S_j(t + 1)$ is minimized.

The queue size was chosen as an optimization target because it is directly connected to the main QoS metrics, such as round-trip latency and loss. However, the analysis and estimation of these metrics in wireless networks is rather complex, and leads to very difficult optimization problems, whereas $Q^P_i(t + 1)$ and $Q^S_j(t + 1)$ can be easily estimated from the known values $Q^P_i(t)$, $Q^S_j(t)$ and $A^P_i(t)$, $A^S_j(t)$ using Lindley's equation [80]:

$$Q_i^P(t+1) = \left\lceil Q_i^P(t) + A_i^P(t) - X_i^P(t) \right\rceil^+, \; i = 1,..., n \tag{68}$$

$$Q_j^S(t+1) = \left\lceil Q_j^S(t) + A_j^S(t) - \sum_{i=1}^{n} X_{ij}^S(t) \right\rceil^+, \; j = 1,..., m \tag{69}$$

where:

$$\lceil x \rceil^+ = \max(0, x) \tag{70}$$

In (68) and (69) the service rates $X^P_i(t)$, $X^S_{ij}(t)$ should be allocated, while the values of $Q^P_i(t)$, $Q^S_j(t)$ and $A^P_i(t)$, $A^S_j(t)$ are observed in all PB_i, SB_j at any time slot t.

We propose to allocate $X^P_i(t)$ such that $Q^P_i(t + 1)$ is minimized. Thus, from (68) it is straightforward to verify that

$$X_i^P(t) = \begin{cases} C_i, & Q_i^P(t) + A_i^P(t) \geq C_i \\ Q_i^P(t) + A_i^P(t), & \text{otherwise} \end{cases}, \quad i = 1,..., n \tag{71}$$

In practice, the network is not fully loaded most of the time. Hence we would have $C_i > X^P_i(t)$ for some values of i. This allows the corresponding primary eNBs to serve some secondary users. The service rates for the secondary users are determined by solving a minimax problem. Define the unknown service rate allocation vectors as

$$\mathbf{X}_i^S(t) = \begin{bmatrix} X_{i1}^S(t) \\ : \\ X_{im}^S(t) \end{bmatrix}, \quad \mathbf{X}^S(t) = \begin{bmatrix} \mathbf{X}_1^S(t) \\ : \\ \mathbf{X}_n^S(t) \end{bmatrix} \tag{72}$$

and the function

$$f\{\mathbf{X}^S(t)\} = \max_{1\le j\le m}\left\lceil Q_j^S(t) + A_j^S(t) - \sum_{i=1}^{n} X_{ij}^S(t)\right\rceil^+ \tag{73}$$

We propose to compute $\mathbf{X}^S$ by solving the optimization problem

$$\begin{aligned} &\text{minimize} \quad f\{\mathbf{X}^S(t)\} \\ &\text{subject to} \quad 0 \le X_{ij}^S(t), \;\; i = 1,...,n, \;\; j = 1,...,m, \\ &\qquad\qquad X_i^P(t) + \sum_{j=1}^{m} X_{ij}^S(t) \le C_i, \;\; i = 1,...,n. \end{aligned} \tag{74}$$

The problem (74) is a convex. It is possible to solve (74) in polynomial time using some interior point method. These are true when the components of $\mathbf{X}^P$ and $\mathbf{X}^S$ are allowed to take real numbers. In reality, however, for LTE the components of $\mathbf{X}^P$ and $\mathbf{X}^S$ are not allowed to take arbitrary real values [169]. This complicates the problem to some extent. Nevertheless, we can use the methods of convex optimization and integer programming to efficiently solve the resulting optimization problems.

3 Resource Allocation in LTE System

To facilitate the description of the proposed resource allocation approach, recall that the main transmission unit over the air interface in an LTE system is called a resource block pair. Each resource block pair has a duration 1ms and is made up of 12 subcarriers, or 180 kHz. The peak service capacity of a resource block depends on the antenna configuration, modulation and coding scheme and the type of cyclic prefix (normal or extended) used in the system. More detailed description of the peak data rate calculation in LTE system can be found in [169].

For 2 × 2 MIMO channel with 64 QAM modulation and normal cyclic prefix the downlink peak data rate of one resource block pair is equal 1.47 Mbis/s. The corresponding uplink peak data rate of the resource block pair (1 × 2 MIMO channel with 64QAM modulation and normal cyclic prefix) is equal to 0.74 Mbits/s [169]. The number of resource blocks corresponding to different spectrum bands in LTE system has already been provided in Table 5 (Chapter 3). The uplink

and downlink peak data rate corresponding to different spectrum bands is shown on Figure 59.

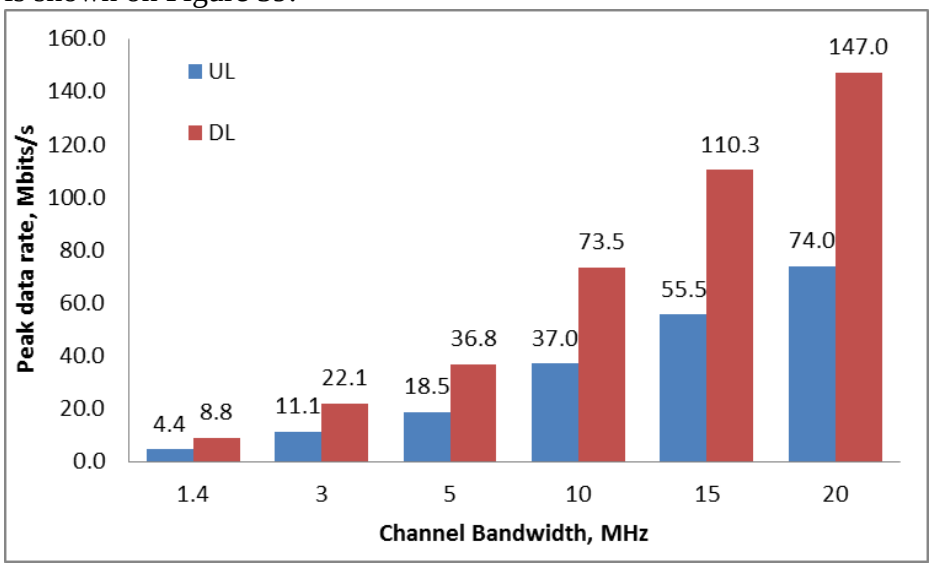


Fig. 59. Peak data rates for different channel bandwidth in standard LTE system [169]

Given the bandwidth of the primary channel of PB_i, we can find the corresponding number of resource blocks N_i. In LTE the resource allocation can be performed only in terms of resource blocks. Thus, at any time slot an eNB can allocate only integer number of resource blocks to each user.

3.1 Algorithm 1 (for Light and/or Smooth Network Traffic)

Basia Idea

Let c_i be the peak data rate of one resource block of PB_i. Let $x^P_i(t)$ be the number of resource blocks allocated to the primary users in PB_i. Then

$$X_i^P(t) = c_i x_i^P(t), \quad C_i = c_i N_i, \quad i = 1,..., n \tag{75}$$

Similarly, if $x^S_{ij}(t)$ is the number of resource blocks that PB_i shares with SB_j at time slot t, then

$$X_{ij}^{S}(t) = c_i x_{ij}^{S}(t), \quad j = 1,..., m \tag{76}$$

According to our strategy described before we allocate $x^{P}_{i}(t)$ in order to minimize Q^{P}_{i}. Thus denoting

$$\overline{N}_i = \frac{Q_i^P(t) + A_i^P(t)}{c_i}, \quad i = 1,..., n \tag{77}$$

it is straightforward to verify that

$$x_i^P(t) = \begin{cases} N_i, & \text{if } \overline{N}_i \geq N_i \\ \lceil \overline{N}_i \rceil, & \text{otherwise} \end{cases}, \quad i = 1,..., n \tag{78}$$

In practice, often we would have spare capacity, i.e. $\overline{N}_i < N_i$ will hold for many values of i. Then we can utilize the spare capacity to serve the secondary users. To find the individual allocations, we solve (74) subject to the constraint that $x^{S}_{ij}(t)$ is an integer. Define the unknown service rate allocation vectors

$$\mathbf{x}_i^S(t) = \begin{bmatrix} x_{i1}^S(t) \\ : \\ x_{im}^S(t) \end{bmatrix}, \quad \mathbf{x}^S(t) = \begin{bmatrix} \mathbf{x}_1^S(t) \\ : \\ \mathbf{x}_n^S(t) \end{bmatrix} \tag{79}$$

and the function

$$\bar{f}\{\mathbf{x}^S(t)\} := \max_{1 \leq j \leq m} \left[Q_j^S(t) + A_j^S(t) - \sum_{i=1}^{n} c_i x_{ij}^S(t) \right]^+ \tag{80}$$

Then we compute $\mathbf{x}^{S}(t)$ by solving the optimization problem

$$\begin{aligned} &\text{minimize} \quad \bar{f}\{\mathbf{x}^S(t)\} \\ &\text{subject to} \quad x_{ij}^S(t) \in \mathbf{Z}^+, \quad i = 1,..., n, \quad j = 1,..., m, \\ &\qquad\qquad x_i^P(t) + \sum_{j=1}^{m} x_{ij}^S(t) \leq N_i, \quad i = 1,..., n. \end{aligned} \tag{81}$$

In (81) $\mathbf{Z}^+$ denotes the set of non-negative integers. The corresponding resource allocation algorithm is fomulated as follows. For each time slot *t*:

1. All SBs/PBs collect and send the values $Q^P_i(t)$, $Q^S_j(t)$ and $A^P_i(t)$, $A^S_j(t)$ to CNM;
2. CNM calculates and sends the optimal integer solutions $x^P_i(t)$, $x^S_{ij}(t)$ to all SBs/PBs;
3. The resources of the primary channel belonging to PB_i are shared by the PB_i which occupies $x^P_i(t)$ resource blocks, and SB(s) which occupy $x^S_{ij}(t)$ resource blocks.

Solution Methodology

In theory, (81) is a hard integer programming problem due to integer restrictions on $x^S_{ij}(t)$. Nevertheless, many efficient methods for solving such integer programming problems exist. Here we propose to solve (81) as follows. First, we relax integer restrictions on $x^S_{ij}(t)$, and solve the convex problems given by

$$\begin{aligned} &\text{minimize} \quad \bar{f}\{\mathbf{x}^S(t)\} \\ &\text{subject to} \quad x^S_{ij}(t) \geq 0, \;\; i = 1,...,n, \;\; j = 1,...,m, \\ &\qquad\qquad x^P_i(t) + \sum_{j=1}^{m} x^S_{ij}(t) \leq N_i, \;\; i = 1,...,n. \end{aligned} \tag{82}$$

Let $\bar{\mathbf{x}}^S(t)$ be the solutions of the problem (82). Note, that the components $\bar{x}_{ij}{}^S(t)$ of the vector $\bar{\mathbf{x}}^S(t)$ are in general non-integer. Then we use $\lfloor \bar{x}_{ij}{}^S(t) \rfloor$ to initialize a branch and bound algorithm to solve (81). It is straightforward to verify that such initialization does satisfy the constraints in (81).

Because (82) is a convex problem, we can compute $\bar{\mathbf{x}}^S(t)$ very quickly using some interior point method. By initializing the branch and bound method at $\lfloor \bar{x}_{ij}{}^S(t) \rfloor$ we are already very close to the solution of (81). Hence, we can expect a quick convergence.

Next, we show that (82) can be solved by solving an integer linear programming problem. Note, that (82) is equivalent to

$$\begin{aligned}
&\text{minimize} \quad y \\
&\text{subject to} \quad y \geq \bar{f}\{\mathbf{x}^S(t)\}, \\
&\qquad x_{ij}^S(t) \geq 0, \;\; i=1,...,n, \;\; j=1,...,m, \\
&\qquad x_i^P(t) + \sum_{j=1}^{m} x_{ij}^S(t) \leq N_i, \;\; i=1,...,n.
\end{aligned} \tag{83}$$

In (83) the constraint $y \geq \bar{f}\{\mathbf{x}^S(t)\}$ holds if and only if (see (70), (80)):

$$y \geq 0, \quad y \geq Q_j^S(t) + A_j^S(t) - \sum_{i=1}^{n} c_i x_{ij}^S(t), \quad j=1,...,m \tag{84}$$

Hence, (83) is equivalent to the linear programming problem

$$\begin{aligned}
&\text{minimize} \quad y \\
&\text{subject to} \; y \geq 0, \\
&\qquad y \geq Q_j^S(t) + A_j^S(t) - \sum_{i=1}^{n} c_i x_{ij}^S(t), \quad j=1,...,m \\
&\qquad x_{ij}^S(t) \geq 0, \;\; i=1,...,n, \;\; j=1,...,m, \\
&\qquad x_i^P(t) + \sum_{j=1}^{m} x_{ij}^S(t) \leq N_i, \;\; i=1,...,n.
\end{aligned} \tag{85}$$

Any standard linear program solver can be used to solve the problem (85). In our simulations we have used the primal-dual interior point [165] method.

3.2 Algorithm 2 (for Heavy and/or Bursty Network Traffic)

Basia Idea

When the traffic is heavy and bursty, then the value of $A^S_j(t)$ and $A^P_i(t)$ may vary significantly from one time step to next time step. For this reason it is often useful to predict what could happen in future, and

somehow incorporate this knowledge in the resource allocation algorithm. In this way we can reduce the possibility of getting 'surprizes' when the size of some queue start to build up suddenly. The philosophy of our allocation strategy is to predict such events and take necessary actions ahead in time to make provisions for sudden bursts.

Suppose, that at any time slot t given the past values of the arrival rates $A^P_i(t - k)$, $A^S_j(t - k)$, $k = 0, 1, 2, \ldots$ we can make a sequence of predictions $\left\{\hat{A}_i^P(t+k \mid t), \hat{A}_j^S(t+k \mid t)\right\}_{k=1}^{l}$ for the time period of the certain length $l \geq 1$ time slots. The notation $\hat{A}_i^P(t+k \mid t), \hat{A}_j^S(t+k \mid t)$ should be read as "the predicted value of $A^P_i(t + k)$, $A^S_j(t + k)$ computed at time t using the observations up to time point t". To generate predictions in this work we used PLR prediction technique applied with AR time-series model (which have been described in Chapter 1).

Here and after we refer to l as the 'prediction horizon'. Here at any timeslot t we compute the optimal allocations $\left\{x_i^P(t+k), x_{ij}^S(t+k)\right\}_{k=0}^{l}$ by solving an optimization problem. This optimal solution accounts for possible events in future within the prediction horizon. Then we implement only $x^P_i(t)$ and $x^S_{ij}(t)$, and move on to the next time step. In this way the allocations become future-aware, and are somewhat immuned to some future events which could cause problems otherwise. This idea of future aware allocation is similar to reinforcement learning [161] and model predictive control [160], popular in machine learning and automatic control, respectively.

In the following the notation $x^P_i(t + k|t)$, $x^S_{ij}(t + k|t)$ denote the value of $x^P_i(t + k)$, $x^S_{ij}(t + k)$ computed based on the result of optimization run at time t. As before, our strategy is to minimize the largest queue sizes in some way, but in this case we consider the queue sizes over the whole prediction horizon. We start by making allocations for the primary users. By using the predicted value $\hat{A}_i^P(t+k \mid t)$ in Lindley's equation [80] recursively for $1 \leq k < l$, we can compute prediction $Q^P_i(t + k + 1|t)$ of the primary queue size $Q^P_i(t + k)$ based on the information available at time t as

$$Q_i^P(t+k+1|t) = \left\lceil Q_i^P(t+k|t) + \hat{A}_i^P(t+k|t) - c_i x_i^P(t+k|t) \right\rceil^+, \quad i = 1,..., n \tag{86}$$

Note that $Q^P{}_i(t + k + 1|t)$ is an implicit function of the allocations $x^P{}_i(t + k|t)$, $k = 0, 1, \ldots, l$. We tune $\{x_i^P(t+k|t)\}_{k=0}^{l}$ to minimize the primary queue lengths. Define

$$\overline{N}_i(t+k|t) = \frac{Q_i^P(t+k|t)}{c_i}, \quad i = 1,..., n \tag{87}$$

Then the pedicted queue lengths are minimized when we take

$$x_i^P(t+k|t) = \begin{cases} N_i, & \text{if } \overline{N}_i(t+k|t) \geq N_i \\ \lceil \overline{N}_i(t+k|t) \rceil, & \text{otherwise} \end{cases}, \quad i = 1,..., n \tag{88}$$

This is important to recognize that computation of $\{x_i^P(t+k|t)\}_{k=0}^{l}$ is recursive in k. First we compute $x^P{}_i(t|t)$, which is essentially same as in (77). This value allows us to calculate $Q^P{}_i(t + 1|t)$ in (86), which then gives $x^P{}_i(t + 1|t)$ using (88). Now using (87), we can find $Q^P{}_i(t + 2|t)$ to get $x^P{}_i(t + 2|t)$ using (88), and so on.

Once the primary allocations are found, we can make secondary allocations using the spare capacity. Computation of the secondary allocations involves computing the vector $\mathbf{X}(t)$ defined (in steps) as follows:

$$\mathbf{x}_j^S(t+k|t) = \begin{bmatrix} x_{1j}^S(t+k|t) \\ : \\ x_{nj}^S(t+k|t) \end{bmatrix}, \mathbf{x}_k(t) = \begin{bmatrix} \mathbf{x}_1^S(t+k|t)^{\mathbf{T}} \\ : \\ \mathbf{x}_m^S(t+k|t)^{\mathbf{T}} \end{bmatrix}, \mathbf{X}(t) = \begin{bmatrix} \mathbf{x}_0(t)^{\mathbf{T}} \\ : \\ \mathbf{x}_l(t)^{\mathbf{T}} \end{bmatrix} \tag{89}$$

If we write $\mathbf{c} = [c_1, c_2, \ldots, c_n]^{\mathbf{T}}$, then note that

$$\sum_{i=1}^{n} c_i x_{ij}^S(t+k|t) = \mathbf{c}^{\mathbf{T}} \mathbf{x}_j^S(t+k|t) \tag{90}$$

We compute $\mathbf{X}(t)$ in such way that the largest secondary queue size over the prediction horizon l is minimized. To see the details, first

apply Lindley's equation [80] recursively for $1 \le k < l$ to obtain the following prediction of $Q^S_j(t + k + 1|t)$:

$$Q_j^S(t+k+1|t) = \left\lceil Q_j^S(t+k|t) + \hat{A}_j^S(t+k|t) - \mathbf{c}^{\mathrm{T}}\mathbf{x}_j^S(t+k|t) \right\rceil^+, \quad j = 1,...,m, k = 1,...,l \tag{91}$$

From (91), $Q^S_j(t + k|t)$ is a function of $\mathbf{X}(t)$. However, we don't show that explicitly to simplify the notation. Let us define $\mathbf{J} = \{1, 2, \ldots, m\}$ and $\mathbf{K} = \{1, 2, \ldots, l\}$, and

$$g\{\mathbf{X}(t)\} = \max_{k \in \mathbf{K},\, j \in \mathbf{J}} Q_j^S(t+k|t) \tag{92}$$

so that $g\{\mathbf{X}(t)\}$ denotes the largest size a secondary queue attainED over the prediction horizon l. Let us denote $\mathbf{I} = \{1, 2, \ldots, n\}$. Then we propose to compute $\mathbf{X}(t)$ by solving the following optimization problem:

$$\begin{aligned} &\text{minimize} && g\{\mathbf{X}(t)\} \\ &\text{subject to} && x_{ij}^S(t+k|t) \in \mathbf{Z}^+, \quad i \in \mathbf{I}, \quad j \in \mathbf{J}, \quad k \in \mathbf{K}, \\ & && x_i^P(t+k|k) + \sum_{j=1}^{m} x_{ij}^S(t+k|t) \le N_i, \; i \in \mathbf{I}. \end{aligned} \tag{93}$$

Solution Methodology

In this section we show that (93) is equivalent to integer linear program. The first step in this derivation is to write (93) as

$$\begin{aligned} &\text{minimize} && \omega \\ &\text{subject to} && \omega \ge g\{\mathbf{X}(t)\}, \\ & && x_{ij}^S(t+k|t) \in \mathbf{Z}^+, \quad i \in \mathbf{I}, \quad j \in \mathbf{J}, \quad k \in \mathbf{K}, \\ & && x_i^P(t+k|k) + \sum_{j=1}^{m} x_{ij}^S(t+k|t) \le N_i, \; i \in \mathbf{I}. \end{aligned} \tag{94}$$

Next we show that the constraint $\omega \ge g\{\mathbf{X}(t)\}$ is equivalent to a set of linear inequalities in $\mathbf{X}(t)$.

Lemma 1. Let us define

$$\gamma_j(t,p) = \hat{A}_j^S(t+p\,|\,t) + \mathbf{c}^{\mathbf{T}}\mathbf{x}_j^S(t+p\,|\,t) \tag{95}$$

Then $\omega \geq Q_j^S(t+k\,|\,t)$ if and only if all inequalities

$$\omega \geq 0 \tag{96}$$

$$\omega \geq \gamma_j(t,k-1) \tag{97}$$

$$\omega \geq \gamma_j(t,k-2) + \gamma_j(t,k-1) \tag{98}$$

:

$$\omega \geq \gamma_j(t,1) + ... + \gamma_j(t,k-2) + \gamma_j(t,k-1) \tag{99}$$

$$\omega \geq Q_j^S(t) + \gamma_j(t,0) + ... + \gamma_j(t,k-2) + \gamma_j(t,k-1) \tag{100}$$

hold.

Proof: From (70) and (91), the inequality

$$\omega \geq Q_j^S(t+k\,|\,t) \tag{101}$$

holds if and only if both the inequalities

$$\omega \geq 0 \tag{102}$$

$$\omega \geq Q_j^S(t+k-1) + \gamma_j(t,k-1) \tag{103}$$

hold. Using (70), (91) and (95), we have

$$Q_j^S(t+k-1) = \max\left\{0, Q_j^S(t+k-2) + \gamma_j(t,k-2)\right\}, \quad j=1,...,m,\ k=1,...,l \tag{104}$$

From (104) we get

$$\begin{aligned} &Q_j^S(t+k-1) + \gamma_j(t,k-1) = \\ &= \max\left\{\gamma_j(t,k-1), Q_j^S(t+k-2) + \gamma_j(t,k-2) + \gamma_j(t,k-1)\right\}, \quad j=1,...,m,\ k=1,...,l \end{aligned} \tag{105}$$

Thus, (103) is equivalent to both the inequalities

$$\omega \geq \gamma_j(t, k-1) \tag{106}$$

$$\omega \geq Q_j^S(t+k-2) + \gamma_j(t, k-2) + \gamma_j(t, k-1) \tag{107}$$

By continuing like above we get the set of inequalities (96) – (100).

□

Let us write (96) – (100) compactly as

$$\omega \geq 0$$

$$\omega \geq \sum_{i=p}^{k-1} \gamma_j(t, i), \quad \forall p \in \{1, 2, ..., k-1\} \tag{108}$$

$$\omega \geq Q_j^S(t) + \sum_{i=0}^{k-1} \gamma_j(t, i)$$

Using (92) we see that the constraint $\omega \geq g\{\mathbf{X}(t)\}$ in (109) is equivalent to the set of inequalities given by

$$\omega \geq 0$$

$$\omega \geq \sum_{i=p}^{k-1} \gamma_j(t, i), \quad \forall p \in \{1, 2, ..., k-1\}, \quad \forall k \in \mathbf{K}, \quad \forall j \in \mathbf{J} \tag{109}$$

$$\omega \geq Q_j^S(t) + \sum_{i=0}^{k-1} \gamma_j(t, i)$$

By replacing $\omega \geq g\{\mathbf{X}(t)\}$ by (109) in (94) we reduce (94) into equivalent integer linear problem, which can be solved efficiently using the branch and bound algorithm.

4 Algorithm Performance

In this section we present the summary of algorithm performance based on results of simulations obtained using OPNET development tool [112]. The simulation model consists of n = 3 primary eNBs and m = 1, 3 secondary eNBs connected to the EPC using 1Mbits/s IP links.

In the model, the CNM functionalities are implemented in EPC. The primary eNBs operate on fixed non-overlapping licensed spectrum

bands b_1 = 5MHz, b_2 = 10MHz, b_3 = 20 MHz. The radio model of the network has been developed according to the ITU-T Recommendation M.1225. Other simulation parameters have been set in coherence with the requirements of the LTE specifications [167, 169] (the simulation parameters of the network model are listed in Table 15).

The user traffic in simulations is made up of three most common network applications: VoIP, video and HTTP. The following models have been used to simulate voice, video and web users:

- The VoIP services model is ON-OFF model with exponentially distributed ON-OFF periods. The mean duration of ON and OFF periods are 0.65s and 0.352s, respectively. The VoIP traffic is generated by using the G.723.1 (12.2 Kbps) codec with a voice payload size 40 bytes and a voice payload interval 30 ms [167].
- Video services are simulated using a high resolution video model with a constant frame size equal 6250 bytes and exponentially distributed frame inter-arrival intervals (with mean equal 0.5s) [167].
- Web users in simulations are HTTP1.1 users generating pages or images with exponential page inter-arrival intervals (mean equal 60sec). It is assumed that one page consists of one object, whereas one image consists of five objects. The object size is constant and equal 1000 bytes [167].

In the standard LTE system the time axis is partitioned in time intervals of the length 1 ms [169]. However, in cognitive LTE network the slot duration should be much longer, considering that within one time slot i) the secondary eNBs should be able to re-adjust to the allocated frequency bands; ii) the network should establish reliable data transmission between the eNBs and the users (that is, the slot duration should be comparable to the mean duration of user sessions in the network). Based on these considerations, the slot duration in simulation model is set to be T_s = 1 sec.

Table 15. Simulation Parameters of the Network Model

Parameter		*Value*
Radio Network Model:	Pass loss	$L=40\log_{10}R+30\log_{10}f+49$, R – distance (km), f – carrier frequency (Hz)
	Shadow fading	Log-normal shadow fading with a st. dev. of 10/12 dB for outdoor/indoor users
	Penetration loss	The average building penetration loss is 12 dB with a st. dev. of 8 dB
	Multipath fading	SCM, Suburban macro
	Cell radius	1 km
	UE velocity	0 km/s
	Tx/Rx antenna gain	10 dBi (pedestrian), 2 dBi (indoor)
	Rx antenna gain	10 dBi (pedestrian), 2 dBi (indoor)
	Rx noise figure	5 dB
	Thermal noise density	–174 dBm/Hz
	Cable/connector/combiner losses	2 dB
PHY profile:	Operation mode	FDD
	Cyclic Prefix Type	Normal (7 Symbols per Slot)
	EPC Bearer Definitions	348kbit/s (Non-GBR)
	Carrier frequency	2GHz
	Subcarrier spacing	15kHz

Admission Control Parameters:	PDCCH symbols per subframe	3
	UL Loading Factor	1
	DL Loading Factor	1
	Inactive Bearer Timeout	20 sec
BSR Parameters:	Periodic Timer	5 subframes
	Retransmission Timer	2560 subframes
L1/L2 Control Parameters:	Reserved Size	2 RBs
	Cyclic Shifts	6
	Starting RBP for Format 1 messages	0
	Allocation Periodicity	5 subframes
RACH Parameters:	Number of Preambles	64
	Preamble Format	Format 0 (1-subframe long)
	Number of RA Resources per Frame	4
	Preamble Retransmission Limit	5 subframes
	RA Response Timer	5 subframes
	Contention Resolution Timer	40 subframes
HARQ Parameters:	Max Number of Retransmissions	3 (UL&DL)
	HARQ Retransmission Timer	8 subframes (UL&DL)

	Maximal Number of HARQ processes	8 per UE (UL&DL)

Considering that the acceptable prediction error should not exceed 0.1%, the length of the prediction window in Algorithm 2 has been set to be equal l = 5 time slots. This value was set after conducting a number of simulations with traffic sources listed above. Figure 59 summarizes the prediction performance as a function of traffic peakedness Z for l = 2, 4, …, 10 time slots. We use mean absolute percentage error [172]

$$MAPE = \frac{100\%}{T}\sum_{t=1}^{T}\left|\frac{A(t)-\hat{A}(t)}{A(t)}\right| \tag{110}$$

to measure the accuracy of prediction, where $A(t)$ and $\hat{A}(t)$ are the actual and predicted arrived traffic at time slot t, respectively; T is the length of simulation (in time slots).

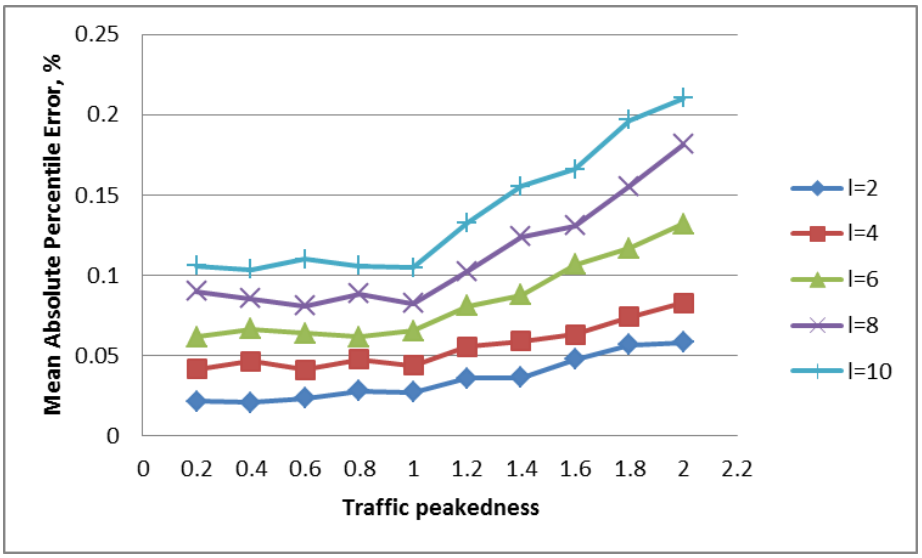


Fig. 59. Plot of *MAPE* as a function of Z for different values of l

Traffic peakedness Z is defined as the variance to mean ratio of the traffic distribution $A(t)$ [168], i.e.

$$Z = \frac{Var\{A(t)\}}{Mean\{A(t)\}} \quad (111)$$

If $Z < 1$ then the traffic is said to be smooth. Otherwise (if $Z > 1$), the traffic is peaked (bursty) and random [168].

To facilitate a fair and comprehensive analysis, we compare the performance of the proposed algorithm with the performance of two most relevant spectrum access techniques applicable to the considered system model. These techniques are:

1. cognitive radio resource management scheme for improving the LTE efficiency described in [170], and
2. dynamic bandwidth access scheme through pricing modeling described in [171].

In the first scheme the spectrum is assumed to be discrete. The total available bandwidth is divided into a number of sub-carriers. The sub-carriers are assigned to eNBs users to maximize the aggregated logarithmic utility given as a function of the bit rate at eNBs. Within the time slot one sub-carrier can be assigned to at most one eNB [170].

In the second scheme a number of PBs provide the wireless access to a number of SBs based on their utility function. The user utility is represented by the function depending on three parameters: 1) the amount of bandwidth which PB is willing to share with SB, 2) the signal to interference ratio of the wireless channel between the SB and PB and 3) the offered price for the bandwidth unit [171].

Here and later in the paper we use the following notation to differentiate between different algorithms:

- RBA (or rate based allocation) for the scheme described in [170];
- PBA (or price based allocation) for the scheme described in [171];
- QBA 1 (or queue size based allocation, Algorithm 1) proposed in this paper.
- QBA 2 (or queue size based allocation, Algorithm 2) with $l = 5$ proposed in this paper.

To evaluate performance of the proposed algorithms, a number of scenarios have been simulated with varying load and traffic peakedness and with different number of secondary eNBs. Description of these scenarios is summarized in Table 16. The traffic sources used in simulations have been listed in the beginning of this section.

Table 16. Simulated Scenarios

Scenario #	*n*	*m*	*Z*	*Users per eNB*
Scenario 1	3	1	0.6 (smooth traffic)	10 (low) ÷ 100 (high load)
Scenario 2	3	3	0.6 (smooth traffic)	10 (low) ÷ 100 (high load)
Scenario 3	3	1	0.2 (smooth) ÷ 2 (bursty traffic)	100 (high load)
Scenario 4	3	3	0.2 (smooth) ÷ 2 (bursty traffic)	100 (high load)

Performance of the network (evaluated in terms of mean packet end-to-end delay and loss) for different scenarios is shown on Figures 60 - 65. In particular, the delay and loss for the users of primary eNBs are plotted on Figures 60, 61. Note that the packet delay and loss in PBs do not depend on the number of SBs in the network. These results demostrate that:

- RBA shows the worst performance for the users of primary eNBs in all simulated scenarios. Service performance of RBA highly depends on the load in eNBs (packet delay and loss increase steeply in Scenarios 1 and 2). The impact of the traffic bursiness on service performance of RBA is much less.
- PBA has better performance than RBA. PBA is highly influenced by the traffic burstiness and in much less degree by the load in eNBs.
- Both QBA1 and QBA2 outperform the other schemes in all simulatied scenarios. QBA1 shows slightly better performance than QBA2 in scenarios with smooth traffic (Scenarios 1 and 2). QBA2 outperforms QBA1 when $Z \geq 1$, i.e. with bursty random traffic.

Figures 62 - 65 demonstrate the network performance for the users of secondary eNBs. As expected, the packet delay and loss for the users of SBs increases if we increase the number of SBs. Results also show that:

- Performance of RBA for the users of secondary eNBs is almost the same as for the users of primary eNBs in all simulated scenarios.
- PBA show the worst performance than all other schemes.

- Both QBA1 and QBA2 outperform the other schemes. QBA2 shows somewhat better performance than QBA1 in all simulated scenarios (i.e. regarless of the load and the traffic peakedness in eNBs).

Based on above observations we can summarize the performance of proposed resource allocation algorithms as follows. Both QBA1 and QBA2 are highly effective in reducing delay and loss for the users of secondary and primary eNBs. For primary eNBs, QBA1 shows slightly better performance that QBA2 with smooth traffic (Z = 0.6). With bursty traffic ($Z \geq 1$) QBA2 outperforms QBA1. For secondary eNBs, QBA2 shows better performance than QBA1 in all scenarios regarless of the load and traffic burstiness in the network.

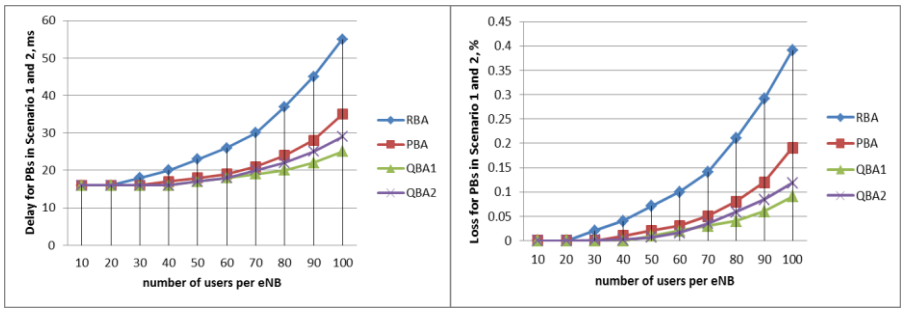


Fig. 60. Mean packet end-to-end delay and loss for users of PBs in scenarios 1 and 2

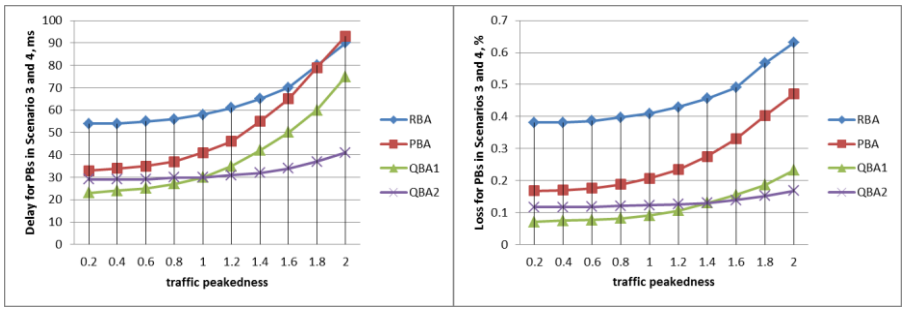


Fig. 61. Mean packet end-to-end delay and loss for users of PBs in scenarios 3 and 4

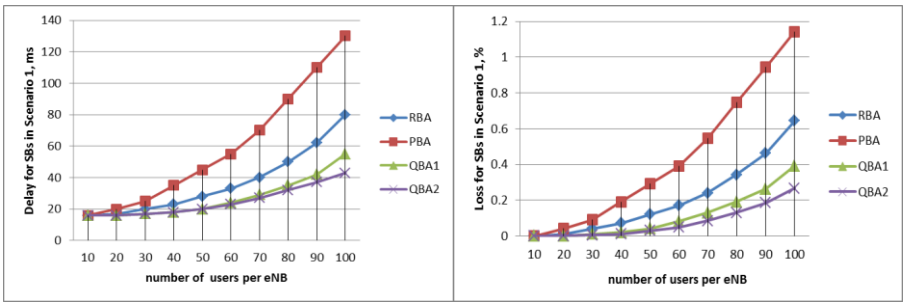


Fig. 62. Mean packet end-to-end delay and loss for users of SBs in scenario 1

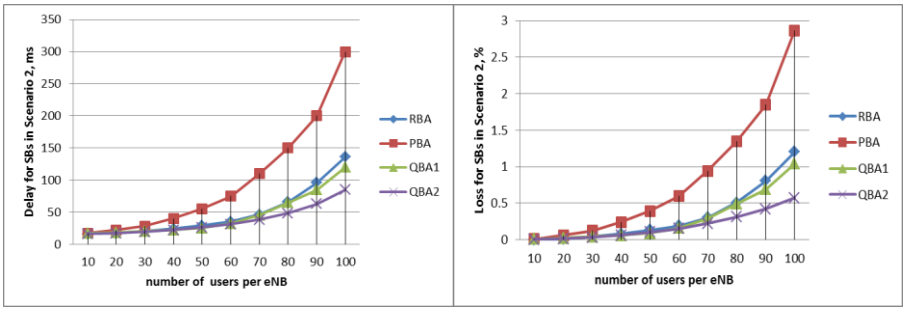


Fig. 63. Mean packet end-to-end delay and loss for users of SBs in scenario 2

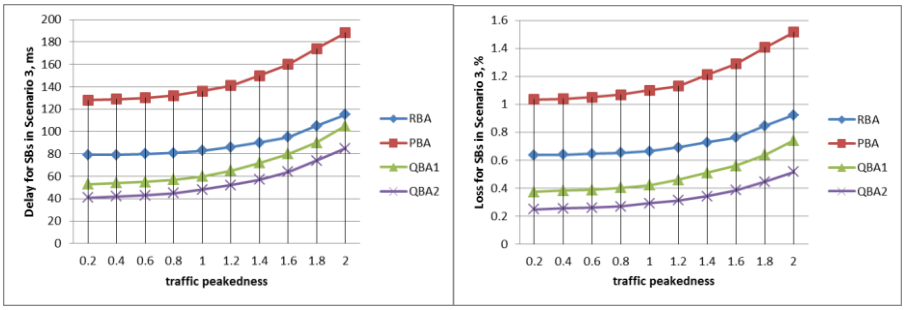


Fig. 64. Mean packet end-to-end delay and loss for users of SBs in scenario 3

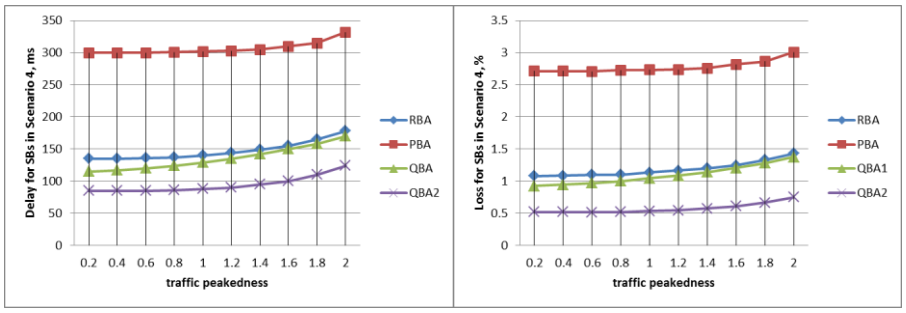


Fig. 65. Mean end-to-end delay and packet loss for users of SBs in scenario 4

CHAPTER 8: Delay Aware Resource Allocation for Secondary Users in Cognitive LTE Network

In this chapter a resource allocation technique for a cognitive LTE network in Scenario 3 is presented. Description of the network deployment scenario, as well as previous research on resource allocation for LTE-based CRN in Scenario 3 had already been summarized in Overview of this thesis. Here we formulate the optimization problem for resource allocation, derive the corresponding DSA algorithm, and summaries the algorithm performance based on simulation model developed in OPNET environment [112]. The corresponding paper will appear in Proceedings of IEEE MASS, October 2014.

1 Introduction

In this paper we consider a problem of dynamic spectrum access (DSA) in the Third Generation Partnership Project (3GPP) long-term evolution (LTE) cognitive radio network (CRN) architecture where the wireless access is provided to the primary (licensed) and secondary (unlicensed) users according to some predetermined policy. Within a CRN the primary users (PUs) get the ultimate prioritized access to licensed spectrum bands, whereas the secondary users (SUs) are served on the best-effort (non-prioritized) basis.

Unlike the other related techniques where the licensed spectrum bands have been assigned to SUs based on external network characteristics (such as signal to noise ratio, interference, traffic load, bit rate, throughput, etc.), we focus on specific design features of LTE radio interface associated with the scheduling process and the limited control channel capacity of the LTE system. In particular, we investigate the reasons limiting the capacity of the system on physical (PHY) and medium access control (MAC) layers, and find the relation between the scheduling delay (which comprises the largest part of the packet end-to-end delay in LTE network) and the number of users in the system.

Based on these results, we derive a simple technique to assign the spectrum for SUs without violating the QoS requirements of PU, and implement this algorithm in LTE-based CRN. We verify the consistent performance of the proposed algorithm by comparing its performance with the performance of other relevant DSA techniques.

This Chapter is organized as follows. In section 2 we formulate the optimization problem for resource allocation and derive the relation between the sceduling delay and the number of users in LTE system. In section 3 we present the proposed algorithm for dynamic spectrum access in LTE-based CRN. The algorithm performance is evaluated in section 4.

2 Optimization Problem

2.1 Problem Formulation

Consider a typical cognitive radio network (CRN) model based on LTE standard network illustrated on Figure 6. It comprises a core networking part and m service providers (SPs) offering the wireless services via a set of respective evolved NodeBs (eNBs) numbered eNB_1, ..., eNB_m. Similar to the standard LTE system, considered network model operates on a slotted time basis: the time axis in the model is partitioned into discrete mutually disjoint intervals of length T_s $\{[tT_s, (t+1)T_s]\}$, $t = 0, 1, 2, \ldots$, with T_s denoting the subframe (in LTE T_s = 1ms), and t denoting the integer values index of T_s.

A CRN provides the wireless access to N primary users (PUs) and X secondary users (SUs). PUs are the licensed network users who pay some prize to their SPs for accessing the wireless services. SUs are unlicensed network users who can access the wireless services for free.

Each eNB operates on a fixed licensed spectrum band and serves a number of PUs, randomly arriving to (and leaving) the network with mean arrival rate λ_{PU} (and mean departure rate μ_{PU}). We denote the spectrum band of the eNB_i by b_i, and the instantaneous number of PUs in eNB_i by n_i. The eNBs can also provide the wireless access to SUs, randomly arriving to (and leaving) the network with mean arrival rate λ_{SU} (and mean departure rate μ_{SU}). Within a CRN the PUs get prioritized access to the spectrum band of SP (eNB) they have arrived

to. The SUs are served on the best-effort (non-prioritized) basis and can be redirected to the other SP (eNB).

We assume that:

1. one SU can connect to at most one eNB;
2. the mean inter-arrival times of PUs and SUs (and the mean inter-departure times of PUs and SUs) are much greater than the subframe duration, i.e. $1/\lambda_{PU} >> T_s$, $1/\lambda_{SU} >> T_s$, $\mu_{PU} >> T_s$, $\mu_{PU} >> T_s$, which is quite reasonable because in real network the mean inter-arrival times (and the mean inter-departure times) of the users are usually much greater than T_s = 1ms;
3. the spectrum bands of eNBs are non-overlapping.

The main goal of the considered CRN is two-folded. Firstly, it should provide the wireless access for SUs. Secondly, it should maintain some QoS levels of PUs. Considering, that the QoS for most of the user applications is measured in terms of the packet end-to-end delay, this goal can be reformulated as follows. Maximize the number of SUs in m eNBs given that the packet end-to-end delay in eNBs does not exceed some predefined limits.

Let x_i be the number of SUs assigned to eNB_i. Let D_i be the average packet end-to-end delay (i.e. the time it takes for a packet to travel from the user through the network to the server, and back) in eNB_i. Let $D^P{}_i$ be the maximum value of the packet end-to-end delay acceptable for eNB_i. Then the corresponding optimization problem for the CRN model illustrated on Figure 6 is given by

$$\max \ f(x) = \sum_{i=1}^{m} x_i$$

$$\text{subject to :}$$

$$\begin{aligned} & D_i \le D_i^P, \ 1 \le i \le m \\ & \sum_{i=1}^{m} x_i \le X \\ & x_i \ge 0, \ x_i \text{ is integer}, \ 1 \le i \le m \end{aligned} \tag{112}$$

where $x = [x_1,\ldots,x_m]^T$ is the vector of non-negative numbers.

To solve the problem (112), we should find the relation between the number of users and the packet end-to-end delay in LTE system. According to [93], the packet end-to-end delay in LTE system is equal:

$$D = D^t + D^b + D^p + D^{eNB} + D^{UE} + D^c + D^{HARQ} + D^{PS} \quad (113)$$

where D^t, D^b, D^p – are the total (uplink and downlink) packet transmission, buffering and propagation delays between the UE and the eNB, respectively; D^{HARQ} – the total (uplink and downlink) packet delay due to hybrid automatic repeat request (HARQ) retransmissions; D^{PS} – the uplink delay due to packet scheduling; D^{eNB} and D^{UE} – processing delays of eNB and the user equipment (UE); D^c – the total (uplink and downlink) packet delay in core network. Figure 66 shows the typical values of different delay components in LTE network [93].

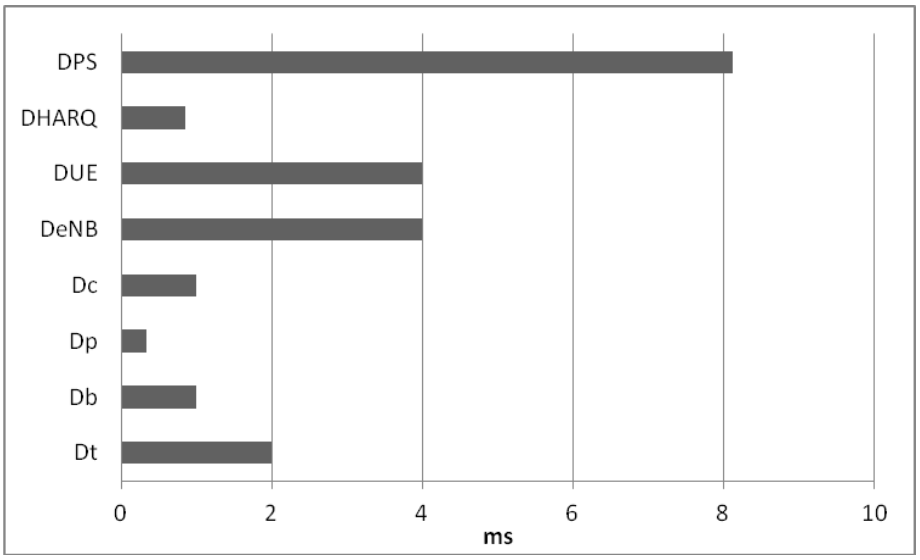


Fig. 66. The typical values of different delay components

Because of the small size of the subframe (the subframe duration in LTE is equal T_s = 1 ms), the transmission and the buffering delay components D^t and D^b are very small in LTE system (D^t = 2 ms, D^b = 1 ms). The propagation delay D^p and the delay in core network D^c depend on the distance between the UE and the eNB, and the eNB and the server, relatively, and usually in orders of 1 ms (in case if the distance between the UE and the server does not exceed 1000 km). The

components D^{eNB} and D^{UE} depend on the processing capabilities of the equipment (typically around 5 ms) [93].

The delay due to HARQ retransmissions depends on the wireless channel quality. The average value of D^{HARQ} can be estimated using [93]:

$$D^{HARQ} = P^{RTX} \cdot T^{HARQ} \tag{114}$$

where P^{RTX} is the probability of the HARQ retransmission; T^{HARQ} – the time interval between the transmission and respective HARQ retransmission (in LTE standard T^{HARQ} = 8 ms). It follows from (114) that delay due to HARQ retransmissions never exceeds 8 ms (in general $D^{HARQ} < 4$ ms) [93].

The delay component D^{PS} is associated with the scheduling process in LTE system. The packet scheduling allows provide the guaranteed wireless channels and maintain some QoS levels for the prioritized network users. On the other hand, the scheduling procedure itself introduces an additional delay for all types of the network users (prioritized and non-prioritized). In LTE the delay due to scheduling is relatively large (in general $D^{PS} \geq 8$ ms) and constitutes the biggest part (≈36%) of the packet end-to-end delay. Unlike the other delay components, the scheduling delay depends on the number of users in eNB [93, 94].

We now return to the primary optimization problem given by (112). Combining (112) and (113), we get the following more detailed formulation of the primary problem:

$$\max \ f(x) = \sum_{i=1}^{m} x_i$$

$$\text{subject to :}$$

$$D_i^t + D_i^b + D_i^p + D_i^{eNB} + D_i^{UE} + D_i^c + D_i^{HARQ} + D_i^{PS} \leq D_i^P, \ 1 \leq i \leq m \tag{115}$$

$$\sum_{i=1}^{m} x_i \leq X$$

$$x_i \geq 0, \ x_i \text{ is integer}, \ 1 \leq i \leq m$$

In (115) D_i^t, D_i^b, D_i^p, D_i^c, D_i^{eNB}, D_i^{UE}, D_i^{HARQ} do not depend on x_i and

can be directly measured at eNB$_i$ (D_i^t, D_i^b, D_i^p, D_i^c, D_i^{eNB}, D_i^{UE} are constants, D_i^{HARQ} depends on the number of HARQ retransmissions in eNB$_i$). The only delay component which depends on x_i and should be restricted in optimization problem is D_i^{PS}.

For convenience let us define

$$\Delta D_i^P := D_i^P - D_i^b - D_i^p - D_i^{eNB} - D_i^{UE} - D_i^c - D_i^{HARQ} \tag{116}$$

Expression (116) is equivalent to the maximum acceptable scheduling delay. Using (116), the optimization problem (115) can be redefined as follows

$$\max \ f(x) = \sum_{i=1}^{m} x_i$$

$$\text{subject to :}$$

$$D_i^{PS} \le \Delta D_i^P, \ 1 \le i \le m \tag{117}$$

$$\sum_{i=1}^{m} x_i \le X$$

$$x_i \ge 0, \ x_i \text{ is integer}, \ 1 \le i \le m$$

Clearly, to solve (117), it is important to find the relation between the scheduling delay and the number of users in LTE system.

2.2 Scheduling Delay and the Number of Users in eNB

In this subsection we find the relation between the scheduling delay and the number of users in eNB. We start from the brief description of the scheduling process in LTE system (more detailed description of the scheduling process can be found for instance in [158]).

In LTE resources are allocated to user equipments (UEs) for uplink and downlink data transmission in terms of RBs. Thus, one UE can be allocated only the integer number of RBs in frequency domain, and these RBs do not have to be adjacent to each other. Resource allocation (scheduling) is carried by the MAC layer packet scheduler in the eNB both for uplink and downlink transmissions [157].The scheduling decisions are made based on the quality of service (QoS), user priority, fairness and instantaneous channel conditions.

The standard dynamic packet scheduling scheme can be described as follows [157]. Within one subframe with duration equal T_s = 1ms:

1. all active UEs generate the scheduling requests (SRs) and send them via the physical uplink control channel (PUCCH) to eNB using the format 1 messages.
2. the eNB receives the PUCCH information, decodes the PUCCH format 1 messages, allocates the resources and sends the scheduling grants (SGs) to UEs via the physical downlink control channel (PDCCH) using the downlink control information (DCI) format 1 messages. The duration of this procedure is equal T_{SR} = 4ms.
3. UEs receive the PDCCH information, decode the DCI format 1 messages, and transmit the uplink data via the physical uplink shared channel (PUSCH). This procedure takes exactly T_{SG} = 4ms.

Because of the finite capacity of the PDCCH and PUCCH, the scheduler is constrained in its freedom of how many users to address in a subframe [131]. Thus, if the number of scheduling requests (SRs) sent in one subframe is not more than the PDCCH/PUCCH capacity, all UEs generating SRs are scheduled and can transmit the uplink data. Otherwise, i.e. if the number of scheduling requests (SRs) sent in one subframe is more than the PDCCH/PUCCH capacity, the scheduling for some UEs generating SRs will be delayed for the next subframe [94, 131].

To estimate the scheduling delay D^{PS} consider a cell consisting of a number of UEs and a eNB. We assume that all UEs in the cell are active all of the time and the number of UEs generating SRs in one subframe N_{SR} is equal to the number of UEs in the cell. Let C^{CCH} be the control channel capacity, i.e. the number of UEs that can be scheduled in one subframe. If $N_{SR} \leq C^{CCH}$ then all UEs in time (8ms after sending the respective SR) [94]. The scheduling delay for all UEs in the cell in this case is equal:

$$if \quad N_{SR} \leq C^{CCH} \ then \ \ D^{PS} = T_{SR} + T_{SG} \tag{118}$$

If $N_{SR} > C^{CCH}$ then exactly C^{CCH} UEs are scheduled in time, while the left (N_{SR} - C^{CCH}) UEs are delayed for the next subframe [94]. The average scheduling delay for all UEs in the cell in this case is equal:

$$if\ N_{SR} > C^{CCH}\ then\ D^{PS} = T_{SR} + T_{SG} + T_s\left(1 - \frac{C^{CCH}}{N_{SR}}\right) \quad (119)$$

Combining (118) and (119) we get the expression for the average scheduling delay in LTE system:

$$D^{PS} = T_{SR} + T_{SG} + T_s\left\lceil 1 - \frac{C^{CCH}}{N_{SR}}\right\rceil^{+} \quad (120)$$

where $\lceil x \rceil^{+} = \max\{0, x\}$.

For the network model shown on Figure 6 the total number of users in eNB$_i$ is equal $n_i + x_i$. Then the number of SRs sent in one subframe is equal to the number of active users in eNB $N_{SR} = n_i + x_i$ by assumption.

In LTE, the value of C^{CCH} can be determined from the bandwidth of the respective eNB denoted via b_i. Recall that SRs are carried via the PUCCH using the PUCCH format 1 messages; SGs are carried via the PDCCH using the DCI format 1 messages. Thus, for eNB$_i$ the capacity of physical control channels in uplink direction is equal to the number of PUCCH sub-channels allocated for PUCCH format 1 messages denoted via $N_i^{PUCCH\_1}$. The PDCCH capacity of eNB$_i$ is equal to the number of control channel elements N_i^{CCE} allocated for the DCI format 1 messages [94]. Then the PDCCH/PUCCH capacity of eNB$_i$ is equal:

$$C_i^{CCH} = \min\{N_i^{PUCCH\_1}, N_i^{CCE}\} \quad (121)$$

The average scheduling delay in eNB$_i$ is equal:

$$D_i^{PS} = T_{SR} + T_{SG} + T_s\left\lceil 1 - \frac{C_i^{CCH}}{n_i + x_i}\right\rceil^{+} \quad (122)$$

And the problem (117) will take the form:

$$\max \ f(x) = \sum_{i=1}^{m} x_i$$

subject to :

$$g_i(x) = \left\lceil 1 - \frac{C_i^{CCH}}{n_i + x_i} \right\rceil^{+} - \frac{\Delta D_i^P - T_{SR} + T_{SG}}{T_s} \le 0, \ 1 \le i \le m \tag{123}$$

$$g_{m+i}(x) = x_i \ge 0, \ 1 \le i \le m$$

$$g_{2m+1}(x) = \sum_{i=1}^{m} x_i - X \le 0$$

x_i is integer, $1 \le i \le m$

3 DSA Algorithm for LTE-based CRN

In this section we present the example of the algorithm implementation in cognitive LTE-based network architecture. The objective of the algorithm is to assign the spectrum to the maximum possible number of SUs subject to certain delay constraints established in eNBs.

We assume that SPs operate on fixed non-overlapping licensed spectrum bands b_1, …, b_m, and the number of available control C_1^{CCH}, …, C_m^{CCH} remain constant. To track the number of PUs and SUs in the proposed CRN model we utilize the modified version of the standard LTE RACH procedure [158] described as follows. For initial access to the network (i.e. at arrival) the PU/SU generates a Primary/Secondary service Initiation Request (PIR/SIR) and sends it in the form of RA-preamble using the spectrum band of any eNB within CRN using the RACH procedure. When the PU/SU leaves the network, it generates a Primary/Secondary service Termination Request (PTR/STR) and sends it in the form of RA-preamble to respective eNB using the RACH procedure.

One of the primary assumptions of the network model was that the mean inter-arrival times of PUs and SUs (and the mean inter-departure times of PUs and SUs) are much greater than the subframe duration, i.e. $1/\lambda_{PU} >> T_s$, $1/\lambda_{SU} >> T_s$, $\mu_{PU} >> T_s$, $\mu_{SU} >> T_s$. Based on this assumption, we propose to make each subsequent spectrum allocation

within the time interval Δt which is less that mean inter-arrival times of PUs and SUs (and the mean inter-departure times of PUs and SUs), but greater than the subframe duration, i.e. $T_s < \Delta t < 1/\lambda_{PU}$, $T_s < \Delta t < 1/\lambda_{SU}$, $T_s < \Delta t < \mu_{PU}$, $T_s < \Delta t < \mu_{SU}$. This will allow decrease the amount of signaling without affecting the algorithm performance.

In CRN all PUs have the prioritized access to all spectrum bands/eNBs comprising the network and therefore they get an immediate access to any spectrum band/eNB. SUs can operate only on the spectrum bands/eNBs that have been allocated to them according to the algorithm that can be briefly described as follows (more detailed description of the algorithm is presented on Figure 67).

Within each time interval Δt:

- All SUs/PUs arriving to the network generate PIRs/SIRs and send them to any eNB within the network. All SUs/PUs leaving the network generate PTRs/STRs and send them to the eNB which they are leaving.
- The eNBs collect all received PIRs, PTRs, SIRs and STRs (we denote all PIRs, PTRs, SIRs and STRs received by the eNB_i via $PIR_i(t)$, $PTR_i(t)$, $SSR_i(t)$ and $STR_i(t)$, respectively), update the number of PUs and SUs, and send them to EPC.
- After receiving $PIR_i(t)$, $PTR_i(t)$, $SSR_i(t)$ and $STR_i(t)$ from all eNBs, the EPC finds the optimal solution to problem (12) given by $x^* = [x^*_1, \ldots, x^*_m]^T$. After this, the EPC redirects the SUs by sending the *index* and the *number* of admitted SUs to all eNBs within the network.
- At each eNB if $x^*_i - x_i(t) \geq 0$ then all SUs are granted the access. Otherwise, the eNB accepts x^*_i SUs and redirects $x_i(t) - x^*_i$ SUs to the admitting eNBs.
- After being located (i.e. accepted or redirected) in the network, the SUs get the wireless access to the spectrum bands of admitting eNBs.

Algorithm 1. DSA Algorithm for LTE-based CRN

At time *t*

PUs/SUs arriving to the cell send *PIR/SIR* to any eNB_i in {1, ..., *m*},
PUs/SUs leaving the cell send *PTR/STR* to eNB_i in {1, ..., *m*},

At all eNB_i in {1, ..., *m*}

1. **Update** $PIR_i(t)$, $PTR_i(t)$, $SIR_i(t)$, $STR_i(t)$
2. **Count** $n_i(t) := n_i(t-\Delta t) + PIR_i(t) - PTR_i(t)$,
   $x_i(t) := x_i(t-\Delta t) + SIR_i(t) - STR_i(t)$,
3. **Send** $n_i(t)$, $x_i(t)$ to EPC

At EPC

1. **Receive** $n_i(t)$, $x_i(t)$ from all eNB_i in {1, ..., *m*}
2. **Find** $x^* = [x^*_1, ..., x^*_m]T$
3. **For all** *i* in {1, ..., *m*}
   if $x_i(t) > x^*_i$
   then
   for all *j* in {*i*+1, ..., *m*}
   if $x_j(t) < x^*_j$
   then $x_i(t) := x_i(t) - min[x_i(t)-x^*_i, x^*_j- x_j(t)]$,
   $x_j(t) := x_j(t) + min[x_i(t)-x^*_i, x^*_j- x_j(t)]$,
   $index := j$,
   $number := min[x_i(t)-x^*_i, x^*_j- x_j(t)]$,
   send *index*, *number* to eNB_i
   else
   $index := i$,
   $number := x_i(t)$,
   send *index*, *number* to eNB_i

At all eNB_i in {1, ..., *m*}

1. **Receive** *index*, *number* from EPC
2. **Send** *index* to *number* SUs

At SUs

1. **Receive** *index* from eNB
2. **Connect to** eNB_{index}

Fig. 67. DSA Algorithm for LTE-based CRN

4 Performance Analysis

4.1 Simulation Model

In this subsection we describe the simulation model of the network shown on Figure 6. The model has been implemented based on the standard LTE-A platform using the OPNET simulation and development package [112]. The wireless networking part of the model consists of m = 7 SPs (eNBs) numbered eNB_1, …, eNB_7. The core networking part comprises the EPC and the server. The SPs operate on fixed non-overlapping licensed spectrum bands and the number of available control remain constant (the values of b_1, …, b_7 and C_1^{CCH}, …, C_7^{CCH} are given on Table 17). In the network $\Delta t = 1000 \times T_s$ = 1s. The radio model of the network has been developed according to the ITU-T Recommendation M.1225. Other simulation parameters are set in coherence with the requirements of the LTE specifications [131, 157, 158] (the simulation parameters of the network model are listed in Table 18).

Table 17. The Bandwidth and the Number of Control Channels in the Model

eNB ID	*Bandwidth*	*Center Frequency*	*Number of Control Channels*
eNB_1	b_1 = 5 MHz	2000 MHz	C_1^{CCH} = 20
eNB_2	b_2 = 5 MHz	2005 MHz	C_2^{CCH} = 20
eNB_3	b_3 = 5 MHz	2010 MHz	C_3^{CCH} = 20
eNB_4	b_4 = 5 MHz	2015 MHz	C_4^{CCH} = 20
eNB_5	b_5 = 10 MHz	2022.5 MHz	C_5^{CCH} = 41
eNB_6	b_6 = 10 MHz	2032.5 MHz	C_6^{CCH} = 41
eNB_7	b_7 = 20 MHz	2047.5 MHz	C_7^{CCH} = 84

Table 18. Simulation Parameters of the Model

Parameter		*Value*
Radio Network Model:	Pass loss	$L=40\log_{10}R+30\log_{10}f+49$, R – distance (km), f – carrier frequency (Hz)
	Shadow fading	Log-normal shadow fading with a st. dev. of 10/12 dB for outdoor/indoor users
	Penetration loss	The average building penetration loss is 12 dB with a st. dev. of 8 dB
	Multipath fading	SCM, Suburban macro
	UE velocity	0 km/s
	Tx/Rx antenna gain	10 dBi (pedestrian), 2 dBi (indoor)
	Rx antenna gain	10 dBi (pedestrian), 2 dBi (indoor)
	Rx noise figure	5 dB
	Thermal noise density	–174 dBm/Hz
	Cable/connector/combiner losses	2 dB
PHY profile:	Operation mode	FDD
	Cyclic Prefix Type	Normal (7 Symbols per Slot)
	EPC Bearer Definitions	348kbit/s (Non-GBR)
	Subcarrier spacing	15kHz
Admission Control Parameters:	PDCCH symbols per subframe	3
	UL Loading Factor	1

	DL Loading Factor	1
	Inactive Bearer Timeout	20 sec
BSR Parameters:	Periodic Timer	5 subframes
	Retransmission Timer	2560 subframes
L1/L2 Control Parameters:	Reserved Size	2 RBs
	Cyclic Shifts	6
	Starting RBP for Format 1 messages	0
	Allocation Periodicity	5 subframes
RA Parameters:	Number of Preambles	64
	Preamble Format	Format 0 (1-subframe long)
	Number of RA Resources per Frame	4
	Preamble Retransmission Limit	5 subframes
	RA Response Timer	5 subframes
	Contention Resolution Timer	40 subframes
HARQ Parameters:	Max Number of Retransmissions	3 (UL & DL)
	HARQ Retransmission Timer	8 subframes (UL & DL)
	Max Number of HARQ processes	8 per UE (UL & DL)

To facilitate fair and comprehensive simulative analysis, we compare the performance of the proposed algorithm with the performance of two

most relevant spectrum access techniques applicable to the considered system model. These techniques are cognitive radio resource management scheme for improving the LTE efficiency described in [170] and dynamic bandwidth access scheme through pricing modeling described in [171].

In the first scheme the spectrum is assumed to be discrete: the total available bandwidth is divided into a number of sub-carriers. The sub-carriers are assigned to the users to maximize the aggregated logarithmic user utility given as a function of the user bit rate. Within the time slot one sub-carrier can be assigned to at most one user [170].

In the second scheme a number of PUs provide the wireless access to a number of SUs based on their utility function. The user utility is represented by the function depending on three parameters: 1) the amount of bandwidth which PU is willing to share with SU, 2) the signal to interference ratio of the wireless channel between the SU and PU and 3) the offered price for the bandwidth unit [171].

Here and after we use the following notation to differentiate performance of different algorithms in simulations: RBA (or rate based allocation) for the scheme described in [170]; PBA (or price based allocation) for the scheme described in [171]; DBA (or delay based allocation) for the scheme proposed in this work.

All algorithms are simulated with identical LTE parameters and under identical network deployment scenarios (such as channel quality, traffic load, use behaviour, etc.).

The user traffic in simulations comprises three most frequently used network applications: VoIP, video and HTTP. The following models are used to simulate voice, video and web users:

- The VoIP services model is ON-OFF model with exponentially distributed ON-OFF periods. The mean duration of ON and OFF periods are 0.65s and 0.352s, respectively. The VoIP traffic is generated by using the G.723.1 (12.2 Kbps) codec with a voice payload size 40 bytes and a voice payload interval 30 ms [167].
- Video services are simulated using a high resolution video model with a constant frame size equal 6250 bytes and exponentially distributed frame inter-arrival intervals (with mean equal 0.5s) [167].

- Web users in simulations are HTTP1.1 users generating pages or images with exponential page inter-arrival intervals (mean equal 60sec). It is assumed that one page consists of one object, whereas one image consists of five objects. The object size is constant and equal 1000 bytes [167].

4.2 Simulations Results

Results below demonstrate the performance of different algorithms collected under different network deployment scenarios. The graphs on the Figures 68 – 72 show mean packet delay in the network with mean number of SUs in the network $X = 100 \div 1000$ UEs. For DBA we limit the maximal allowed delay to be $D_1^P = ... = D_7^P = 100$ ms for all eNBs. Figures 68, 69 show the performance for mean number of PUs n = 1000 UEs; Figures 70, 71 show the performance for mean number of PUs n = 2000 UEs.

Results demonstrate that PBA and DBA show better performance for PUs because of the prioritized access of PUs offered in the network, whereas the delay for PUs and SUs in RBA are almost the same (the spectrum resource in RBA are assigned based on user bit rate without prioritizing). Results also show that the performance of DBA is much better than performance of RBA and PBA both for PUs and SUs which is mainly explained by the fact that the main component of delay in LTE network is related to scheduling, and the algorithm restricts the number of SUs subject to delay constrains of PUs.

Performance of DBA can be better demonstrated using the graphs on Figure 72 showing the mean number of SUs served by CRN with maximal allowed delay $D_1^P = ... = D_7^P$ ranging from 0 to 100 ms and X = 3000 SUs. From this graphs it clearly follows how the number of SUs served by CRN is related to the delay constraints in DBA.

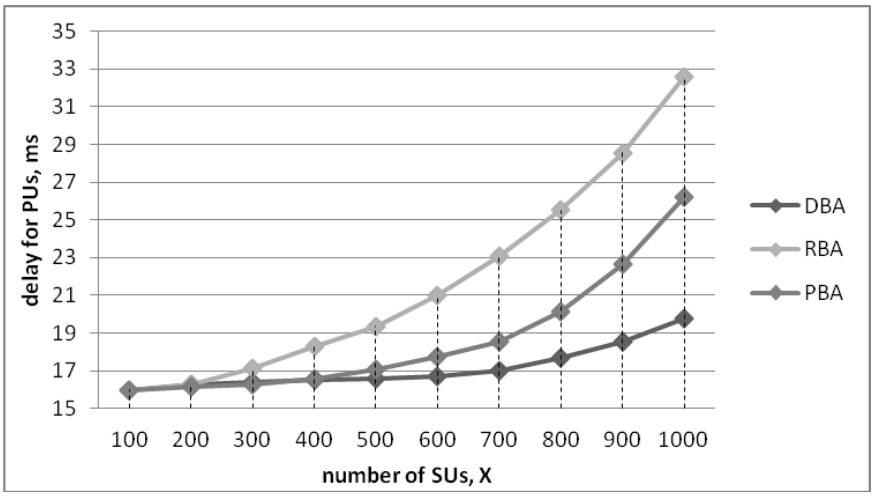


Fig. 68. Mean packet delay for PUs with n = 1000 UEs

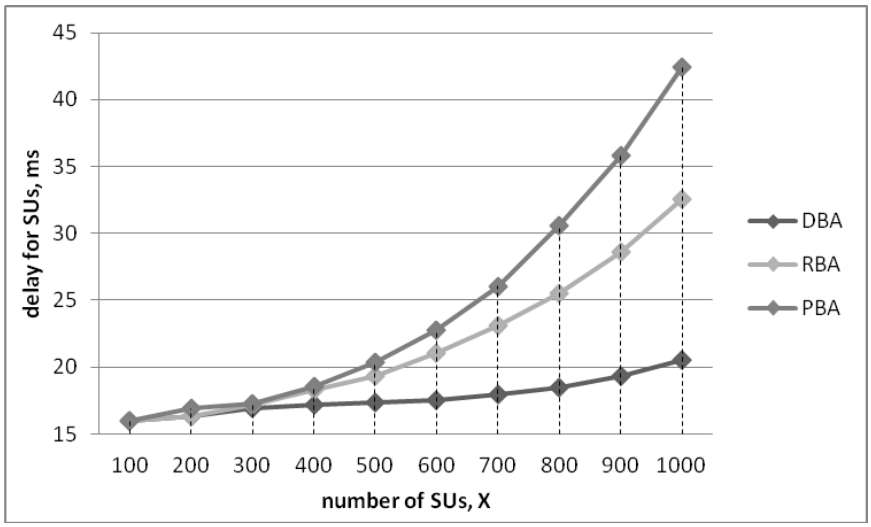


Fig. 69. Mean packet delay for SUs with n = 1000 UEs

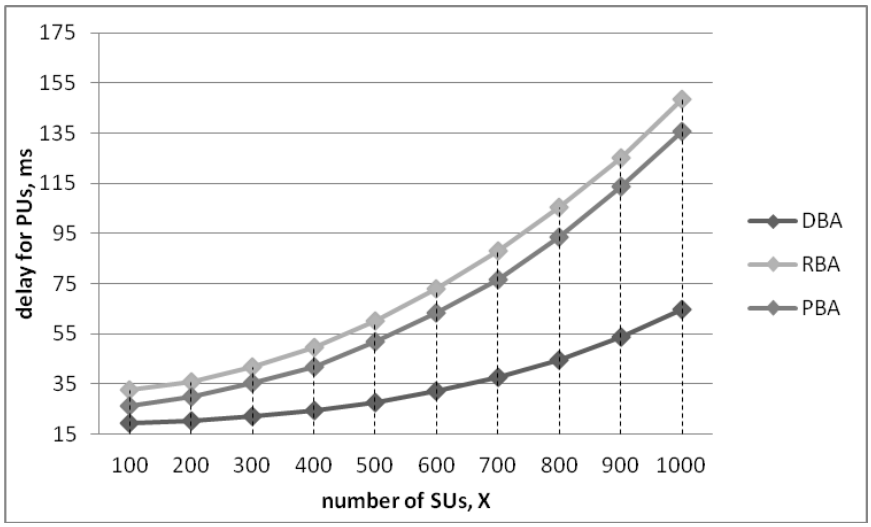


Fig. 70. Mean packet delay for PUs with n = 2000 UEs

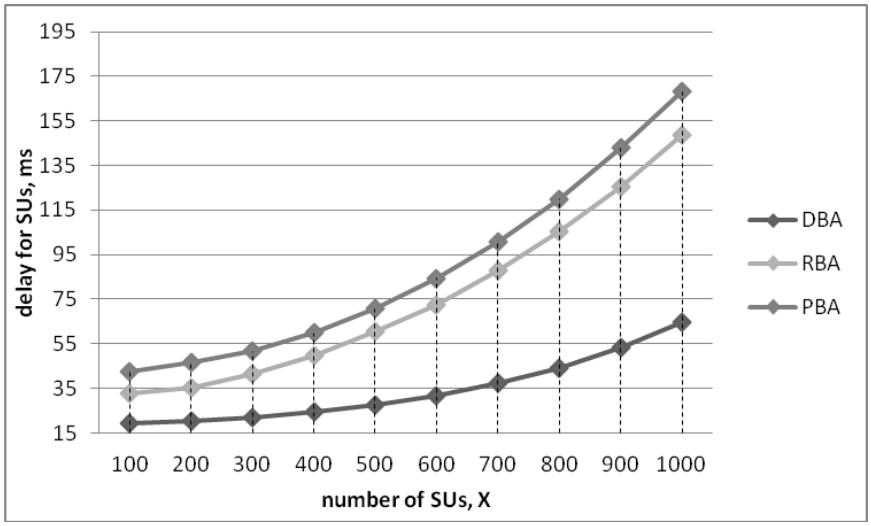


Fig. 71. Mean packet delay for SUs with n = 2000 UEs

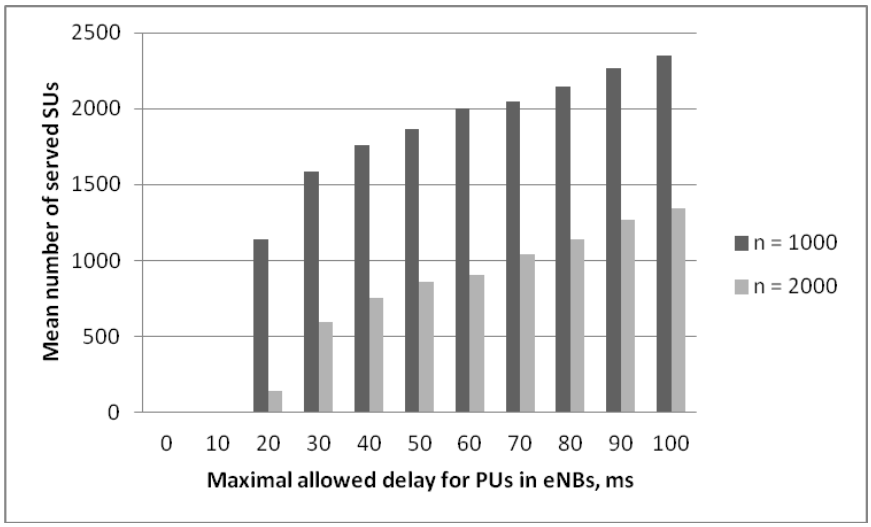


Fig. 72. Mean number of served SUs with X = 3000 UEs

CONCLUSIONS

This thesis contains a collection of various resource management techniques to provide increased spectrum utilization and enhanced end-to-end QoS for users of future wireless networks. Although we consider the application of these techniques to specific wireless network interfaces (Wi-Fi and LTE), most of them can be deployed in any OFDMA-based network. In this thesis we consider three different network deployment scenarios, and offer some general network architecture and resource allocation policy which can be implemented using any of the proposed algorithms to improve the overall capacity and service performance of the network.

The main advantage of the algorithms proposed in this thesis is possibility of pratical implementation of these algorithms upon existing wireless networking platforms. For instance, in most cases we use the arrival rate and the size of the queues at the nodes of the network as observable system parameters. In real Wi-Fi and LTE networks, these parameters are readily available at the MAC queues of user terminals and base stations in both uplink and downlink directions. After collecting the necessary parametric information, the nodes send the parametric values to a central processor (which can be implemented at a base station, EPC or a server) using IP links of high date rate. This way of information exchange between the network nodes and a central processor eliminates the need for additional control signaling over the wireless medium, and enables fast and practically error-free transmission of the control information used for resource allocation.

After receiving the parametric information, a central processor calculates the optimal resource allocation, and sends the results to the correponding nodes (via IP links). If (for some reason) the current observation(s) is not received, the last available data is used by a central processor to obtain an optimal solution. Thus, the proposed resource allocation approach is robust, since the dynamics of resource allocation (the interval between two consecutive resource allocations) is much faster than the dynamics of the change in parameteric values. Finally, most of the proposed methods (for traffic prediction and resouse allocation) have low or moderate computational complexity,

and therefore can be applied to large networks consisting of many nodes.

Bibliography

1. Cognitive Radio, Spectrum and Radio Resource Management. Working Group 6. White Paper. Wireless World Research Forum 2004.
2. P.Demestichas, L.Papadopoulou, V.Stavroulaki, M.Theologou, G.Vivier, G.Martinez, F.Galliano, "Wireless beyond 3G: Managing Services and Network Resources", IEEE Computer, Vol. 35, No. 8, Aug. 2002.
3. End to End Reconfigurability (E2R), IST-2003-507995 E2R, http://www.e2r.motlabs.com.
4. A. Nosratinia, T. Hunter, A. Hedayat, "Cooperative communication in wireless networks", IEEE Communications Magazine (2004). Vol. 42 (10), pp. 74-80.
5. P. Demestichas, N. Koutsouris, G. Koundourakis, K. Tsagkaris, A. Oikonomou, V. Stavroulaki, L. Papadopoulou, M. Theologou, G. Vivier, K.El-Khazen, "Management of networks and services in a composite radio context", IEEE Wireless Commun. Mag., Vol. 10, No. 4, Aug. 2003, pp. 44-51.
6. P. Demestichas, V. Stavroulaki, "Issues in introducing resource brokerage functionality in B3G, composite radio, environments", IEEE Wireless Communications Magazine, Vol. 11, No. 10, October 2004.
7. P. Demestichas, G. Vivier, K.El-Khazen, M. Theologou, "Evolution in wireless systems management concepts: from composite radio to reconfigurability", IEEE Communications Magazine, Vol. 42, No. 5, pp. 90-98, May 2004.
8. P.Demestichas, V.Stavroulaki, L.Papadopoulou, A.Vasilakos, M.Theologou, "Service configuration and distribution in composite radio environments", IEEE Transactions on Systems, Man and Cybernetics Journal, vol. 33, No. 4, pp. 69-81, Nov. 2003.
9. Software Defined Radio forum: www.sdrforum.org.
10. IST project SCOUT (Smart user-centric communications environment), www.ist-scout.org.
11. E. C. Van der Meulen, "Three-terminal communication channels," in Advances in Applied Probability, vol. 3, 1971, pp. 120–154.
12. T. M. Cover and A. A. E. Gamal, "Capacity theorems for the relay channel," IEEE Transactions on Information Theory, vol. 25, no. 5, pp. 572–584, Sep 1979.
13. R. Gallager, Communications and Cryptography: Two Sides of One Tapestry, ser. in Engineering & Computer Science. Kluwer, 1994.
14. A. Sendonaris, E. Erkip, and B. Aazhang, "Increasing uplink capacity via user cooperation diversity," in Proc. of IEEE International Symposium on Information Theory (ISIT), Aug. 1998, p. 156.
15. Cooperative wireless networking beyond store-and-forward: Perspectives for PHY and MAC design. Working Group 3. White Paper. Wireless World Research Forum 2006.

16. M. Janani, A. Hedayat, T. E. Hunter, and A. Nosratinia, “Coded cooperation in wireless communications: space-time transmission and iterative decoding,” IEEE Transactions on Signal Processing, vol. 52, no. 2, pp. 362–371, Feb. 2004.
17. Nosratinia, T. E. Hunter, and A. Hedayat, “Cooperative communication in wireless networks,” IEEE Communications Magazine, vol. 42, no. 10, pp. 74–80, Oct. 2004.
18. T. E. Hunter, S. Sanayei, and A. Nosratinia, “Outage analysis of coded cooperation,” IEEE Transactions on Information Theory, vol. 52, no. 2, pp. 375–391, Feb. 2006.
19. H. Shan, W. Zhuang, and Z. Wang, “Distributed cooperative MAC for multi-hop wireless networks,” IEEE Commun. Mag., vol. 47, no. 2, pp. 126–133, Feb. 2009.
20. Z. Zhang, J. Shi, H.-H. Chen, M. Guizani, and P. Qiu, “A cooperation strategy based on nash bargaining solution in cooperative relay networks,” IEEE Trans. Veh. Technol., vol. 57, no. 4, pp. 2570–2577, July 2008.
21. J. Huang, Z. Han, M. Chiang, and H. V. Poor, “Auction-based resource allocation for cooperative communications,” IEEE J. Sel. Areas Commun., vol. 26, no. 7, pp. 1226–1237, Sep. 2008.
22. Q.-Q. Zhang, W.-D. Gao, M.-G. Peng, and W.-B. Wang, “Partner selection strategies in cooperative wireless networks with optimal power distribution,” J. China Universities of Posts and Telecommunications, vol. 15, no. 3, pp. 47–50,58, 2008.
23. A. Nosratinia and T. E. Hunter, “Grouping and partner selection in cooperative wireless networks,” IEEE J. Sel. Areas Commun., vol. 25, no. 2, pp. 369–378, Feb. 2007.
24. Y. J. Zhang and K. B. Letaief, “Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDM systems,” IEEE Trans.Wireless Commun., vol. 3, no. 5, pp. 1566–1575, Sep. 2004.
25. Simon Haykin, “Cognitive Radio: Brain-Empowered Wireless Communications”, IEEE Journal on Selected Areas in Communications, Vol. 23, No. 2, Feb. 2005, pp. 201-220.
26. Hsien-Po Shianh and Mihaela van der Schaar, “Queuing-Based Dynamic Channel Selection for Heterogeneous Multimedia Applications over Cognitive Radio Networks”, IEEE Transactions on Multimedia, Vol. 10, No. 5, Aug. 2008, pp.896-909.
27. Amir Ghasemi and Elvino S. Sousa, “Optimization of spectrum sensing for opportunistic spectrum access in Cognitive Radio Networks”, IEEE CCNC’07, Jan. 2007, pp. 1022 – 1026.
28. Danijela Cabric, Shridhar Mubaraq Mishra, Robert W. Brodersen, “Implementation issues in spectrum sensing for cognitive radios”, IEEE Signals, Systems and Computers, Nov. 2004, pp. 772-776.
29. Danijela Cabric, Artem Tkachenko and Robert W. Brodersen, “Spectrum sensing measurements of pilot, energy and collaborative detection”, IEEE MILCOM’06, Oct. 2006, pp. 1-7.
30. Kwang-Chang Chen and Ramjee Prasad, “Cognitive Radio Networks”, John Wiley&Sons Ltd., 2009, 359 p.

31. Manoj B. S., Ramesh R. Rao and Michele Zorzi, “On the Use of Higher Level Information for Cognitive Networking”, Proceedings of IEEE GLOBECOM 2007, pp. 3569 – 3573.
32. Yiping Xing, R. Chandramouli, Stefan Mangold, and Sai Shankar N., “Dynamic Spectrum Access in Open Spectrum Wireless Networks”, IEEE Journal on Selected Areas in Communications, Vol. 24, No. 3, Mar. 2006, pp. 626-637.
33. Ian F. Akyildiz, Won-Yeol Lee, Mehmet C. Vuran, Shantidev Mohanty, “NeXt Generation / Dynamic Spectrum Access / Cognitive Radio Wireless Networks: A Survey”, Computer Networks, No. 50, 2006, pp. 2127–2159.
34. Carlos Cordeiro, Kiran Challapali, Sai Shankar, and Dagnachew Birru, “IEEE 802.22: An introduction to the first wireless standard based on cognitive radios”, Journal of Communications, Vol.1, No. 1, Apr. 2006, pp. 38-47.
35. Seyed A. Zekavat, Xiukui Li, “Ultimate dynamic spectrum allocation via user-central wireless systems”, Journal of Communications, Vol. 1, No. 1, Apr. 2006, pp.60-67.
36. Charles C. Wang, Gregory J. Pottie, “Variable bit allocation for FH-CDMA wireless communication systems”, IEEE Transaction on Communications, Vol. 50, No. 10, Oct. 2002, pp. 1637-1644.
37. Draft Standard for Information Technology Telecommunications and Information Exchange Between Systems – Local and Metropolitan Area Networks – Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 2: High-speed Physical Layer (PHY) extension in the 2.4 GHz band Corrigendum 1, IEEE std 11b-1999/Cor 1-2001.
38. Sai Shankar N, Carlos Cordeiro, and Kiran Challapali, “Spectrum agile radio: Capacity and QoS implementations of dynamic spectrum assignment”, Proceedings of IEEE GLOBECOM 2005, 2005, pp. 2510-2516.
39. Sirin Tekinay and Bijan Jabbari, “Handover and channel assignment in mobile cellular systems”, IEEE Communications Magazine, Nov. 1991, pp. 42-45.
40. Yan Zhang, “Dynamic Spectrum Access in Cognitive Radio Wireless Networks”, Proceedings of IEEE ICC 2008, pp. 4927-4932.
41. Waqas Ahmed, Jason Gao, Hong Zhou and Michael Faulkner, “Throughput and Proportional Fairness in Cognitive Radio Networks”, IEEE ATC’09, Oct. 2009, pp. 248-252.
42. Ryan W. Thomas, Daniel H. Friend, Luiz A. DaSilva, and Allen B. MacKenzie, “Cognitive Networks: Adaptation and Learning to Achieve End-to-End Performance Objectives”, IEEE Communications Magazine, Topics in Radio Communications, IEEE, Dec. 2006, pp. 51-57.
43. Simon Dobson, Spyros Denazis, Antonio Fernandez, Dominique Gaiti, Erol Gelenbe, Fabio Massacci, Paddy Nixon, Fabrice Saffre, Nikita Schmidt, France Zambonelli, “A Survey of Autonomic Communications”, ACM Transactions on Autonomous and Adaptive Systems, Vol. 1, No. 2, Dec. 2006, pp. 223-259.
44. Mohamed Ahmed, Vinay Kolar, Marina Petrova, Petri Mahonen and Stephen Hailes, “A Component-based Architecture for Cognitive Radio Resource Management”, Proceedings of 4th International Conference on Cognitive Radio

Oriented Wireless Networks and Communications (CROWNCOM'09), Hannover, Germany, June 2009.
45. Erol Gelenbe, Michael Gellman, and Pu Su, "Self-Awareness and Adaptability for Quality of Service", IEEE ISCC'03, 2003, vol.1, pp. 3-9.
46. Erol Gelenbe, Ricardo Lent, and Arturo Nunez, "Self-Awareness Networks and QoS", Proceedings of the IEEE, Sep. 2004, vol.92, pp. 1478-1498.
47. Erol Gelenbe and Peixiang Liu, "QoS and Routing in the Cognitive Packet Network", IEEE WoWMoM'05, Jun. 2005, pp. 517-521.
48. Hang Su, Xi Zhang, "Cross-Layer Based Opportunistic MAC Protocols for QoS Provisioning over Cognitive Radio Wireless Networks", Selected Areas in Communications, IEEE Journal, vol.26, issue 1, Jan. 2008, pp. 118-129.
49. Muraleedharan Rajani, Lisa Ann, "Increasing QoS and Security in 4G Networks Using Cognitive Intelligence", Globecom Workshop, IEEE, Nov. 2007, pp. 1-6.
50. Xiangying Dong, Jiajia Wang, Yong Zhang, Mei Song, Ruijun Feng, "End-to-end QoS Provisioning in Future Cognitive Heterogeneous Networks", Procceding of ICCTA 2009, IEEE, 2009, pp. 425-429.
51. IEEE 802 LAN/MAN Standards Committee 802.22 WG on WRANs (Wireless Regional Area Networks), IEEE, retrieved 2009.
52. C. Stevenson et al., "IEEE 802.22: The First Cognitive Radio Wireless Regional Area Network Standard", IEEE Communications Magazine, Vol 47 (1), pp. 130-138, 2009.
53. K. Pahlavan, P. Krishnamurthy, Principles of Wireless Networks, Prentice Hall, 2002.
54. Future Directions in Cognitive Radio Network Research, Report of NSF workshop, 2009.
55. M. Timmers, S. Pollin, A. Dejonghe, A. Bahai, L. Van der Perre and F. Catthoor, Accumulative Interference Modeling for Distributed Cognitive Radio Networks, Journal of Communications, Vol. 4(3), 2009.
56. W. Ren, Q. Zhao, and A. Swami. Power control in cognitive radio networks: How to cross a multi-lane highway. IEEE Journal on Selected Areas in Communications (JSAC): Special Issue on Stochastic Geometry and Random Graphs for Wireless Networks, 2009.
57. Q. Zhao. Spectrum opportunity and interference constraint in opportunistic spectrum access. In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2007, pp. 605–608.
58. C. Cordeiro, K. Challapali, M. Ghosh, Cognitive PHY and MAC layers for dynamic spectrum access and sharing of TV bands, in Proceedings of IEEE International Workshop on Technology and Policy for Accessing Spectrum, 2006, p. 222.
59. Q. Zhao, L. Tong, A. Swami and Y. Chen, Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework, IEEE Journal on Selected Areas in Communications, Vol. 25(3), 2007, pp. 589-600.
60. R. Urgaonkar and M. J. Neely, Opportunistic Scheduling with Reliability Guarantees in Cognitive Radio Networks, in Proc. IEEE INFOCOM, 2008, pp. 1301-1309.

61. S. Huang, X. Liu and Z. Ding, Opportunistic Spectrum Access in Cognitive Radio Networks, in Proc. IEEE INFOCOM 2008, pp. 1427-1435.
62. H.-P. Shiang and M. van der Schaar, Queuing-Based Dynamic Channel Selection for Heterogeneous Multimedia Applications over Cognitive Radio Networks, IEEE Transactions on Multimedia, Vol. 10 (5), 2008, pp. 896-909.
63. K.-L. A.Yau, P. Komisarczuk, T. Paul, Enhancing network performance in Distributed Cognitive Radio Networks using single-agent and multi-agent Reinforcement Learning, In Proc. IEEE Conference on Local Computer Networks, 2010.
64. T. Jiang, Reinforcement Learning-based Spectrum Sharing for Cognitive Radio, Ph.D. Thesis, University of York, 2011.
65. V. Jacobson, "Congestion Avoidance and Control". In Proc. SIGCOMM'88. Aug. 1988, Vol. 18(4), pp. 314-329.
66. F. Kelly, "Charging and rate control for elastic traffic", Europ. Trans. Telecom, Vol 8 (1997), pp. 33-37.
67. J. Rosenberg, et al., "SIP: Session Initiation Protocol", RFC 3261, June 2002.
68. M. Handley, V. Jacobson, "SDP: Session Description Protocol", RFC 2327, April 1998.
69. IEEE 802.11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. (2007 revision). IEEE-SA. 12 June 2007.doi:10.1109/IEEESTD.2007.373646.
70. 3GPP TS36.300. Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Overall Description.
71. B. Mehdi, Spectrum sharing for future mobile cellular systems, Ph. D. Thesis, University of Oulu, 2009.
72. M. Felegyhazi, M. Cagalj, S.S. Bidokhti, J.P. Hubaux, Noncooperative multi-radio channel allocation in wireless networks, Proceedings of IEEE INFOCOM'07, 2007, pp. 1442-1450.
73. K.-L. A.Yau, P. Komisarczuk, T. Paul, Enhancing network performance in Distributed Cognitive Radio Networks using single-agent and multi-agent Reinforcement Learning, In Proc. IEEE Conference on Local Computer Networks, 2010.
74. T. Jiang, Reinforcement Learning-based Spectrum Sharing for Cognitive Radio, Ph.D. Thesis, University of York, 2011.
75. J. Jin, W. H. Wang, M. Palaniswami, Utility max-min fair resource allocation for communication networks with multipath routing, Computer Communications, vol. 32 (17), 2009, pp. 1802-1809.
76. Z. Cap, E.W. Zegura, Utility max-min: an application oriented flow control: fundamentals, algorithms and fairness, Proceedings of INFOCOM, 1999, pp. 793-801.
77. J. Acharya and R.D. Yates, Dynamic spectrum allocation for uplink users with heterogeneous utilities, IEEE Transactions on Wireless Communications, vol. 8 (3), 2009, pp. 1405-1413.

78. Z.X. Chen, K. Sun, J. Yuan, Y. Wang, Max-utility subcarrier allocation for heterogeneous services in wireless OFDM system, Journal of China Universities of Posts and Telecommunications, vol. 16 (1), 2009, pp. 51-57.
79. S. Shenker, Fundamental design issues for the future Internet, IEEE Journal on Selected Areas of Communications, vol. 13 (7), 1995, pp. 1176-1188.
80. D. V. Lindley (1952), "The theory of queues with a single server", Mathematical Proceedings of the Cambridge Philosophical Society, 48 (2), pp. 277–289.
81. 3GPP TR 25.913, Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN), Release 8.
82. Gilberto Berardinelli et. al., OFDMA vs. SC-FDMA: Performance comparison in local area IMT-A Scenarios, IEEE Wireless Communications, Vol. 15 (5), 2008.
83. V. Osa, C. Herranz, J. F. Monserrat and X. Gelabert, Implementing opportunistic spectrum access in LTE-advanced, EURASIP Journal on Wireless Communications and Networking, 2012.
84. A. Alizadeh, S. M.-S. Sadough, S. A. Ghorashi, Relay Selection and Resource Allocation in LTE-Advanced Cognitive Relay Networks, Vol. 1 (4), 2011, pp. 303-310.
85. Y. Chen, C. Cho, I. You and H. Chao, A cross-layer protocol of spectrum mobility and handover in cognitive LTE networks, in Proc. of Simulation Modelling Practice and Theory, 2011, pp. 1723-1744.
86. Q. Zhao, L. Tong, A. Swami and Y. Chen, Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework, IEEE Journal on Selected Areas in Communications, Vol. 25 (3), 2007, pp. 589-600.
87. R. Urgaonkar and M. J. Neely, Opportunistic Scheduling with Reliability Guarantees in Cognitive Radio Networks, in Proc. IEEE INFOCOM, 2008, pp. 1301-1309.
88. S. Huang, X. Liu and Z. Ding, Opportunistic Spectrum Access in Cognitive Radio Networks, in Proc. IEEE INFOCOM, 2008, pp. 1427-1435.
89. H.-P. Shiang and M. van der Schaar, Queuing-Based Dynamic Channel Selection for Heterogeneous Multimedia Applications Over Cognitive Radio Networks, IEEE Transactions on Multimedia, Vol. 10 (5), 2008, pp. 896-909.
90. S.-Y. Lien and K.-C. Chen, Statistical Traffic Control for Cognitive Radio Empowered LTE-Advanced with Network MIMO, in Proc. IEEE INFOCOM, 2011, pp. 80-84.
91. S. Wang et.al., "A Characterization of Delay Performance of Cognitive Medium Access", IEEE Transaction on Wireless Communications, Vol. 11 (2), 2012, pp. 800 – 809.
92. L. Yong et.al., "Pricing-based Spectrum Access Control in Cognitive Radio Network with Random Access", in Proc. IEEE INFOCOM, 2011, pp. 2228-2236.
93. "LTE for UMTS: Evolution to LTE-Advanced" by Harri Holma and Antti Toskala, John Wiley and Sons (2011). 576 pp.
94. Anna Larmo et. al., "The LTE link-layer design". IEEE Comm. Mag. 47 (4), pp. 52-59, 2009.
95. D. Kivanc, G. Li, and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," IEEE Trans. Wireless Commun., vol. 2, no. 6, pp. 1150-1158, Nov. 2003.

96. M. Ergen, S. Coleri, and P. Varaiya, “QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless systems," IEEE Trans. Broadcasting, vol. 49, no. 4, pp. 362-370, Dec. 2003.
97. G. Song and Y. L., “Cross-layer optimization for OFDM wireless networks–part I-II," IEEE Trans. Wireless Commun., vol. 4, no. 2, pp. 614-634, Mar. 2005.
98. H. Yin and H. Liu, “An efficient multiuser loading algorithm for OFDM based broadband wireless systems," in Proc. IEEE GLOBECOM, 2000.
99. Malek Boussif , Nestor Quintero, Francesco D. Calabrese, Claudio Rosa, and Jeroen Wigard, “Interference Based Power Control Performance in LTE Uplink”, in Proc. of IEEE ISWCS 2008, pp. 698-702.
100. Haiming Wang, Dajie Jiang “Performance Comparison of Control-less Scheduling Policies for VoIP in LTE UL”, in Proc. of WCNC 2008, pp. 2497-2501.
101. R. W. Thomas, D. H. Friend and L. A. DaSilva and A. B. MacKenzie, “Cognitive Networks Adaption and Learning to Achieve End-to-End Performance Objectives”, IEEE Communications Magazine, December 2006, pp. 51-57.
102. E. H. Ong and J. Y. Khan, “Cooperative Radio Resource Management Framework for Future IP based Multiple Radio Access Technologies Environment”, Computer Networks, vol:54, no:7, May 2010, pp. 1083-1107.
103. K-C Chen and R. Prasad, Cognitive Radio Networks, John Wiley & Sons, 2009.
104. Markus Dillinger, Kambiz Madani, Nancy Alonistioti, Software Defined Radio: Architectures, Systems and Functions, Wiley & Sons, 2003, 454 p.
105. G. Janacek and L. Swift. Time series: forecasting, simulation, applications. Ellis Horwood Limited, 1993.
106. T. Soderstrom and P. Stoica. System identification. Prentice Hall International (UK) Ltd, 1989.
107. Y. C. Ho. On the stochastic approximation method and optimal filtering theory. Journal of Mathematical Analysis and Applications, Vol. 6, pp. 152-154, 1963.
108. V. Paxson and S. Floyd. Wide-Area Traffic: The Failure of Poisson Modeling. IEEE/ACM Transactions on Networking, 3(3), pp. 226-244, June 1995.
109. H. E. Hurst. Long-term storage capacity of reservoirs, Transactions on American Scoiety of Civil Engineers, pp. 770 – 808, 1951.
110. H. Akaike. Fitting autoregressive models for prediction. Ann. Inst. Statist. Math., Vol. 21, pp. 243-247, 1969.
111. Q. Xia, X. Jin and M. Hamdi, “Active Queue Management with Dual Virtual Proportional Integral Queues for TCP Uplink/Downlink Fairness in Infrastructure WLANs”, IEEE Trans. On Wireless Communications, Vol. 7 (6), 2008, pp. 2261-2271.
112. OPNET website: www.opnet.com
113. S. Senkindu and H. A. Chan, “Enabling end to end quality service in a WLAN-Wired Network”, IEEE 2008.
114. P. Gopalakrishnan, D. Famolari and T. Kodama, Improving WLAN Voice Capacity through Dynamic Priority Access”. In Proc. GLOBECOM 2004.
115. LTE Physical Layer Framework for Performance Verification. (2007) 3GPP R1-070674.

116. Puttonen J. et. al. (2008) 'Voice-over-IP Performance in UTRA Long Term Evolution Downlink', in Proc. IEEE VTC'S08.
117. Persson F. (2007) 'Voice over IP Realized for the 3GPP Long Term Evolution', in Proc.IEEE VTC'F07.
118. Jiang D. et. al. (2007) 'Principle and Performance of Semi-Persistent Scheduling for VoIP in LTE System', in Proc. WiCom'07, pp. 2861–2864.
119. Puttonen J. et. al. (2008). 'Persistent Packet Scheduling Performance for Voice-over-IP in Evolved UTRAN Downlink', in Proc. PIMRC'08.
120. Fan Y. et. al. (2008) 'Efficient Semi-Persistent Scheduling for VoIP on EUTRA Downlink', in Proc. IEEE VTC'F08.
121. Wong C. Y. et. al. (1999) 'Multiuser OFDM with adaptive subcarrier, bit and power allocation', IEEE J.Select. Areas Commun., Vol. 17, No. 10, pp. 1747-1758.
122. Kivanc D., Li G., and Liu H. (2003) 'Computationally efficient bandwidth allocation and power control for OFDMA', IEEE Trans. Wireless Commun., Vol. 2, No. 6, pp. 1150-1158.
123. Ergen M., Coleri S., and Varaiya P. (2003) 'QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless systems', IEEE Trans. Broadcasting, Vol. 49, No. 4, pp. 362-370.
124. Song G. and Li G. (2005) 'Cross-layer optimization for OFDM wireless networks–part I-II', IEEE Trans. Wireless Commun., Vol. 4, No. 2, pp. 614-634.
125. Yin H. and Liu H. (2000). 'An efficient multiuser loading algorithm for OFDM based broadband wireless systems', in Proc. IEEE GLOBECOM 2000.
126. Boussif M. et. al. (2008) 'Interference Based Power Control Performance in LTE Uplink', in Proc. IEEE ISWCS'08, pp. 698-702.
127. Wang H. and Jiang D. (2008) 'Performance Comparison of Control-less Scheduling Policies for VoIP in LTE UL', in Proc. of WCNC'08, pp. 2497-2501.
128. Holma H. and Toskala A.. (2011) LTE for UMTS: Evolution to LTE-Advanced, John Wiley and Sons.
129. E-UTRA and E-UTRAN; Overall Description. 3GPP TS 36.300. Stage 2 (Release 8).
130. Larmo A. et. al. (2009) 'The LTE link-layer design', IEEE Comm. Mag., Vol. 47, No. 4, pp. 52-59.
131. Physical Channels and Modulation. 3GPP TS 36.211. (Release 8).
132. Kela P. et.al. (2008) 'Dynamic Packet Scheduling Performance in UTRA Long Term Evolution Downlink', in Proc. IEEE ISWPC'08.
133. E-UTRA; MAC protocol specification . 3GPP TS 36.321. (Release 8).
134. Multiplexing and Channel Coding. 3GPP TS 36.212. (Release 8).
135. Ghosh A. and Zhang J. (2010). Fundamentals of LTE. The Prentice Hall communications engineering and emerging technologies series.
136. Number of Control Symbols. (2007). 3GPP TSG-RAN WG2 Meeting #57bis, R2-071227.
137. ITU-T Recommendation M.1225. (1997) Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000. Question ITU-R 39/8.
138. Rosenberg J. et al. (2002) SIP: Session Initiation Protocol. RFC 3261.

139. Handley M. and Jacobson V.(1998) SDP: Session Description Protocol. RFC 2327.
140. Multimedia Telephony; Media handling and interaction. (2007) 3GPP TS 26.114.
141. Service Aspects and Service Capabilities. (2002) 3GPP TS 22.105. Version 5.1.0.
142. Vieira P., Queluz P. and Rodrigues A. (2008) LTE spectral efficiency using spatial multiplexing MIMO formacro-cells', in Proc. IEEE ICSPCS'08.
143. Pokhariyal A., Kolding T.E. and Mogensen P.E. (2006) 'Performance of Downlink Frequency Domain Packet Scheduling For the UTRAN Long Term Evolution', The 17th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Helsinki, Sep. 2006.
144. Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios. (2009) 3GPP TR 36.942. Version 8.2.0 (Release 8).
145. S. Glisic, B. Lorenzo (2009), "Advanced Wireless Networks: Cognitive, Cooperative & Opportunistic 4G Technology", 2nd Edition, John Wiley & Sons Ltd., 892 pp.
146. D. Gözüpek, F. Alagöz, "Genetic algorithm-based scheduling in cognitive radio networks under interference temperature constraints". Int. J. Commun. Syst. (2010). DOI: 10.1002/dac.1152.
147. W. Wang, W. Wang, Q. Lu, T. Peng, "An uplink resource allocation scheme for OFDMA-based cognitive radio networks". Int. J. Commun. Syst. (2009). Vol. 22(5), pp. 603-623.
148. D. T. Ngo, C. Tellambura, H. H. Nguyen, "Resource Allocation for OFDM-Based Cognitive Radio Multicast Networks". In Proc. WCNC'09. April 2009, pp. 1-6.
149. J. Neel, R.M. Buehrer, B.H. Reed, R.P. Gilles, Game theoretic analysis of a network of cognitive radios, In Proc.MWSCAS-2002, 2002, vol 3, pp. 409-412.
150. R.W. Thomas, Cognitive Networks, PhD thesis, Virginia Polytechnic Institute and State University, 2007.
151. M. Felegyhazi, M. Cagalj, S.S. Bidokhti, J.P. Hubaux, Noncooperative multi-radio channel allocation in wireless networks, In Proc. INFOCOM'07, 2007, pp. 1442-1450.
152. G. He, S. Gault, M. Debbah, E. Altman, Distributed power allocation game for uplink ofdm systems, In Proc. WiOPT, 2008, pp. 515-521.
153. J. Jin, W. H. Wang, M. Palaniswami, Utility max-min fair resource allocation for communication networks with multipath routing, Comp. Comm., vol. 32 (17), 2009, pp. 1802-1809.
154. J. H. Mo and J. Walrand, "Fair End-to-End Window-based Congestion Control", IEEE/ACM Transactions on Networking. Vol.9, pp.556-567.
155. H. W. Kuhn, A. W. Tucker (1951), "Nonlinear programming". In Proc. 2nd Berkeley Symposium, Berkeley: University of California Press. pp. 481–492.
156. Sergio Benedetto and Ezio Biglieri (1999). Principles of Digital Transmission: With Wireless Applications. Springer.
157. 3GPP TS 36.321. E-UTRA; MAC protocol specification (Release 8).
158. P. Kela et.al., "Dynamic Packet Scheduling Performance in UTRA Long Term Evolution Downlink," in Proc. ISWPC'08, May 2008.
159. 3GPP TS 36.211. Physical Channels and Modulation (Release 8).

160. H. -P. Shiang and M. van der Schaar, Queuing-Based Dynamic Channel Selection for Heterogeneous Multimedia Applications Over Cognitive Radio Networks, IEEE Transactions on Multimedia, Vol. 10 (5), 2008, pp. 896-909.
161. R.S. Sutton and A.G. Barto, Reinforcement Learning: An Introduction. Cambridge, UE: MIT Press, 1998.
162. Y. Nesterov, Smooth minimization of non-smooth functions, Mathematical Programming, Vol. 103 (1), 2005, pp. 127-152.
163. N.Z. Shor, Minimization Methods for Non-differentiable Functions. Springer-Verlag, 1985.
164. Y. Nesterov and A. Nemirovsky, Interior Point Polynomial Methods in Convex Programming. SIAM, 1994.
165. S.P. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, 2004.
166. IEEE 802.11g-2003: Further Higher Data Rate Extension in the 2.4 GHz Band. IEEE. 2003-10-20. Retrieved 2007.
167. IPOGUE, Internet study 2007 and 2008/2009, research report.
168. Erik A. van Doorn, Some Aspects of the Peakedness Concept in Teletraffic Theory, Journal of Information Processing and Cybernetics, Vol. 22, 2/3, pp. 93 – 104, 1986.
169. 3GPP TS 36.213, Evolved Universal Terrestrial Radio Access (E-UTRA).
170. A. Saatsakis, Cognitive Radio Resource Management for Improving the Efficiency of LTE Network Segments in the Wireless B3G World, DySPAN, 2008, pp. 1-5, 2008.Spectrum Policy Task Force, FCC Report 02-135, 2002.
171. M.R.Hassan, G. Karmakar, J. Kamruzzaman, Dynamic bandwidth access to cognitive radio ad hoc networks through pricing modeling, APCC, 2011, pp. 12-17.
172. J. V. Mynsbrugge, Bidding Strategies Using Price Based Unit Commitment in a Deregulated Power Market, K.U.Leuven, 2010.
173. G. Bianchi, IEEE 802.11-Saturation throughput analysis, IEEE Commun. Lett., Vol. 2, 1998, pp. 318–320
174. A. V. Barbosa, M. F. Caetano, J. L. Bordim, The Theoretical Maximum Throughput Calculation for the IEEE802.11g Standard, International Journal of Computer Science and Network Security (IJCSNS), Vol. 11 (4), 2011, pp. 136 – 143.
175. G. Bianchi, Performance Analysis of the IEEE 802.11 Distributed Coordination Function, IEEE Journal on Selected Areas in Communications, Vol. 8, 2000, pp. 535-547.
176. J. Jun, P. Peddabachagari, M. Sichitiu, Theoretical Maximum Throughput of IEEE 802.11 and its Applications, In Proc, IEEE International Symposium on Network Computing and Applications, 2003, pp. 249-256.
177. V. Osa, C. Herranz, J. F. Monserrat and X. Gelabert. Implementing opportunistic spectrum access in LTE-advanced. EURASIP Journal on Wireless Communications and Networking, 2012:99, 2012.
178. A. Alizadeh, S. M. S. Sadough and S. A. Ghorashi. Relay selection and resource allocation in LTE-advanced cognitive relay networks. International Journal on Communications Antenna and Propagation, 1(4):303–310, 2011.

179. Y. Chen, C. Cho, I. You and H. Chao. A cross-layer protocol of spectrum mobility and handover in cognitive LTE networks. In Simulation Modelling Practice and Theory, 1723–1744, 2011.
180. P. Tran-Gia, D. Staehle, K. Leibnitz, Source Traffic Modeling of Wireless Applications, International Journal of Electronics and Communications, Vol. 55 (1), 2001, pp. 27 – 36.
181. A. Iyer, C. Rosenberg, A. Karnik, What is the right model for wireless channel interference?, IEEE Trans. Wireless Comm., Vol. 8 (5), 2009, pp. 2662 – 2671.
182. J. Elias et. al., Non-Cooperative Spectrum Access in Cognitive Radio Networks: a Game Theoretical Model, Computer Networks, Vol. 55 (17), 2011, pp. 3832–3846.
183. M. Maskery et. al., Decentralized Dynamic Spectrum Access for Cognitive Radios: Cooperative Design of a Non-Cooperative Game, IEEE Trans. on Communications, Vol. 57 (2), 2009, pp. 459 – 469.
184. B. Wang et. al., Game theory for cognitive radio networks: An overview, Computer Networks, Vol. 54 (14), 2010, pp. 2537–2561.
185. D. Fudenberg and J. Tirole, Game Theory. Cambridge, MA: MIT Press, 1991.
186. K. Akkarajitsakul et. al., Game theoretic approaches for multiple access in wireless networks: A survey, Communications Surveys Tutorials, IEEE, Vol. 13 (3), 2011, pp. 372 –395.
187. O. E. Ferkouss and W. Ajib, Game theory based resource allocation for cognitive radio networks, in Proc. GLOBECOM, 2012, pp. 1174 – 1179.
188. J. Elias et. al., Joint Pricing and Cognitive Radio Network Selection: a Game Theoretical Approach, in Proc. WiOPT, 2012, pp. 49 – 53.
189. X. Kang, R. Zhang, M. Motani, Price-Based Resource Allocation for Spectrum-Sharing Femtocell Networks: A Stackelberg Game Approach, IEEE Journal on Selected Areas in Communications, Vol. 30 (3), 2012, pp. 538 – 549.
190. N. Karamchandani et. al., Cooperation in multi-access networks via coalitional game theory, in Proc. Communication Control, and Computing (Allerton), 2011, pp. 329 –336.
191. H. Xiaolei et. al., A Coalition Formation Game for Energy-Efficient Cooperative Spectrum Sensing in Cognitive Radio Networks with Multiple Channels, In Proc. GLOBECOM, 2011, pp. 1 – 6.
192. C.-G. Yang et. al., Optimal Power Control for Cognitive Radio Networks Under Coupled Interference Constraints: A Cooperative Game-Theoretic Perspective, IEEE Trans. on Vehicular Tech., Vol 59 (4), 2010, pp. 1696 – 1706.
193. A. T. Hoang and Y.-C. Liang, Downlink Channel Assignment and Power Control for Cognitive Radio Networks, IEEE Trans. on Wireless Comm., Vol. 7 (8), 2008, 06 – 3117.
194. CDMA2000 Evaluation Methodology Version 1.0 (Revision 0).