

RADIO COMMUNICATIONS

RADIO COMMUNICATIONS

Edited by
ALESSANDRO BAZZI

Published by In-Teh

In-Teh

Olajnica 19/2, 32000 Vukovar, Croatia

Abstracting and non-profit use of the material is permitted with credit to the source. Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published articles. Publisher assumes no responsibility liability for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained inside. After this work has been published by the In-Teh, authors have the right to republish it, in whole or part, in any publication of which they are an author or editor, and the make other personal use of the work.

© 2010 In-teh

www.intechweb.org

Additional copies can be obtained from:

publication@intechweb.org

First published April 2010

Printed in India

Technical Editor: Martina Peric

Cover designed by Dino Smrekar

Radio Communications,

Edited by Alessandro Bazzi

p. cm.

ISBN 978-953-307-091-9

Preface

In the last decades the restless evolution of information and communication technologies (ICT) brought to a deep transformation of our habits. The growth of the Internet and the advances in hardware and software implementations modified our way to communicate and to share information.

Although conceived for study and working scopes, the presence of ICT infrastructures and services are now pervasive in our life. Their availability is felt as a need that cannot be constrained to a place: people want to have them anytime, anywhere. And the need for wireless connections starts from daily communication and embraces all fields of human life, in an ever increasing way: transportation systems, home management, healthcare, emergency operations. This brought to the large interest and great evolution that characterized radio technologies in the last years all over the world. Since the development of the first wireless systems, great technological advances were performed: spread spectrum and multi-carrier techniques allowed to make transmissions over larger bandwidths with higher data rates for the final users in challenging scenarios where line of sight is not guaranteed and multipath severely affects signal propagation; new coding techniques were developed to increase the efficiency of radio transmissions; advanced link adaptation and power control algorithms were implemented in order to allow optimized trade-off between reliability and data rate; and many other topics shall be cited.

Cellular systems, initially deployed to allow telephony in mobility for urgency needs have now reached a penetration of almost 100% of inhabitants in developed countries, with the ability to exchange data at more than 10 Mbit/s; at the same time, wireless access to local networks can now be performed at almost 100 Mbit/s. But the need for higher data rates does not stop and it jointly proceeds with the development of new services, thus requiring the investigation and implementation of new technologies and new radio resources management solutions. In the near future, all available technologies will be jointly used by smart devices, following the always best connected paradigm: users will connect to an heterogeneous wireless network, without the need to know which technology they are effectively using. The adoption of multiple antennas and beamforming techniques will reduce the effects of interference and increase the achievable data rates. Relaying and cooperation among devices will allow better performance with lower costs and energy consumption.

In this book, an overview of the major issues faced today by researchers in the field of radio communications is given through 35 high quality chapters written by specialists working in universities and research centers all over the world. Various aspects will be deeply discussed: channel modeling, beamforming, multiple antennas, cooperative networks, opportunistic

scheduling, advanced admission control, handover management, systems performance assessment, routing issues in mobility conditions, localization, web security. Advanced techniques for the radio resource management will be discussed both in single and multiple radio technologies; either in infrastructure, mesh or ad hoc networks. In particular, the book is organized following a bottom up structure, starting from the physical level up to the whole system. More precisely, the physical layer aspects are discussed through chapters 1 to 10, with particular emphasis to MIMO systems; from chapter 11 to chapter 18, data link layer and radio resource management are handled without focusing to a specific technology: backward error correction, scheduling, relaying; a deeper investigation of systems performance, with particular reference to WiMAX and WiFi systems, is then given through chapters 19 to 25; heterogeneous wireless networks are then examined through chapters 26 to 29; the cross-layer issues of localization and web security, which are of increasing interest in modern wireless technologies, are finally discussed in the remaining part of the book.

Although these pages will not cover all the aspects of such a wide topic, the reader will find a helpful overview on what is happening now and what is expected in an early future in the field of radio communications. It must be remarked that this book was possible thanks to the valuable work of all authors involved and to the IN-TECH technical staff.

Alessandro Bazzi*

*Alessandro Bazzi works at WiLab (www.wilab.org), an organization with affiliates belonging to the Italian National Council for Researches CNR (ICT Department, IEIIT Institute) and Italian Universities of Bologna and Ferrara. WiLab has expertise in the field of telecommunications systems with particular emphasis on wireless systems; research activity is mainly performed within the context of important national and international projects, in relation with some of the most important manufacturers and service providers, and in strict cooperation with colleagues coming from the main universities in Europe and United States.

Contents

Preface	V
1. Radio-communications architectures Antoine Diet, Martine Villegas, Geneviève Baudoin and Fabien Robert	001
2. Analytical SIR for Cross Layer Channel Model Abdurazak Mudesir and Harald Haas	037
3. The Impact of Fixed and Moving Scatterers on the Statistics of MIMO Vehicle-to-Vehicle Channels Ali Chelli and Matthias Pätzold	051
4. Planar Antenna Array Hybrid Beamforming for SDMA in Millimeter Wave WPAN Sau-HsuanWu, Lin-Kai Chiu, Ko-Yen Lin and Ming-Chen Chiang	065
5. A Distributed Multilayer Software Architecture for MIMO Testbeds José A. García-Naya, M. González-López and L. Castedo	077
6. Recent Developments in Channel Estimation and Detection for MIMO Systems Seyed Mohammad-Sajad Sadough and Mohammad-Ali Khalighi	099
7. Cooperative MIMO Systems in Wireless Sensor Networks M. Riduan Ahmad, Eryk Dutkiewicz, Xiaojing Huang and M. Kadim Suaidi	123
8. Optimal Cooperative MIMO Scheme in Wireless Sensor Networks M. Riduan Ahmad, Eryk Dutkiewicz, Xiaojing Huang and M. Kadim Suaidi	151
9. Single/Multi-User MIMO Differential Capacity Daniel Castanheira and Atilio Gameiro	167
10. Low Dimensional MIMO Systems with Finite Sized Constellation Inputs Rizwan Ghaffar and Raymond Knopp	185
11. Advanced Hybrid-ARQ Receivers for Broadband MIMO Communications Tarik Ait-Idir, Houda Chafnaji, Samir Saoudi and Athanasios Vasilakos	211
12. Cooperative ARQ: A Medium Access Control (MAC) Layer Perspective Jesús Alonso-Zárate, Elli Kartsakli, Luis Alonso and Christos Verikoukis	227

13. A Hybrid Feedback Mechanism to Exploit Multiuser Diversity in Wireless Networks Yahya S. Al-Harhi	247
14. Opportunistic Access Schemes for Multiuser OFDM Wireless Networks Cédric Gueguen and Sébastien Baey	265
15. Bidirectional Cooperative Relaying Prabhat Kumar Upadhyay and Shankar Prakriya	281
16. A Novel Amplify-and-Forward Relay Channel Model for Mobile-to-Mobile Fading Channels Under Line-of-Sight Conditions Batool Talha and Matthias Pätzold	307
17. Resource Management with Limited Capability of Fixed Relay Station in Multi-hop Cellular Networks Jemin Lee and Daesik Hong	321
18. On Cross-layer Routing in Wireless Multi-Hop Networks Golnaz Karbaschi, Anne Fladenmuller and Sébastien Baey	339
19. Mobile WiMAX Performance Investigation Alessandro Bazzi, Giacomo Leonardi, Gianni Pasolini and Oreste Andrisano	361
20. Throughput-Enhanced Communication Approach for Subscriber Stations in IEEE 802.16 Point-to-Multipoint Networks Chung-Hsien Hsu and Kai-Ten Feng	385
21. Holdoff Algorithms for IEEE 802.16 Mesh Mode in Multi-hop Wireless Mesh Networks Bong Chan Kim and Hwang Soo Lee	399
22. Call Admission Control Algorithms based on Random Waypoint Mobility for IEEE802.16e Networks Khalil Ibrahimi, Rachid El-Azouzi, Thierry Peyre and El Houssine Bouyakhf	419
23. Queueing-Model-Based Analysis for IEEE802.11 Wireless LANs with Non-Saturated Nodes Shigeo Shioda and Mayumi Komatsu	441
24. Increasing the Time Connected to Already Deployed 802.11 Wireless Networks while Traveling by Subway Jaecouk Ok, Pedro Morales, Masateru Minami and Hiroyuki Morikawa	457
25. Asymmetric carrier sense in heterogeneous medical networks environment Bin Zhen, Huan-Bang Li, Shinsuke Hara and Ryuji Kohno	473
26. Multi-Agent Design for the Physical Layer of a Distributed Base Station Network Philippe Leroux and Sébastien Roy	493

27. Inter-RAT Handover Between UMTS And WiMAX Bin LIU, Philippe Martins, Philippe Bertin and Abed Ellatif Samhat	523
28. MHD-CAR: A Distributed Cross-Layer Solution for Augmenting Seamless Mobility Management Protocols Faqr Zarrar Yousaf, Christian Müller and Christian Wietfeld	551
29. Mobility in IP Networks: From Link Layer to Application Layer Protocols and Architectures Thienne Johnson, Eleri Cardozo, Rodrigo Prado, Eduardo Zagari and Tomas Badan	573
30. Positioning in Indoor Mobile Systems Miloš Borenović and Aleksandar Nešković	597
31. Location in Ad Hoc Networks Israel Martin-Escalona, Marc Ciurana and Francisco Barcelo-Arroyo	619
32. Location Tracking Schemes for Broadband Wireless Networks Po-Hsuan Tseng and Kai-Ten Feng	639
33. Wireless Multi-hop Localization Games for Entertainment Computing Tomoya Takenaka, Hiroshi Mineno and Tadanori Mizuno	651
34. Measuring Network Security Emmanouil Serrelis and Nikolaos Alexandris	673
35. A testing process for Interoperability and Conformance of secure Web Services Spyridon Papastergiou and Despina Polemi	689

Radio-communications architectures

Antoine Diet*, Martine Villegas**, Geneviève Baudoin** and Fabien Robert**

**Université Paris-Sud 11, DRÉ, UMR 8506*

***Université Paris-Est, ESYCOM, ESIEE Paris
France*

1. Introduction

Wireless communications, i.e. radio-communications, are widely used for our different daily needs. Examples are numerous and standard names like BLUETOOTH, WiFi, WiMAX, UMTS, GSM and, more recently, LTE are well-known [Baudoin et al. 2007]. General applications in the RFID or UWB contexts are the subject of many papers. This chapter presents radio-frequency (RF) communication systems architecture for mobile, wireless local area networks (WLAN) and connectivity terminals. An important aspect of today's applications is the data rate increase, especially in connectivity standards like WiFi and WiMAX, because the user demands high Quality of Service (QoS). To increase the data rate we tend to use wideband or multi-standard architecture. The concept of software radio includes a self-reconfigurable radio link and is described here on its RF aspects. The term multi-radio is preferred. This chapter focuses on the transmitter, yet some considerations about the receiver are given. An important aspect of the architecture is that a transceiver is built with respect to the radio-communications signals. We classify them in section 2 by differentiating Continuous Wave (CW) and Impulse Radio (IR) systems. Section 3 is the technical background one has to consider for actual applications. Section 4 summarizes state-of-the-art high data rate architectures and the latest research in multi-radio systems. In section 5, IR architectures for Ultra Wide Band (UWB) systems complete this overview; we will also underline the coexistence and compatibility challenges between CW and IR systems.

2. Transceiver aspects for radio-communications

2.1 Radio communications signals

Radio-communications applications deal with communicating and non-communicating links with their different parameters. People expect high quality from their different services (QoS) whatever the telecommunications system used. For example, voice (low data rate) or visio-phone and multimedia download (high data rate) are assumed to be present on the new generation mobile phone. This reveals the co-existence and interaction goals between mobile communication systems (GSM, GSM EDGE, UMTS...) and connectivity standards (BLUETOOTH, WiFi, WiMAX...) [Baudoin et al., 2007]. Thanks to Impulse Radio

Ultra Wide Band (IR-UWB), other fields of interest such as Radio Frequency Identification (RFID) and localization systems are examples of where radio-communications transceivers are currently being designed. Each kind of application can be classified by the resulting radio signals emitted/received. Determining factors are (i) the use of power efficient or spectrum efficient modulation schemes, (ii) the frequency and type of carrier signal used: Continuous Wave (CW) or Impulse Radio (IR) based signals and (iii) the data rate needed (defining a major subdivision of CW based systems). Depending on the choice of the factors involved, the design faced by the RF architect can be varied and challenging. We differentiate three types of cases in this chapter, as illustrated in Fig. 1.

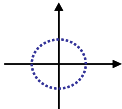
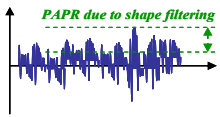
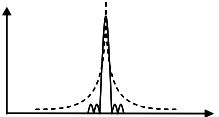
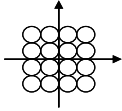

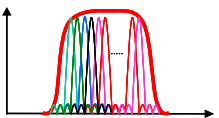
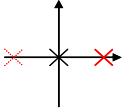
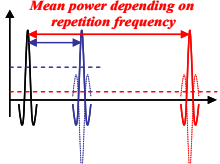
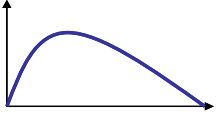
		Typical Modulation Scheme	Time signal (with shaping filter for CW)	Frequency spectrum	PROS.
CW	NB-CW				Power efficient
	WB-CW				Spectrum efficient
IR	IR-UWB				Power saving Spread spectrum

Fig. 1. Main types of radio-communications signals

1) NarrowBand CW (NB-CW) systems, like GSM (GMSK), EDGE (D-QPSK), BLUETOOTH, RFID tags, etc... These systems are often power efficient modulation schemes (EDGE is an exception) because these applications use a low data rate transfer. Major problems of the NB-CW architecture involve with the coexistence and the signal protection against interferers or blocking signals. Some spread-spectrum techniques are often added to improve the communication range (frequency-hopping in the case of BLUETOOTH). The NB-CW case is also considered to be the classic radio communications link and reference system because it corresponds to the popular AM/FM radio broadcasts. NB-CW systems are often considered as using constant envelope signals. This could be true only if FSK modulation schemes are performed, but it is also possible to use x-QAM low symbol rate modulations (EDGE). Additionally, the shaping filter (root raised cosine filter) implies amplitude variations (non-constant envelope) even on FSK modulated signals (GSM).

2) WideBand CW (WB-CW) systems, like: UMTS (W-CDMA), WiMAX (OFDM enhanced), LTE (OFDM based), WiFi (OFDM), UWB (OFDM version)... These systems correspond to high data rate transfer, often for multimedia applications. We can see an increasing need in this field of interest due to the large number of new standards which can be found. Due to the bandwidth limitations for each standard, the modulation scheme is often spectrum efficient (x-QAM) and the use of multi-carrier transmission is usually performed, e.g., OFDM or MC-CDMA. Since there are high amplitude variations of such signals, the transceiver architecture is designed in function with the unavoidable non-linear effects (NL) caused by the power amplification block. Emitting a high PAPR (Peak to Average Power Ratio) has always been a well-known transmitter challenge. The linearization of such a transmitter is mandatory and will be discussed in part 3. WB-CW transceivers also need a wideband design for all of its key elements: antennas, LNA (Low Noise amplifier), HPA (High Power Amplifier) and mixers. This often results in lower performance of the above-mentioned blocks than for NB-CW systems.

3) IR-UWB systems such as UWB localization systems, RFID-UWB... These systems are special because they are based on (one of the) spread spectrum techniques in order to protect the information. The idea is to spread the information in frequency while lowering the emitted power. The use of UWB (3.1-10.6 GHz in the USA) was highly discussed in order to evaluate its co-existence with NB-CW and WB-CW. What is important here is how the power amplification is processed differently for this type of communication. An average power is defined in function of the Pulse Repetition Frequency (PRF), given by the specifications of the IR-UWB standard. For a fixed emitted power: the shorter the pulse, the higher the instantaneous power and bandwidth. Transceivers for these signals are based on impulse generators and energy detectors or correlation receivers, see section 5.

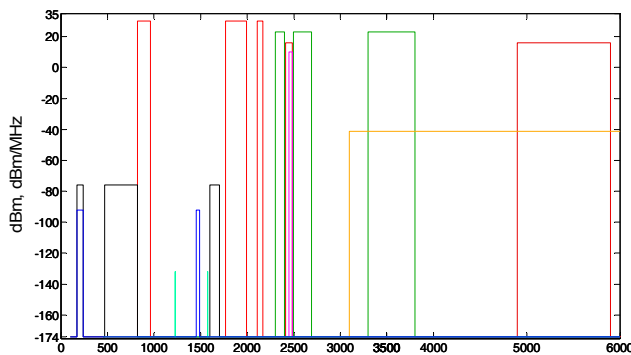


Fig. 2. Telecommunications spectrum sharing and power limitations up to 6 GHz

The three types described can represent every kind of radio-communications signal. The characteristics which have an impact on the architecture are mainly the centre frequency (choice of the technology) and bandwidth (circuits' topologies and performance limitations), and also the PAPR for CW signals. Fig. 2 qualitatively summarizes the power limitations and frequency specifications for some telecommunications standards up to 6 GHz: cellular, WiMAX, WiFi, Bluetooth, UWB and DVB-H. The goal of RF architecture is to emit and receive such signals with no alteration of the information (no constellation distortion for CW). Designing the architecture implies other considerations such as noise, linearity, efficiency and

systems co-existence and immunity. The receiver part of a transceiver has to correctly identify information without adding too much noise, even when high power unwanted signals are close (in the frequency domain). The transmitter part has to linearly amplify the signal in order to emit as far as possible, respecting the standard power limitations (spectrum mask) for co-existence. Architectures for NB-CW, WB-CW and IR-UWB use some basic elements (blocks) that will be described in the next sub-section. Multi-radio is interpreted as a possible reconfigurable architecture for most of the signals presented (mainly NB-CW and WB-CW). This helps drive improvements on classic structures.

2.2 Basic elements and their imperfections

Transceiver architectures for radio-communications signals is defined from the Digital to Analog Converter (DACs, in the baseband section) to the transmitter (Tx) antenna and, respectively, from the receiving (Rx) antenna to the Analog to Digital Converter (ADCs). Each of the Tx and Rx sections deal with unavoidable tradeoffs such as linearity/efficiency (Tx), noise/gain (Rx). Other transceivers in the spectrum vicinity are supposed to correctly receive (co-existence) and/or emit their signals (immunity) without lowering QoS. The basic functions (blocks) in radio-communications are conversion (digital/analog), high frequency transposition and modulation, filtering, power or low noise amplification and radiation/sensing (Tx/Rx antenna). Here, we are describing a system and how it relates to these blocks in radio-communications architecture. We will focus on their imperfections and their impact on the system performance (noise, spectrum distortion...). Sometimes, for CW standards, the influence of noise or spectral re-growths is quantified by certain criteria such as the EVM and/or the ACPR as defined in Fig. 3 [Baudoin et al., 2007][Villegas et al., 2007].

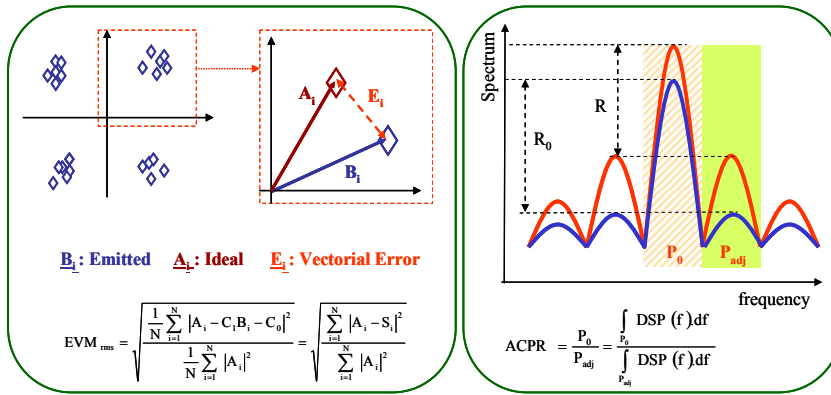


Fig. 3. EVM and ACPR definitions

- DACs and ADCs will not be presented in details because their performance in terms of bandwidth (up to 100 MHz) and resolution (up to 16 bits) is almost sufficient for today's radio-communications signals. The main limitations are the current consumption of fast converters and the difficulty in designing near-GHz Sigma Delta ($\Sigma\Delta$) encoders, often in the context of polar transmitters as is presented in section 4. It should be noticed that for WB-CW signals, the bandwidth limitation is the criterion of choice and a source of important spectrum degradation (approximated qualitatively by a windowing effect).

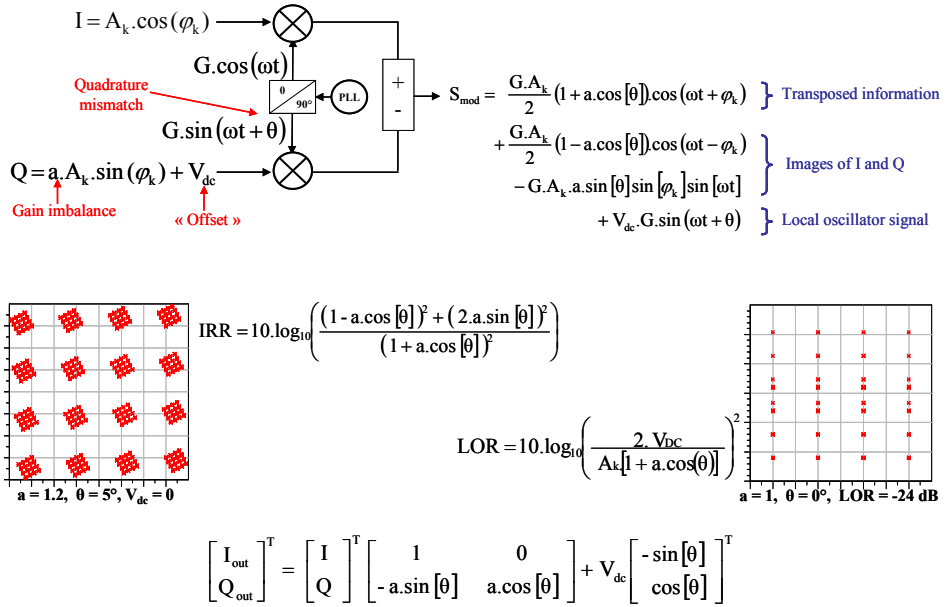


Fig. 4. IQ modulator equations and the effects of its imperfections

- The IQ modulator is the block providing the transposition of the information at high frequency (up-converter) or at baseband/intermediate frequency (down-converter). We describe the case of the up-converter for simplicity. It needs the baseband information components I and Q (I and Q channels) and a carrier frequency signal, provided by a frequency synthesizer. Key components are the non-linear multipliers provided thanks to passive (PIN diodes...) or active circuits (Gilbert cell...). Every component should be carefully designed with regards to the synthesizer frequency value. Moreover, the noise added by the multipliers will be impossible to filter. The main difficulty for the modulator is the perfect matching between I and Q paths. As is reported in Fig. 4 (at one fixed frequency), gain and phase imbalances (a and θ) create unwanted images of the information. Also the presence of an offset V_{DC} can create the emission of the synthesized carrier frequency (called Local Oscillator, LO). All of these imperfections result in the distortion of the information and are quantified by the IRR and LOR power ratio (the actual performance is in the range of -50 dB). These imperfections are not seen on the spectrum because their equivalent added noise is inside the main lobe. The biggest challenge concerns the modulation of wideband signals because it is hard for the modulator to perform and to be frequency independent (over the entire bandwidth). An example of this can be seen in the conversion gain (ratio between the RF output power and the baseband input power). A variation of this gain produces an unwanted AM on the emitted signal and can distort the information. Another problem is the image frequency. This effect is due to the multiplication of signals in a modulator. A multiplication of harmonic signals results in two signals whose frequencies are, respectively, the sum and the difference of the input frequencies. This duality is mathematically illustrated and discussed for the case of the receiver in Fig. 16, section 3. At the outgoing emission, an image is created at RF and

increases the EVM. At the reception (only for heterodyne architectures), an image frequency different from RF can be sensed and interpreted as unwanted information which increases the EVM, too. This is illustrated in section 3.3 with the presentation of the Hartley and Weaver image rejection receiver architectures for NB-CW. To conclude, the effects of the IQ modulator imperfections are: non-linearities, noise and the possibility of image frequencies. Selective filtering and stable reference signals are needed to improve the system (easier for NB-CW than for WB-CW).

- Frequency synthesizers are blocks which produce a stable reference signal for transposition. RF architectures need flexibility and stability in the choice of their reference frequency. It is usually not possible to provide this simultaneously with a simple oscillator circuit such as Quartz (stable), SAWR... The stabilization and synthesis by a Phase Lock Loop (PLL) is usually necessary. The PLL is a looped system whose different designs will not be discussed here, only its system's characteristics, which are reported in Fig. 5. The worst imperfection of the PLL is its phase noise. An example of the resulting phase noise profile of a synthesizer is illustrated in Fig. 5 (N=1). The different transfer functions of each noise of the sub-blocks are reported in the same figure. The great challenge of the PLL is to keep the stability performance of the synthesized signals for different values of N. It is also possible to modulate the signal in frequency by directly adding the baseband information on the Voltage Controlled Oscillator (VCO) input. This is called an "over the loop" modulation. For CW systems, synthesizers are used to produce the different Local Oscillators (LOs) needed for transposition(s) and for channel selection (fine tuning of the LO value). PLL can be used to transpose the information, but only for angular modulations schemes: PM, PSK, FSK or FM. This is not usually used for constant envelope WB-CW signals for noise and stability reasons.

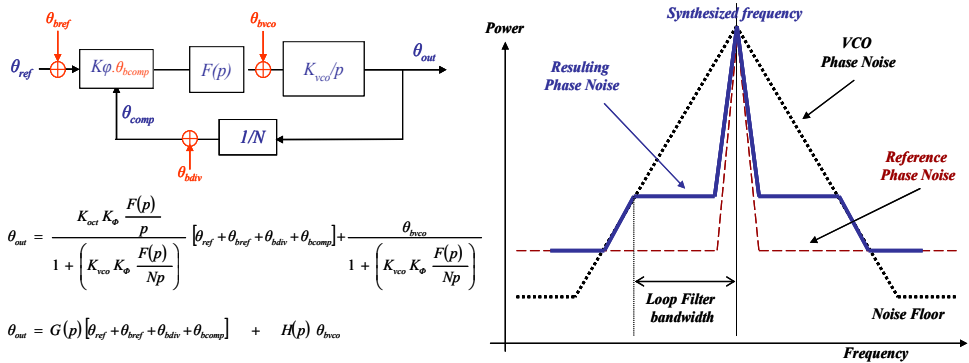


Fig. 5. PLL functional blocks and frequency synthesizer phase noise

- RF filters are essential in communications chains for information selectivity (interferences, noise, image-filtering for example). They are used at emission (limiting spectral re-growths), reception (rejecting unwanted signals) and for channel selection (high selectivity for discrimination and for image rejection). Their system characteristics are well-known: attenuation/rejection, selectivity, ripple, group delay... Different technologies are used depending on the frequency, implying different sources of imperfections.

At GHz frequencies, LC filtering is preferred for its low noise property but the components' sensitivity (especially for integrated technology) implies low order filters. For higher selectivity, active or high speed digital filters can be used, but often need a frequency transposition section due to the circuit bandwidth and the sampling rate limitations. Moreover, these filters consume power and add much noise (circuit or quantification).

Whatever the technology used, the wider the bandwidth the higher the ripple (oscillation of the transfer function). This ripple problem introduces an unwanted amplitude variation. In order to reduce this, the order is increased as much as possible and some prototype functions like Butterworth or Cauer are chosen for the design. Attenuation in the rejection band is worst in these cases. Additionally, the group delay of the filter is mandatory for non-distortion of the information. For NB-CW, it is usually not a problem but the phase response has to be linear for WB-CW and IR-UWB systems. In the case of a multi-band system, being linear in each sub-band is sufficient. These remarks point out that filters for WB-CW are more difficult to design than filters for NB-CW due to the performance limitations over the bandwidth.

To conclude, the RF filters imperfections are modeled by their noise factors (noise added, distortions and non constant group delays), insertion losses (due to the ripples and mismatches) and selectivity (finite attenuation of unwanted frequencies).

- Power Amplifiers (PA) in RF architectures are designed to linearly amplify radio-communication signals with the highest efficiency possible and with the lowest spectrum re-growths or added noise (see spectrum mask/ACPR and EVM criteria to respect). Since the active components of the amplifier are operating at maximum power, non-linear (NL) compression/conversion and memory effects are unavoidable (Fig. 6). The PA design, identified as "class of operation", impacts the performances [Villegas et al., 2007][Diet et al., 2004-2007]. Efficiency and linearity of the PA are mandatory for NB-CW and WB-CW architecture because the signal is transmitting information continuously. Low efficiency reduces battery lifetime and increases the dissipated power and the temperature of the circuit. Low linearity affects the quality of the signal. For IR-UWB systems, only peak performances in time are needed (see section 5). We are focusing upon the PA impact on the architecture in the case of the CW system. The most difficult case is for WB-CW due to the bandwidth, and the usual high PAPR of chosen modulation schemes (high data rate). A PA class of operation is determined by the hypothesis of transistor saturation (current source or switch). There are two families of PA classes: the switched mode (SW) and the continuous wave (CW) or biased mode. The different load-lines are illustrated in Fig. 6 (right).

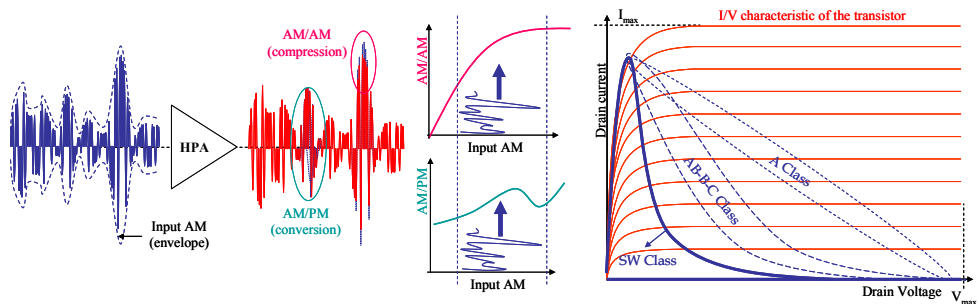


Fig. 6. AM/AM, AM/PM effects (without memory effects) and PA class load-lines

Due to the need of polarization, CW classes (namely A, B, AB and C) present lower efficiency than SW classes (D, S, E and F). SW classes need a switching of the transistor and cannot reproduce an amplitude modulation for that reason. Their linearity is also worse than in CW classes because of this switching operation. Moreover, it is possible to restore the amplitude information by adjusting the voltage supply. This is theoretically linear in the case of SW classes and there is a tracking effect for CW classes. For AM signals the average efficiency will rely on the statistical properties of the signal itself. Additionally, it is important to consider that efficiency is given as a peak value for CW classes. An improvement in efficiency is gained if saturation/clipping on the peak values is introduced in order to increase the average power of the output signal for the same power dissipated by the amplifier. WB-CW signals for high data rate applications present such a high PAPR that the amplification by a CW class PA requires a power back off (very low efficiency) or a linearization technique to reduce the NL effects of compression (AM/AM) and conversion (AM/PM), see Fig. 6. Techniques which are interesting for the designer in the case of wideband and high PAPR signal are those providing the highest efficiency of the entire architecture including the PA (see section 3). Polynomial modeling of the AM/AM can be done by the PA response to 1 or 2 frequency signals, called 1 or 2 tons. These indicators are the 1 dB compression point (P1dB) and the 3rd order Interception Point (IP3), as defined in Fig. 7.

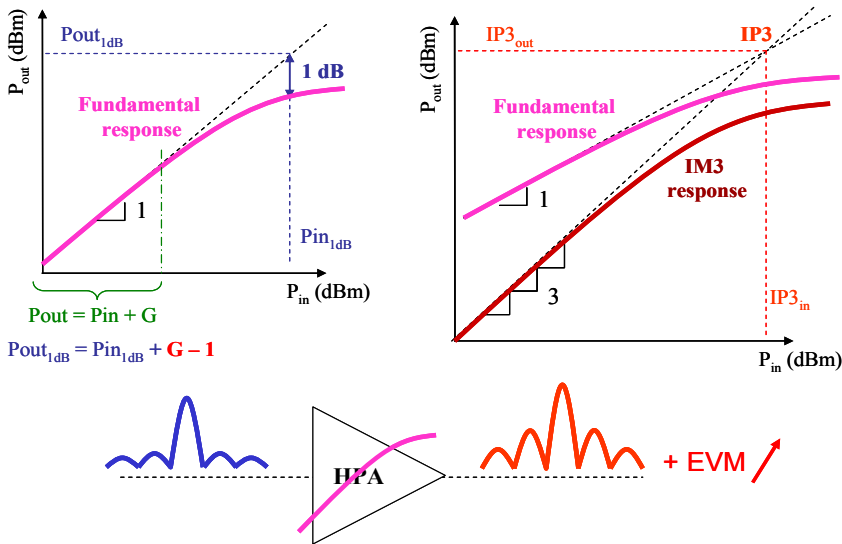


Fig. 7. Compression effect modeling for 1 ton (P1dB) and 2 tons (IP3)

These are considered to illustrate the main PA imperfections. AM/AM and AM/PM measurements, when it is possible to take them, can be the best way of characterizing the effects of the PA on the architecture: EVM and noise factor increase and spectral re-growths (ACPR, spectrum mask). In CW systems, the PA has to amplify modulated signals (a sum of several tons). For NB-CW signals, the PA behavior is well-characterized by the P1dB and the input power is usually set at this value for the linearity/efficiency trade-off (if the PAPR is not too high). In the case of WB-CW, the IP3 is a good representation of NL effects on the

spectrum if the frequency separation of the two tons is coherent with the modulated bandwidth (symbol rate frequency). Moreover, the conversion effects are, by far, more complex to analyze. It is almost impossible to determine an equivalent of the P1dB or the IP3 for AM/PM. This relies upon the influence of transistor technology. For wideband and high PAPR signals, the conversion effect increases the EVM significantly and often destroys the information.

SW classes are based on the hypothesis that the transistor switches perfectly (no power dissipated). The filtering is mandatory in RF applications, except if the SW PA is dedicated to the amplification of a square signal (class D). The load-line of every SW PA class tends to be that of an ideal switch, which is impossible in practice due to the physical realities of the transistor (resistive and capacitive output effects). Although the switching cannot be perfect, SW classes present higher efficiency than CW classes, see the 100% efficiency class E PA [Sokal & Sokal, 1975][Raab et al., 2003][Diet et al., 2008]. Moreover, they are used for AM signals only with a recombination process: supply modulation or amplification of an envelope coded signal (PWM, $\Sigma\Delta$) [Robert et al., 2009][Suarez et al., 2008]. Using supply modulation creates NL effects of compression and conversion on the output signal. These effects are named Vdd/AM and Vdd/PM [Diet et al., 2004]. For new RF architectures, high efficiency (SW) classes are preferred due to their high efficiency, but the challenge then shifts to the linearity of the amplified signals. Whatever the CW system is, NL effects of the PA are unavoidable and are sources of important imperfections (i.e., loss of linearity). In the case of high PAPR signals, a correction or a modification of the architecture is needed, as is presented in the following section.

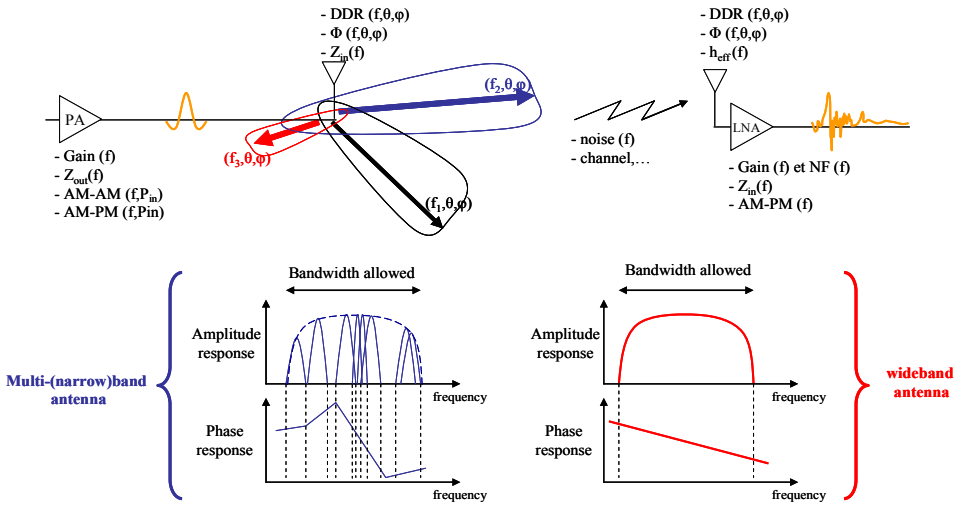


Fig. 8. Antenna system characteristics in radio-communications

- Antenna characteristics from a "system" point of view are reciprocal interfaces between electrical and radiated signals. Depending on the application, antennas have to be adapted to their environment (omni-directivity of their radiation pattern, polarization...). Fig. 8 summarizes the different system parameters of an antenna that can influence the

performance of the radio-communications link: spatial and frequency variations of the radiation pattern, bandwidth limitations and phase distortion. Some of the imperfections are equivalent to those of the filter, but they depend on the signal direction and the channel. For wideband systems, i.e. WB-CW and IR-UWB, the use of a UWB antenna is mandatory for keeping the phase information unaltered [Schantz, 2005]. For NB-CW and separated multi-band signals using a wide bandwidth (see Fig. 8), the use of a multi-band antenna is sufficient. The latter can present phase linearity only for used sub-bands [Diet et al., 2006]. To conclude, the communications channel composed of the emitting and receiving antennas and the propagation channel is the source of amplitude and phase distortion, with a statistical dependence on time and space. Noise is added and the techniques of signal protection and antenna diversity (e.g., MIMO) are exploited as much as possible to improve the system's range.

3. Architecture basics

RF architectures are adapted to radio-communications signals and provide emission and reception. Linearity improvements are needed when power amplification NAs can destroy the information. This section first focuses on the transmitter section where different major modifications are described. We especially focus on CW systems because of their important challenges about efficiency and linearity.

3.1 Classic architectures for RF transmitters

Radio-communications architecture is composed of the above-mentioned blocks in section 2 and provides the emission and reception of the signal. There exist three kinds of basic architectures, the classic ones, designed for NB-CW systems (the oldest application of which is radio broadcasting): homodyne, heterodyne and PLL modulated.

- The homodyne architecture means that the I and Q signals (the information after DACs) is transposed directly from baseband to RF see Fig. 9. If the modulation scheme is not x-QAM, there is only one frequency transposition in the transmitter. This type of architecture is the simplest combination of function blocks and theoretically requires the minimum number of components.

If the signal is transposed directly to RF, the frequency synthesizer output is at the same frequency value, and so is the HPA and the antenna. In a compact system (mobile phones for example), the spatial proximity is the cause of unavoidable coupling between the synthesizer, the HPA and the antenna. This problem of electromagnetic coupling (EMC, EM Compatibility) is that it highly distorts the signal (EVM increase). EMC effects can be reduced by circuit spatial optimization, when possible, and more efficiently with a shielding of the considered block (LO, HPA). Additionally, a ground plane acting as a reflector can be added between the antenna and these elements (HPA and/or LO synthesizer) to reduce the amount of radiated waves toward the circuit. Choosing the homodyne structure results from multiple trade-offs concerning the EVM, the simplicity and the size of the system.

- The heterodyne architecture, as represented in Fig. 9, means that the frequency transposition is achieved in two or more steps. In fact, heterodyne means that the frequency synthesizers are not at RF values. Any combination is possible, but for a minimum use of components only two transpositions are usually performed. As the frequencies are different, the coupling effect is highly reduced in this architecture. In heterodyne structures,

there are more components than in homodyne ones (with at least one filter and one additional mixer) and imperfection sources are added. For example, the phase noise is a function of the number of synthesizers. An additional mixer also means the possibility of an image frequency that can distort the emitted information if the intermediate frequency (IF) filter is not selective enough. This IF filter is traditionally an external SAW one (because of its selectivity). This increases the cost and complexity of building this system. An advantage of the heterodyne architecture is that the need of frequency flexibility for channel selection can be more easily achieved with two synthesizers than with one (homodyne case).

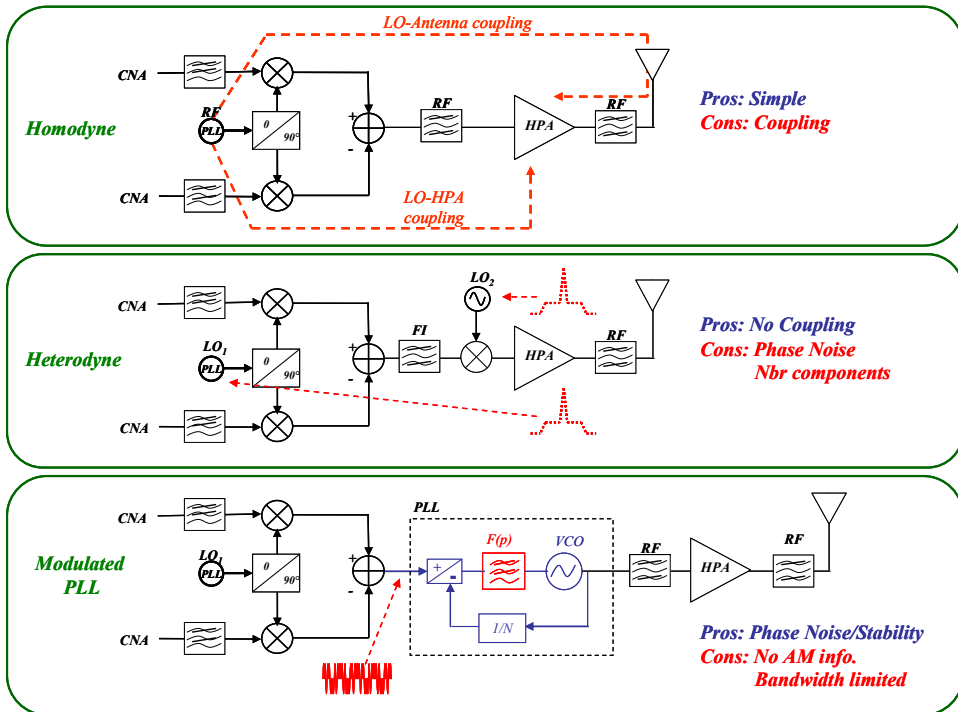


Fig. 9. Classic architectures for RF transmitters

- Architectures using modulated PLL can benefit from PLL advantages. This corresponds to the direct modulation of a synthesizer (with $N=1$). The signal of a PLL is stable and its noise profile depends on the loop filter bandwidth. It is possible to modulate the PLL by introducing small variations of the reference frequency which will be stabilised by the loop reaction. This is called a modulation "in the loop" and is used for narrow-bandwidth modulations (low symbol rate). On the contrary, the input voltage of the VCO can be directly modulated to produce a wider bandwidth modulation (modulation "over the loop") but this could affect the stability of the PLL. As is understood here, only angular modulations are possible with modulated PLL architectures. If the signal to transmit has AM information, this latter should be reintroduced after the PLL. To conclude, architectures using modulated PLL are very interesting with regards to their noise property (less noise

than an IQ modulator) and their design for non-constant envelope signals implies some important modifications of the architecture.

The classic architectures presented are widely used for NB-CW signals. While the efficiency and linearity of the transmitter is not significantly affected, these well-known structures are preferred for their simplicity. In the case of WB-CW systems, the bandwidth and the probable increase of the PAPR (due to high data rate) lower the performance of such architectures. For power amplification in particular, high PAPR values of OFDM and other multi-carrier signals cause such compression and distortion/conversion effects that the information cannot be interpreted at reception. Additionally, standard limitations are, by far, not respected. The first choice is to perform a PA back-off, but this drives it to unacceptably low values for the architecture efficiency. To achieve a linear transmitter, linearization techniques are provided. The next sub-section is dedicated to their descriptions.

3.2 Linearization techniques for the transmitter

Wireless communications require highly efficient and compact transceivers, whatever the signal characteristics are. Transmitter architecture, at worst, must meet design constraints of: providing high efficiency and linearity for a wideband and high PAPR signal (or high dynamic). Power amplification of WB-CW multi-carrier signals (WiFi, WiMAX, LTE...) introduces crippling Non-Linearities (NLs) in amplitude and in phase in the communication system. The linearization of such this kind of transmitter is mandatory. Identifying a type of architecture for such signals requires a careful study of linearization techniques and their performance. A linearization technique is beneficial only if it provides linearity with the maximum efficiency possible. There are several linearization techniques, depending on the PAPR of the signal, the added complexity and the increase in size and consumption of the system that can be accepted by the RF designers [Villegas et al., 2007]. Many criteria characterize the technique used: static/dynamic processing, adaptability, frequency (digital, baseband, IF or RF), correction of memory effects, complexity, stability, resulting efficiency, size increase... Herein, we basically classify these techniques in three types: (i) correction techniques, (ii) anticipation of NLs and (iii) those based on a decomposition and recombination of the signal, often dedicated to wideband signals.

Examples of correction techniques are (A) feed-back, (B) feed-forward and the anticipated technique of (C) pre-distortion, illustrated in Fig. 10. Their point in common is to modify the modulated signal as close as possible to the PA (before or after). The architecture considerations here, do not include the modulator nor the baseband signal processing. To linearize, we need a carefully selected NL model of the PA (Volterra series, Wiener or Saleh model...). Adaptability to the signal amplitude can be introduced in order to compensate for the model's lack of accuracy and the PA memory effects (a temperature influence can be considered) [Baudoin et al., 2007]. Each structure contains a major defect. (A) Feed-back reduces the gain of the amplification and introduces a bandwidth limitation due to the transfer function of the loop (stability and dynamic response). The feed-back can be performed on the amplitude (Polar feed-back) or on I and Q quadrature components of the signal (Cartesian feed-back) and both are dedicated to narrowband signals. (B) Feed-forward requires a significant increase in signal processing and RF blocks in the transmitter, with the hypothesis of a precise matching between NLs and reconstructed transfer functions. The improvement in linearity will be costly in terms of consumption and size

(integration criterion). The advantages are stability and the possibility to process wideband signals. The most interesting of the three techniques is (C) pre-distortion because of its flexibility: the anticipation can be done in the digital part and, by doing so, can provide adaptability of the technique if using a feed-back loop (with an additional DAC). The digital pre-distortion represents an additional and non-negligible consumption of a Digital Signal Processor (DSP) and often requires a look-up table [Jardin & Baudoin, 2007]. The signal is widened in frequency because of the non-linear law of the pre-distorter (as for NL effects of compression on the spectrum), requiring baseband and RF parts to be wideband designed. Interesting improvements of pre-distortion have been made with OFDM WB-CW signals in [Baudoin et al., 2007] [Jardin & Baudoin, 2007].

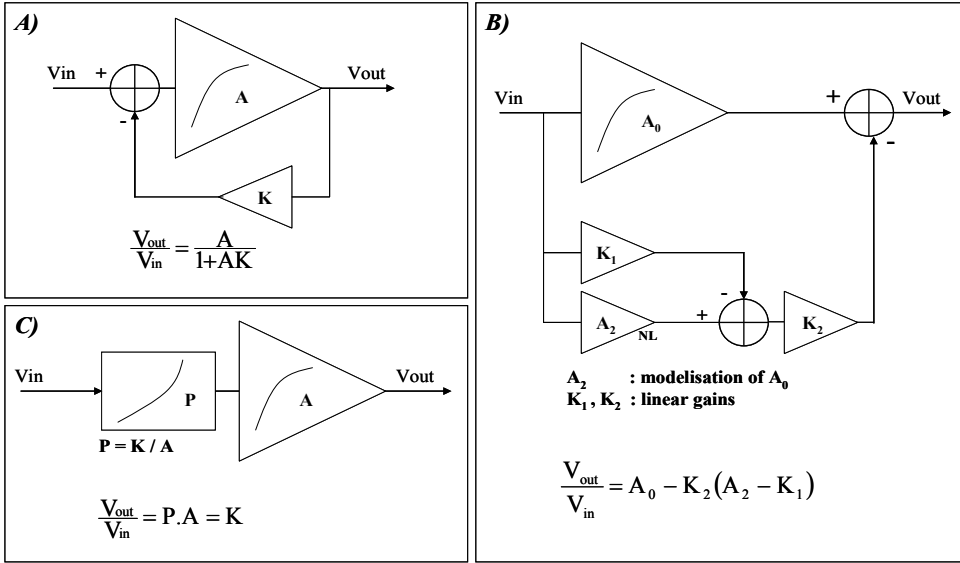


Fig. 10. Correction (A and B) and anticipation (C) linearization techniques

Other techniques presented are based on a vectorial decomposition of the signal. The goal is to drive high efficiency switched mode RF PAs with constant envelope (constant power) signals, avoiding AM/AM and AM/PM [Raab et al., 2003] [Diet et al., 2003-2004]. These techniques are used when NL effects are so great that feed-back or pre-distortion cannot sufficiently improve the linearity. We can consider the problem of linearization in the communication chain from the digital part to the antenna (front end). This drives one to completely modify the architecture and its elements' specifications in baseband, IF/RF and power RF. After the amplification of constant envelope parts of the signals, the challenge is to reintroduce the variable envelope information with lower NLs than in a direct amplification case, keeping high efficiency of the architecture. Basic examples of these techniques are the LINC (LInearization with Non-linear Components) and the EER (Envelope Elimination and Restoration) methods (and their recent evolutions) [Cox, 1974] [Kahn, 1952] [Baudoin et al., 2003-2007] [Diet et al., 2004] [Suarez et al., 2008].

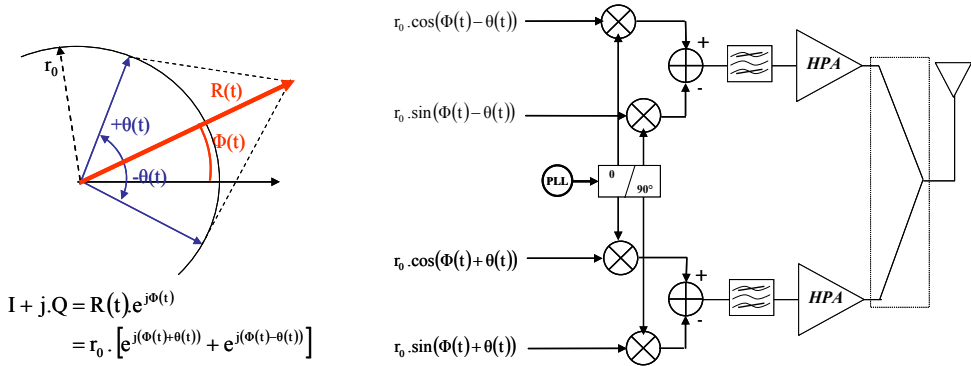


Fig. 11. LINC decomposition and recombination at RF power amplification

The LINC principle relies on a decomposition of the modulated signal into two constant envelope signals as is shown in Fig. 11. The decomposition can be computed by a Digital Signal Processor (DSP) or by combining two VCOs in quadrature PLL configuration: CALLUM. This latter configuration is an interesting architecture but presents the possibility of instability and additional manufacturing costs. The amplification of the two constant envelope signals implies the design of two identical HPAs at RF frequency, and this often causes signal distortion due to imbalance mismatches. Also the HPA must be wideband because the signal decomposition is a non-linear process (widening of the spectrum), and the phase modulation index is increased. Whatever the decomposition technique is (LINC/CALLUM), the problem is that the efficiency is directly determined by the recombination step: a sum of the powers. It is very difficult to avoid losses at RF while designing a RF power combiner.

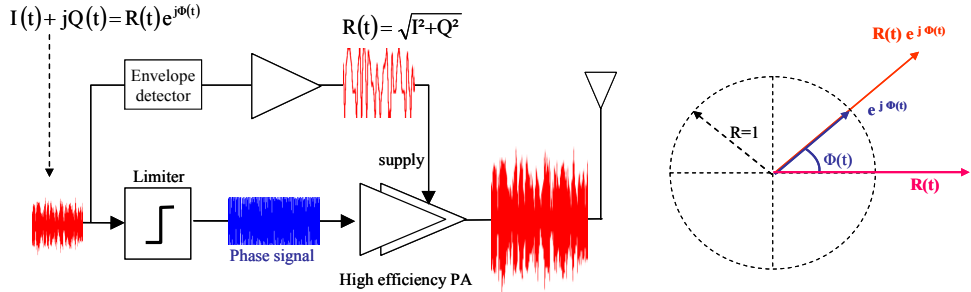


Fig. 12. Principle of the EER technique [Kahn, 1952]

Another decomposition technique was proposed by Kahn in 1952 and this is basically an amplitude and phase separation technique (polar): Envelope Elimination and Restoration (EER). This method was first proposed for AM signals as represented in Fig. 12. The advantage of EER is that it drives the RF PA with a constant envelope modulated signal (carrying the phase information), enabling the use of a SW high efficiency amplifier [Raab et al. 2003] [Sokal & Sokal, 1975] [Diet et al., 2005-2008]. The difficulty is to reintroduce the amplitude information linearly using the variations of the PA voltage supply. This implies a

power amplification of the envelope signal at a frequency equal to the symbol rate (lower than RF). The recombination can be done with a SW class PA because the output voltage is linearly dependant on the voltage supply for this PA mode. Two difficulties are to be considered in such a linearization technique: (i) synchronization between the phase and the amplitude information and (ii) linear and efficient amplification of the amplitude before the recombination (directly impacting the overall efficiency), as reported in [Diet, 2003-2005]. Recently, a lot of work has been done on the EER based architectures, often named as “polar” ones [Nielsen & Larsen, 2007] [Choi et al., 2007] [Suarez et al., 2008] [Diet et al., 2008-2009]. The generation of the amplitude and phase components can be expected to be done digitally thanks to the power of DSPs, as is shown in Fig. 13. As was previously discussed in [Diet et al., 2003], the bandwidths of the envelope and phase signals are widened due to NL processing and make it necessary to design the circuit for three to four times the symbol rate (as for LINC or any other NL decomposition). Fortunately, a clipping in frequency and on the envelope is possible, increasing the EVM and ACPR under acceptable levels. These polar architectures are suited for new high data rate standards where efficiency of the emitter and linearization are mandatory. Also, the multi-standard and multi-radio concepts have helped polar architectures to evolve in multiple ways. For example, recombination on the PA’s input signal is possible because the amplitude information can modulate the phase signal (RF) and can be restored by the band-pass shape function of the following blocks: PA + emission filter + antenna. The emitted spectrum is the criterion of quality to be considered carefully, because the PWM or $\Sigma\Delta$ envelope coding are the source of useless and crippling spectral re-growths. The efficiency is also penalized by the power amplification of such frequency components, but this is counter-balanced by the advantages of high flexibility of this architecture [Robert et al. 2009]. The digitally controlled PA and the mixed-mode digital to RF converter are key parameters in the evolution of these polar architectures [Suarez et al., 2008] [Robert et al., 2009] [Diet et al., 2008].

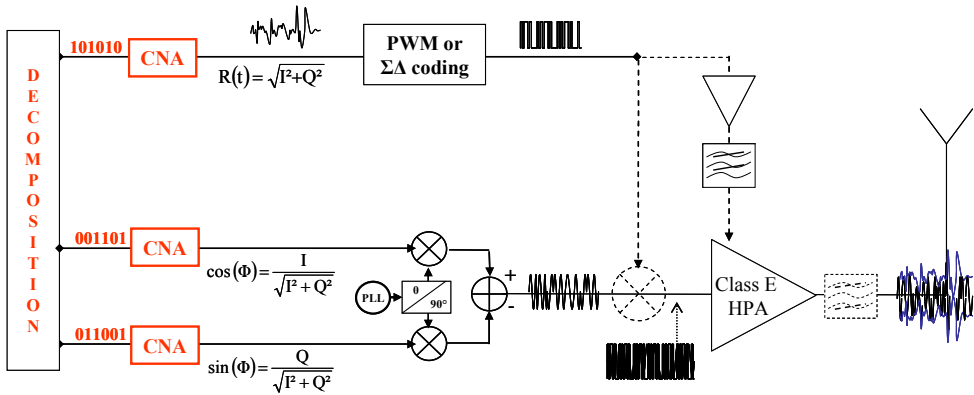


Fig. 13. Recent improvements of EER/ polar architectures for wideband OFDM signals

We have presented the main linearization techniques. Actual needs, in terms of transmitter linearity and efficiency for high data rate applications, have caused the RF designer to consider RF architectures based on combined techniques. For example, digital pre-distortion is an improvement on polar based architectures, as is shown in section 4. Another example

is described in [Diet et al., 2008] where a combination of cascaded EER and LINC techniques can theoretically provide an architecture which cancels the PAPR influence, see Fig. 14.

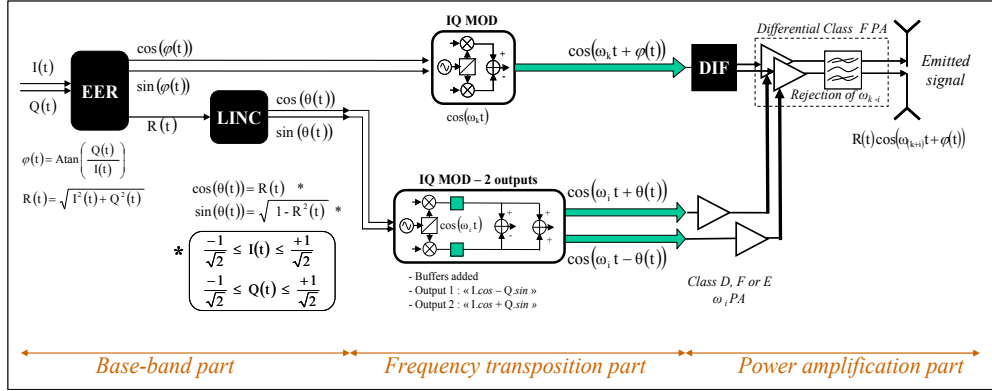


Fig. 14. EER-LINC method for high PAPR signals

This architecture needs some new components (e.g., a balanced switched PA with frequency transposition). This also represents an increase in complexity, but proposes new ways to improve WB-CW signals transmitters. In EER-LINC, the envelope information is converted to two angular modulated signals. The RF balanced PA (also called differential PA) is in SW class and is supply-modulated with the two above-mentioned signals. The fact that the PA is differentially supplied allows for the combining operation and a transposition at the same time (due to the multiplication). As the antenna is supposed to be differential, there is no balun after the PA.

This review of linearization techniques reveals that the different parts of an efficient and linear transmitter cannot be designed separately: baseband, frequency transposition and power amplification (and antenna for wideband systems). The modification of the architecture for global performance improvements must be done, considering each block's impact, digital or analog (and their imperfections). To conclude this theoretical sub-section, linearized architectures are mandatory for the major part of actual and future CW systems, which are WB-CW (high data rate). The highest performance can be reached if a combination of different techniques is exploited: pre-distortion and EER seem to be the most popular.

3.3 Receiver architectures

The challenges of the receiver architectures are the noise level, the presence of other channels and blocking/interference signals. This is summarized by the immunity and the coexistence of the standards, illustrated in Fig. 15.

The information received is at such a low level that other RF emissions can mask it. It is easily possible to saturate the receiver if a signal is too strong in the vicinity of the spectrum. Other characteristics of the receiver are its sensitivity and selectivity. The sensitivity is the lowest level of power that can be received and demodulated correctly (providing a Bit Error Rate sufficient for interpretation). This latter, is altered by the total amount of noise added by the receiver itself. The noise factor (F) expresses an equivalent of a white Gaussian noise

addition, for each block. It is possible to compute the global noise factor for the receiver thanks to Friis formula, see Fig. 15. This formula shows that amplifying the signal as close as possible to the antenna, at the front-end, is better for minimizing the total added noise. This is the goal of the Low Noise Amplifier (LNA) whose design is optimized for noise and not for gain or output power performance. Unfortunately, the presence of high power signals (blocking) often implies starting the reception chain with a selective band-pass filter, as is reported for the Rx architectures in Fig. 15. At this point, the signal bandwidth is another limiting factor for the design of the receiver because it directly impacts the reception filter design and the LNA. The selectivity and sensitivity performance are also affected. At reception, the decrease in the signal quality is expressed by the EVM and the BER. This degradation results from the noise in the system (noise factor) but other imperfections are due to the architecture design: block interactions, CEM, image frequency and so on. In digital radio-communications, the channel coding (block and convolution types) improves the robustness of the system against the noise, but this point will not be developed here.

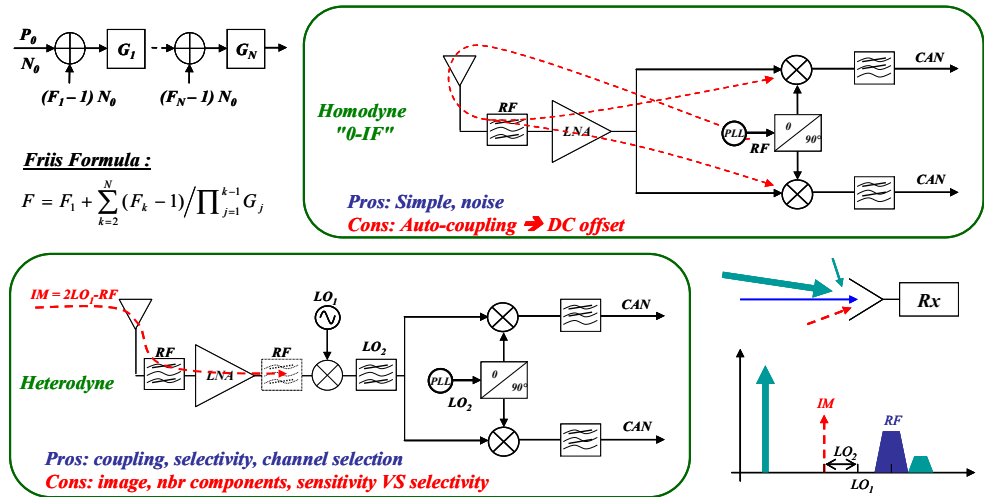


Fig. 15. Considerations for receiver architectures

The topic of this sub-section is to present the main types of receiver architectures. As there is no PA NL effects (Tx architecture), there is no need for linearization techniques. If the receiver is saturated, it suffices to reduce the gain of the LNA (if possible) or to attenuate the signal (back-off). The basic receiver structures are similar to that of the classic transmitter structures, that is to say homodyne and heterodyne types as represented in Fig. 15.

- Homodyne receivers are composed of a direct IQ demodulator which comes after the LNA. RF filters, before and after the LNA, tend to avoid a masking effect by other signals received. The filter technology for RF is subject to a trade-off between the selectivity and the losses, which increase the noise factor of the filter. Swapping the LNA and the RF filter in the receiver chain creates a limitation of the received band. This is set in function with disturbing signals and the resulting noise factor (sensitivity). This architecture is simple, as it requires few components, and theoretically limits the total amount of noise added. The

transposition needs a Local Oscillator (LO) signal that is stronger than the received signal. The coupling effects between the Rx antenna and the frequency synthesizer is very important and a shielding is mandatory to reduce CEM effects. Moreover, if a part of the LO signal is sensed by the antenna, it creates an offset (DC) by auto-mixing of the RF signal. This DC component may saturate the ADCs and is added to the baseband signal. The sensitivity of the homodyne receiver is not as low as expected due to this coupling effect whose impact is greater than that for Tx architecture.

To summarize, homodyne architecture is very attractive for its simplicity if the integration of the synthesizer is not causing crippling CEM effects. The addition of RF filtering can reduce the offset but reduces the receiver sensitivity (noise factor increased).

- Heterodyne receivers need two (or more) transpositions of the signal. The number of components is important and represents an additional current consumption compared to the homodyne structure (simpler). The coupling effect due to the strong LO_1 and LO_2 signals, see Fig. 15, is minimized due to the different frequency values. This architecture needs more filters than the homodyne one but this can be useful for improving the receiver selectivity: Intermediate Frequency (IF) filters can reach higher selectivity than in RF. Moreover, additional filters inevitably decrease the signal-to-noise ratio, as well as the receiver performance. Sensitivity and selectivity are the sources of the challenges faced by the receiver designer. Additionally, the mixers' operation frequency and the IF filters enable different manufacturing technologies. A great challenge of the heterodyne structure is the possibility of receiving unwanted image information. An image is a signal producing an IF signal after the first mixer ($IF = LO_2 = LO_1 - RF$, as reported in Fig. 15). A signal whose frequency is " $IM = 2LO_1 - RF$ " produces, after mixing, an output signal whose frequency is " $LO_1 - RF = IF$ ". The receiver should reject this IM component in order to protect the receiver from this perturbation. Without selectivity, the receiver cannot attenuate IM components. If the addition of a filter after the antenna is not sufficient to do this, an improvement is possible with "image-rejection" architectures. The principle here is to eliminate the image signal by an addition of 180° phase shifted copy. The considered improvements were proposed by Hartley and Weaver. This is possible if the total power level of received signals (information at RF and image at IM) are not saturating the receiver. An RF filter is mandatory even if these architectures are used.

Hartley and Weaver improvements are called "image-rejection" (heterodyne) receiver architectures. The different structures are summarized in Fig. 16.

As is reported, these structures are supposed to cancel the received signal whose frequency value is IM. The Hartley structure uses an all-pass filter with a phase shift of 90° . For NB-CW systems, this filter can be realized by R-C structures. For WB-CW systems, a filter with such frequency independent characteristics is called a Hilbert filter. The realization of this filter is a source of imperfections which directly impact the image rejection property. In order to avoid filter design difficulties, Weaver proposed a structure in which the 90° phase shift is achieved by a second frequency transposition. This heterodyne structure can directly output the baseband signal (information), as expressed in Fig. 16. The cost of this improvement is the number of added components, which is almost twice as much when compared to a classic heterodyne structure. If the signal is IQ modulated, it is possible to create quadrature image rejection architecture by using two more mixers, as is illustrated by the IQ Weaver schematic in Fig. 16. The increase in size, complexity and consumption is balanced by the image rejection property, if IQ mismatches are low enough. Hartley,

Weaver and IQ Weaver were first designed for NB-CW systems. Their use for WB-CW can be discussed if IQ modulator performance and filters enables it. The bandwidth of the filter implies the choice of LO values. Some receivers, whose required bandwidths are too wide, are not possible to design. "Image-rejection" architectures are an interesting alternative to reduce selectivity constraints in the receiver design, at the cost of additional components and additional noise (lower sensitivity).

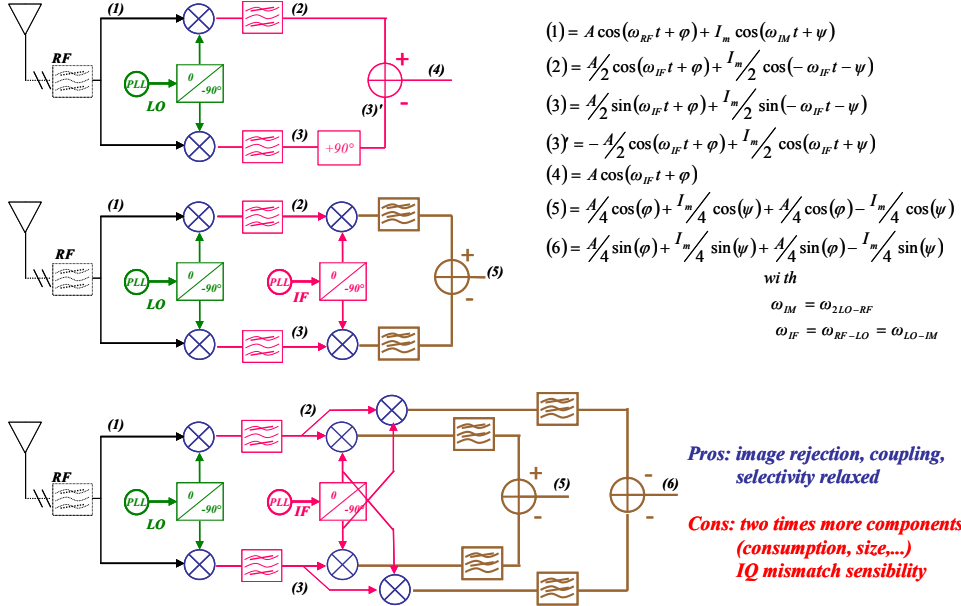


Fig. 16. Hartley (top), Weaver (middle) and IQ Weaver (bottom) architectures.

Another interesting improvement of heterodyne architecture is a group of Low-Intermediate Frequency (low-IF) receivers whose goal is to use ADCs at IF and not in baseband. There are two main possibilities for signal processing at IF frequencies: (i) using a poly-phase filter in a modified IQ Weaver architecture or (ii) using fast ADCs with digital signal processing on the baseband information. These ADCs imply a large increase in current consumption in function with the IF value. Nowadays, solution (ii) is more popular due to the state-of-the-art ADC performance (resolution versus sampling rate) and the high improvement possibilities provided by digital algorithms. Low-IF architectures with band-pass ADCs represent a sub-group of heterodyne architectures and are popular for its potential flexibility.

To conclude this section in the case of CW systems, there are two approaches for receiver architectures: (i) homodyne (also called "zero-IF") with limitations of LO coupling effects and (ii) low-IF heterodyne structures, with a growing interest in IF ADCs.

4. High data rate state-of-the-art transmitters

4.1 Front-end considerations

Many new standards for high data rate communications have appeared recently or are under development whether in the frequency range below 6 GHz or in the millimeter wave range (60 GHz radio in particular). In the first case, the data rates are in the range of several tens to several hundreds of Mbps and in the second case they can be in the range of several Gbps. There are many challenges for the design of high data rate RF Transceivers. Among the most critical are:

- Designing power transmitters with a good linearity and efficiency for wideband, high PAPR signals with a large range of necessary transmit power control.
- Designing mobile transceivers using OFDM and multi-antenna MIMO and smart antenna techniques in order to achieve very high performance (throughput and BER) in mobile channels with large delay spreads while maintaining low power consumption and reduced size and cost.
- Designing flexible and scalable devices able to accommodate for many standards, frequency bands and modes

Multi-carrier (OFDM and OFDMA for multiple user access) and multiple-antenna (MIMO) techniques have emerged as enabling technologies for beyond 3G and 4G high data rate communication systems. Many new standards use OFDM and MIMO approaches. Examples include: IEEE 802.11n for wireless local area networks, IEEE 802.16d and IEEE 802.16e fixed and mobile WiMAX or IEEE 802.20 for wireless metropolitan area networks, IEEE 802.22 for wireless regional networks, 3GPP LTE (Long Term Evolution) for beyond 3G cellular networks. OFDM and MIMO have also generated new challenges in term of transmitter architectures with good efficiency and linearity, and in term of integration of MIMO transceiver and antennas in mobile user terminals. One of the main advantages of OFDM in mobile wireless channels is the simplification of the channel equalization in comparison with single carrier modulation. Indeed, for an OFDM modulation with N carriers, the original high data rate M-QAM data stream with a rate R is split into N parallel lower speed streams with a rate R/N that modulate one of the N carriers with a M-QAM modulation. In practice, the N parallel single carrier modulation is achieved thanks to a baseband IFFT. For a given data rate R , the bandwidth of the M-QAM/OFDM signal is similar to that of a single carrier M-QAM signal. An OFDM symbol comprises N QAM original symbols. A guard interval (GI) is introduced between successive OFDM symbols in order to prevent inter-symbol interference. The duration of the GI is of the order of the delay spread of the impulse response of the channel. Therefore, the larger the number of carriers, the more efficient the system. Unfortunately, the PAPR value of the modulated signal also depends on the number of carriers. The larger the number of carriers, the larger the PAPR value, and the more difficult it is to design a power transmitter with a high efficiency and a high linearity. In order to accommodate for multiple users, a set of carriers can be allocated to each user resulting in the OFDMA principle. The LTE standard proposes an uplink (mobile emitter) variant called single-carrier FDMA (SC-FDMA) which has a smaller PAPR value. In a SC-FDMA emitter, the OFDM modulator is preceded by a DFT, which increases the baseband complexity.

MIMO technology uses multiple antennas at the transmitter and the receiver. The obtained diversity and spatial multiplexing allow for better BER performance and increased data rate or link range without increasing the bandwidth or the transmitted power. In comparison

with a SISO system (Single antenna at the emitter and at the receiver) the maximal achievable increase of the data rate depends on the minimum number of antennas in the transmitter and the receiver. For example, for 2 Tx and 2 RX antennas, the data rate can be multiplied by 2 at the maximum. There is a compromise between the diversity gain and the multiplexing gain. In the WiMAX standard, different modes are defined depending on whether one wants to increase the diversity (Space time bloc code STBC approach referred to as MIMO matrix A) or the spatial multiplexing (referred to as MIMO matrix B). It is possible to use MISO approach (multiple Tx and single Rx antennas) and use transmit diversity with space time coding (called MIMO matrix A). WiMAX also includes possibility for uplink collaborative MIMO technique. This is intended to allow for 2 separate user devices, each with a single transmit antenna, to communicate on the same frequency with a base station using 2 antennas. The MIMO approach can be associated with beam-forming to control the direction and shape of the radiation pattern.

MIMO techniques allow for tremendous improvements in throughput performances. But it is a challenge to integrate many antennas and transceivers in a small mobile user device. With a multiple antenna base station and a single antenna user device it is possible to achieve some improvement but a multiple antenna user device is necessary to really take advantage of MIMO technique. In the "WiMAX Wave 2" device certification, 2 receive antenna systems are mandatory. Another kind of challenge for high data rate front-end transmitters is the case of millimeter wave (mmWave) mobile communications. At 60 GHz, the available unlicensed bandwidth is very large and the mmWave technology is a good candidate for indoor very high data rate (several Gbits/sec) communication systems. The high data path loss allows for high frequency re-use. The small wave-length allows for the use of very small antennas and integrating multiple antennas for MIMO or beam-forming approaches can be easily achieved. The standard IEEE 802.15.3c is intended to provide Gbits data rate at distance of the order of a few meters. For fixed equipments, the situation at mmWave is quite similar to the preceding one (at frequency below 6 GHz) and OFDM technique is widely used. But the design of mobile devices with low power consumption constraint is still very challenging. The OFDM approach with high speed DAC/ADC and baseband processors has generally crippling power consumption. Among the challenges are: design of PAs with a sufficient output power, efficiency and linearity; power consumption of DAC and ADC with Gigahertz sampling frequencies; design of a physical layer with a low complexity and sufficient performance. Two approaches are commonly proposed: single carrier QPSK and UWB techniques.

4.2 Polar architectures

One of the main challenges for OFDM based mobile transmitters is to design power transmitters with a good linearity and efficiency for wideband, high PAPR signals with a large range of transmit power control. For example, for the WiMAX mobile standard:

- The signal bandwidth is scalable up to 10 MHz.
- For a full OFDM WIMAX signal with a 1024 FFT and a 16-QAM mapping, the PAPR (or effective power ratio with a probability of 10^{-3}) is approximately 12 dB.
- The Peak transmit power is typically 23 dBm for subscriber terminals.
- The power control of the transmitted signal (TPC) that compensates for variations in signal strength (e.g., distance variation) must be monotonic and able to cover a range of at least 45 dB by steps of 1 dB with a relative accuracy of 0.5 dB.

For common PAs operating in CW classes (A, AB, B or C), the power efficiency is better for high power output values (close to the P1dB) than for low output values while the linearity is usually better for small power values. In order to transmit a modulated signal with a good linearity, the maximum instantaneous power of the signal must be kept smaller than the P1dB. Therefore, for a given PAPR, the average input power is PAPR dB below P1dB. The amplifier is operated with a back-off depending on the value of the PAPR and for large PAPR values; the efficiency can be very small. The common efficiency obtained with class AB PA for WiMAX signal is typically smaller than 20% while it would be at least twice bigger for constant envelope signals such as GSM signals.

Another parameter to take into account is the average output power. For the same PAPR, the power amplifier efficiency also depends on the average output power. The power supply of the PA should be adjusted in order to take into account the desired average output power and keep a constant efficiency on a large range of average output power.

Different kinds of polar architectures have been proposed as candidate solutions for high PAPR management or for adjustment of power supply to the average power. The terminology is not very stable. But, one can distinguish:

- Polar lite architectures.
- Polar feedback loop.
- Dynamic power supply and amplifier gain control with dynamic biasing, envelope or power tracking for linear PA, drain (collector) modulation for non-linear PA.
- EER and Sampled-EER architectures.

Whatever the modulation, the complex envelope z of the modulated signals can be expressed by its cartesian coordinates: real and imaginary parts usually called quadrature or I and Q components or by its polar coordinates: amplitude ρ and phase ϕ , see (1). As the amplitude ρ and phase ϕ are obtained by very non-linear operation from I and Q , their bandwidth is much higher than that of I and Q , as discussed in section 3.

$$z(t) = I(t) + jQ(t) = \rho(t)\exp(j\phi(t)), \quad \rho(t) = \sqrt{I(t)^2 + Q(t)^2}, \quad \tan(\phi(t)) = \frac{Q(t)}{I(t)}. \quad (1)$$

When the amplitude $\rho(t)$ is constant, the polar decomposition is interesting since only one signal ($\phi(t)$, or its time derivative) has to be digital to analog converted. GSM Transmitters can take advantage of this characteristic. As seen in section 3.1, the modulation of the frequency synthesis loop is a common GSM architecture that benefits of the good noise floor of the VCO and suppresses the need for external filtering. Since the envelope of the modulated signal is constant, the PA can operate in saturated CW or SW class (high efficiency). An average power controller must be added. This architecture is also called **translational (or tracking offset) loop**. The modulation of the PLL becomes very difficult for large bandwidth signals such as WCDMA. Therefore this approach is mostly used for GSM/EDGE signals. The name "polar transmitter" is sometimes limited to architectures in which the phase/frequency modulation is applied directly to the RF carrier by a modified PLL, using different techniques such as a "2-point modulation" [Durdodt 2001]. But we will use the adjective "polar" whenever the signal is decomposed in polar coordinates.

The translational loop technique was extended for EDGE-GSM signals using the so-called **polar lite architecture**. The 8-PSK EDGE-GSM modulated signal is decomposed in polar coordinates. The idea is to keep the benefits of the translational loop architecture but to introduce an AM capability. In such architecture, the signal is decomposed in polar coordinates. The phase modulates the loop and the amplitude multiplies the output of the

modulated loop by controlling the gain of a high dynamic range variable-gain amplifier (VGA). Since the signal at the input of the PA is envelope-varying, this architecture does not support a saturated PA. It uses a linear PA and has no particular efficiency advantage. It is sometimes associated with some dynamic power supply technique in order to increase the efficiency. An example is given in [Staszewski et al 2005].

The **polar feedback loop architecture** is a derivation of the polar lite architecture with the advantage of using a saturated PA. In the polar feedback loop, the PA is fed with a constant envelope signal and is modulated in amplitude. A feedback path takes a portion of the PA output signal (with a coupler) and an error signal is calculated between the ideal and actual output. The phase of the error drives the input of the PA and the error magnitude is used for the amplitude modulation of the PA. The feedback loops provides some linearization to the transmitter. But as in any feedback loop, the stability constraint limits the possible bandwidth. An example of a polar feedback loop architecture for GSM and EDGE is given in [Sowlati et al, 2004]. In that example, the PA efficiency is 54% in GSM mode at 33dBm output power and 37% in EDGE mode at 27dBm output power.

Many techniques are possible for **dynamic power supply**: dynamic biasing (at gate or drain), envelope or power tracking for linear PA, drain (collector) modulation for non-linear PA. Dynamic power supply using DC-DC switching regulators are interesting in terms of efficiency but they are still limited in term of signal bandwidth to a few tens of MHz. For envelope tracking (ET), the DC-DC is fed with the amplitude of the modulated signal (Fig. 17). ET was not presented in section 3 because it is not strictly a decomposition technique of linearization, but an improvement of the PA output power (and consequently the efficiency). This optimisation of PA biasing can increase of more than 40% the efficiency of a class A PA. In some standards the power-control dynamic range must be very wide (i.e., 80 dB for WCDMA). The amplitude modulation (envelope restoration) can be applied on a VGA or directly on the PA which leads to a better efficiency.

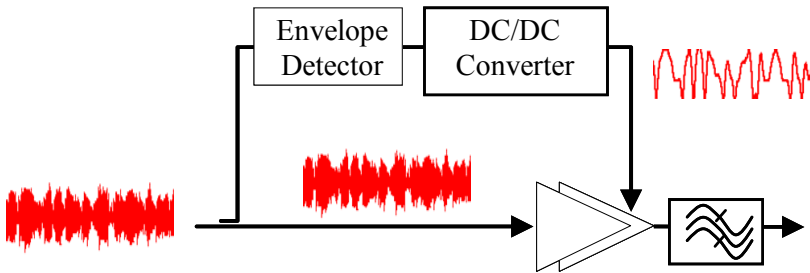


Fig. 17. Envelope tracking

The principle and interest of **EER architectures** and their variants under different names such as polar architectures were explained in section 3.2. The critical points for this architecture are: (i) the time mismatch between the envelope and the RF phase signals, (ii) the NL of the envelope restoration, (iii) the distortion caused by difficulty of biasing the PA when the amplitude of the envelope signal is very small, (iv) the leakage of the RF PA input to the output, (v) the influence of the necessary limitation of the frequency bandwidth of the envelope and phase signals. It can be shown [Baudoin et al., 2003] that the signal to noise ratio due to a time mismatch τ between these two paths is in a first approximation inversely

proportional to τ and to the bandwidth of the envelope signal. Therefore, the wider the signal bandwidth, the smaller should be the time mismatch. For an OFDM signal, such as in WiFi, the time mismatch τ must be kept smaller than typically 2 ns (which is a small percentage of the QAM symbol duration) in order to fulfil the specifications of the standard. The time mismatch can be corrected by adaptive techniques and the NL of the envelope restoration can be compensated by adaptive pre-distortion techniques. But, as already stated for dynamic power supply, it is difficult to design DC-DC converters with a bandwidth superior to a few tens of MHz. This is all the more critical as the envelope signal bandwidth is wider than the original modulated signal bandwidth. For example, an OFDM modulated WiFi signal has a frequency bandwidth close to 20 MHz. But the bandwidths of its envelope and phase signals cannot be filtered to a bandwidth smaller than respectively 40 MHz and 100 MHz if one wants to meet the specifications of the standard. The polar EER approach can be illustrated by *Tropian's Timestar* (TM) RF IC supporting GSM/GPRS, EDGE and WCDMA signals [Wendell et al., 2003] [McCune et al., 2005] or *Sequoia Communications Inc. SEQ7400* chip that supports HSDPA, GSM and EDGE and WCDMA [Groe et al., 2007-2008].

In polar architectures, the recombination of amplitude and phase signals can be done whether by PA amplitude modulation (EER architecture) or before the PA input. We will now consider the second approach, and we will focus on architectures using sampled signals and switched RF amplifiers (typically class D, E, F and their variants) [Jeong et al., 2007] [Hibon et al., 2005] [Berland et al., 2006] [Nielsen et al., 2007]. We will call these architectures "polar sampled architectures". The major motivations for using sampled signals and switched amplifiers are the ease of integration of digital circuits and the very high theoretical efficiency of switched PA. In polar sampled architectures, the envelope signal is "sampled" (converted/coded) by a pulse width modulation (PWM) or a 1-bit Sigma-Delta ($\Sigma\Delta$) modulator with bipolar output $\pm A$ [Murmann et al., 2007]. This two level signal multiplies the RF phase modulated signal before the switched PA. The result is a constant envelope signal. The switched RF PA is fed with this constant envelope signal that controls the switching of the PA. The output of the PA must be filtered in order to suppress the PWM or $\Sigma\Delta$ noise and to recover the modulated signal. Fig. 18 illustrates this principle.

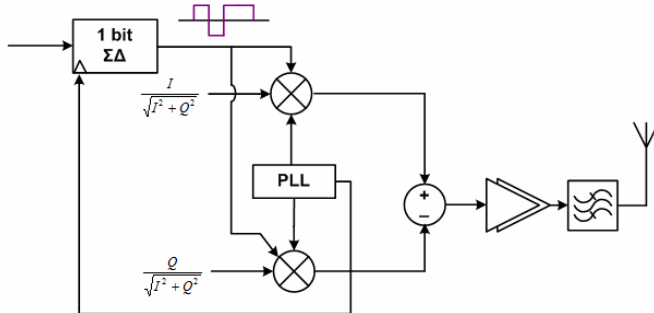


Fig. 18. Polar sampled architecture with envelope-phase recombination before the PA

In comparison with direct $\Sigma\Delta$ architectures [Rode et al 2003] in which the modulated RF signal is directly covered by a 1-bit $\Sigma\Delta$ coder before the switched amplifier, the coding of the envelope signal is interesting because it allows using smaller clock rates for the $\Sigma\Delta$ coder. One difference with EER architecture is the position and type of the filter used to eliminate

the noise of the $\Sigma\Delta$ or PWM modulator. As illustrated in Fig. 13, EER architectures use a low-pass filter at the output of the envelope BF amplifier to recover the envelope signal. But when the recombination of envelope and phase signals is done before the RF PA, the noise must be eliminated after the RF PA with a band-pass RF filter (Fig. 18). For a given standard, this RF band-pass filter should be unique and correspond to the full uplink bandwidth. It should not be specific to a given channel. Therefore the over-sampling ratio of the $\Sigma\Delta$ modulator must be calculated in reference to the full uplink bandwidth and not to channel bandwidth (bandwidth of the modulated signal). In [Andia et al., 2008] the possibility of using BW filters for WiMAX standard has been studied.

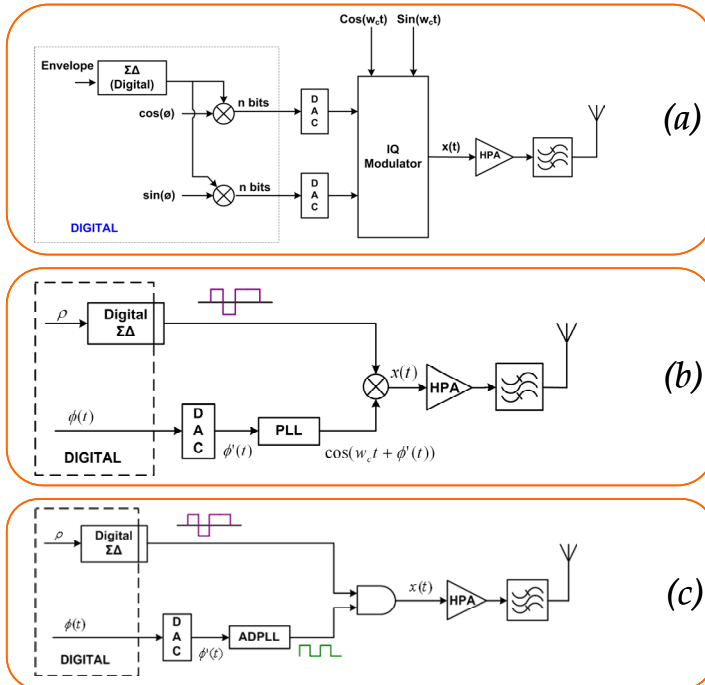


Fig. 19. (a) Polar Sampled Architecture, PSA, with 2 DACs. (b) PSA with a single DAC and an analog mixer. (c) PSA with a single DAC and with digital mixing.

Different structures have been proposed for polar sampled architectures [Suarez et al., 2008], see Fig. 19. In the structure (a) of this figure, $\Sigma\Delta$ modulator output is digital, as well as $I(t)$ and $Q(t)$ ($\rho\cos(\phi)$ and $\rho\sin(\phi)$). Therefore, two DACs are necessary before the IQ modulator. DACs sampling frequency is chosen according to the $\Sigma\Delta$ frequency. It has to be high enough to avoid $\Sigma\Delta$ noise overlapping. Targeted communication standards require high Sigma-Delta frequencies and therefore significant sampling frequency for DACs. In the structure (b), envelope and phase signals ($\rho(t)$ and $\phi(t)$) are calculated and processed independently. The output of the low-pass sigma-delta modulator is analog. The digital phase signal is converted by a DAC and then modulated to the carrier frequency (f_c). Finally constant envelope and phase signals are recombined.

The advantage of this approach compared to the first one is that it only requires one DAC. Furthermore, DAC frequency requirements are mitigated. In the first architecture, modulation to the carrier frequency is performed by the IQ modulator bloc with help of an analog mixer. Similarly, this second architecture makes use of an analog multiplier to recombine envelope and phase. The third architecture, structure (c), uses a digital carrier and replaces the analog mixer by a digital one, like an AND gate for example. In this case, an All-Digital Phase Locked Loop (ADPLL) whose input is the phase signal is used. The digital mixing is advantageous because it offers all the typical advantages of a digital signal treatment and digital IC integration but it introduces harmonics of the carrier frequency in the output spectrum with replication of the spectrum every $3 \cdot f_c$ (Fig. 20). If digital mixing is implemented in a polar $\Sigma\Delta$ architecture, frequency ratio between carrier frequency and sigma-delta frequency has to be carefully chosen to avoid $\Sigma\Delta$ noise overlapping.

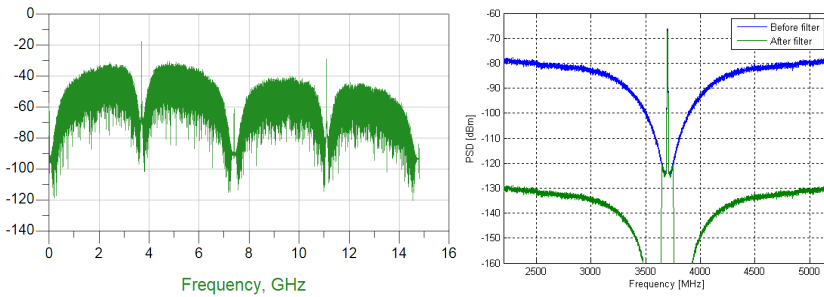


Fig. 20. Power spectral density of the signal at the PA input for a PSA with digital mixing (right), WiMAX case. Zoom on the power spectral density before and after the RF filter (left)

Fig. 20 corresponds to the third architecture, (c) in Fig. 19, and gives the power spectral density (PSD) of the signal at the input of the PA in the case of a WiMAX signal with a 3.7 GHz carrier and for a second order $\Sigma\Delta$ modulator, using a clock equal to 3.7 GHz. Fig. 20 also shows a zoom on the PSD before and after the RF filter. In that case the considered filter bandwidth is 100 MHz (while the channel bandwidth is 10 MHz) and the over-sampling ratio is equal to $(3.7 \text{ GHz})/(200 \text{ MHz})=18$.

4.3 Digital architectures

The performance of ADCs and processors, and research activities on RF-DACs (e.g., Sigma Delta) argue for integrating more and more functions in the digital part of a transmitter. Based on the software radio concept, the actual idea is to “digitalize” the design of some basic blocks such as the modulator, DACs RF and PA (D-PAs). Brand new cellular devices lead to the use of two or three different standards in the same mobile platform. As explained in a previous part, it requires stringent spectrum performance under very high speed data treatments. Each element of the transmitter must be designed with regards with several parameters: (high) frequency, dynamic power control and consumption (leading to high efficiency). Due to the use of Sigma Delta modulators, WB-CW standards such as WiMAX 3.5 and 5.8 GHz need an over-sampling frequency in the range of 15 GHz (more than four times the carrier frequency).

Using CMOS 90 nm technology is a current solution to address this challenge, while reducing the number of analog blocks. In this sub part we present two digital architectures, representing the digitalization trend.

The first architecture presented is based on a classic direct conversion architecture using only one frequency transposition. The “Direct Digital to RF Modulator” (DDRM) architecture [Eloranta et al] was developed as a basis toward further architecture digitization. In this architecture the system is digitalized the closest possible to the amplifier, which is still an analog part of the transmitter, see Fig. 21.

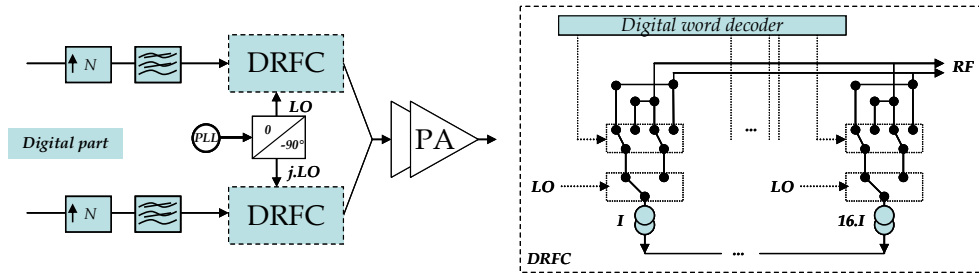


Fig. 21. DDRM architecture (left) and DRFC (right) [Eloranta et al]

The advantage of being digital is that it limits the imperfections due to variations in the process. In analog architectures, the multiple filtering blocks see their basic characteristics varying, and so the architecture performance varies as well. This leads to the use of a calibration loop. Digitization provides size optimization and good stability of the circuit. As the first stage is an over-sampling stage, there is no baseband signal and no need to filter with high selectivity before the DACs. In this architecture, mixing and D/A conversion are performed by a single block: “Digital to RF converter” (DRFC), a kind of RF-DAC. Weighted Gilbert cells are parallelized. No baseband signal in the architecture implies a very low LO leakage (DC offset). The principle is that the data is over-sampled at the carrier frequency by switches. The signal amplitude is coded into a digital word (MSB to LSB). At the output of Gilbert cells, the current is proportional to the code word. The linearity performance and the signal resolution increase with the number of parallel cells. Due to this parallelization, IQ imbalance is limited and can only result from the average cell imbalances. Power control can be achieved by a bias current variation, which reduces the output current from each unit cell. As it uses no filter, the choice of the converter frequency is paramount. Indeed, the only filtering applied (SINC) is the zero-order hold. The frequency must be chosen so that baseband harmonics are cancelled thanks to the zeros of the filter response.

Thanks to CMOS evolutions, this architecture was optimized to address spectrum cohabitation issues [Pozsgay et al] under the name of “Sigma Delta – RFDAC” architecture, illustrated in Fig. 22. In this architecture, the gain control can take place throughout the transmitter. The first power control appears after the first up-sampling filter. It is then a low speed dynamic control, resulting from multiplying IQ signals with a binary word. This word size depends on the control resolution. A second power control stage is achieved, by deleting successive LSBs of the signal (6 dB steps). Several Delta-Sigma modulators in MASH structures are then used and I and Q channels are duplicated (I and I', Q and Q'). I

and Q signals go through a modulator, and signals I' and Q' are first delayed before a second modulator. This will result in the recombination of mixed signals around the carrier, and the creation of notches in desired frequency bands (certainly used by other standards). An additional 6 dB power control can be done by by-passing I' and Q', and the corresponding RF-DAC, losing the advantage of notches as a result.

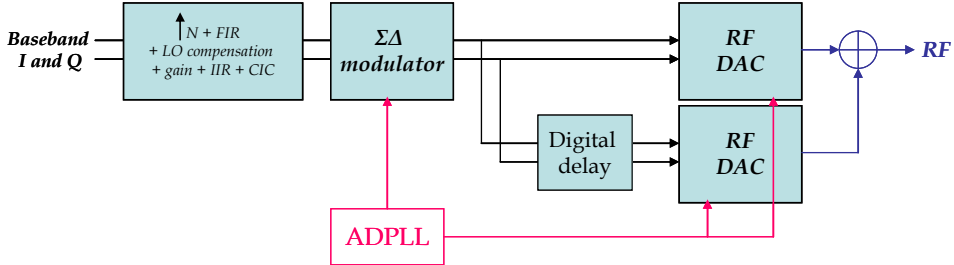


Fig. 22. Sigma Delta - RF-DAC architecture [Pozsgay et al]

The second architecture presented is based on an EER-like architecture. As will be explained, there is no complete separation between phase and amplitude in the transmitted signal processing. The architecture was developed by [Parikh et al], see Fig. 23. IQ signals are over-sampled and filtered in order to place the spectral re-growths far enough (zero order hold) and setting notches at multiples of the over-sampling frequency.

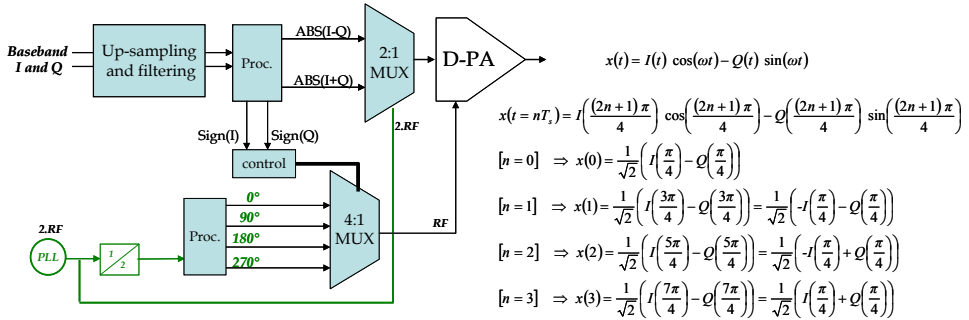


Fig. 23. Digital Quadrature Modulator [Parikh et al]

Two signals are created: "abs(I + Q)" and "abs(I-Q)". Moreover, we determine phase control signals (2 bits) with the signs of I and Q (complex quadrant). Amplitude signals are alternately sent to a digital differential amplifier (DDPA) at a rate of twice the carrier frequency. Thus we get a higher resolution of the variation of the original signal envelope. This can be seen as an over sampling by four of the IQ symbols leading to a better estimation of the amplitude, see Fig. 23. In Fig. 24, we can look at what would happen if I and Q take the value 4 and 3 respectively. This coded amplitude signal feeds a Digital Power Amplifier (DPA). This block will mix the signal around a carrier depending on the instantaneous sign of both I and Q. Knowing I and Q signs, gives us information on the phase (linked to the quadrant position). A part of the phase information can be taken from "abs(I+Q)" and "abs(I-Q)" signals.

This amplitude-coded signal can then be modulated by an RF carrier with an estimated phase. As seen in Fig. 23 and Fig. 24, four carrier signals with four different phases (0° 90° 180° and 270°) feed the DDPA depending on I and Q signs (2-bit control signals). As we obtain a NZR signal, the filtering condition after the DDPA will be less stringent than when using I and Q directly. One of the carrier frequency signals and the coded amplitude are fed to the DDPA. The carrier frequency signal is logical ("0" or "1" values) and activates one or another of the two differential pairs in the DDPA, in function with the sign signals. If the carrier is set to "0", "1", respectively, the output is alternately " $-\text{abs}(I-Q)$ " and " $-\text{abs}(I+Q)$ ", or " $\text{abs}(I+Q)$ " and " $\text{abs}(I-Q)$ " (see Fig. 24). The output network of the DDPA is a band-pass filter around the carrier frequency, to reconstruct the original signal. Looking at the spectrum, performance is well suited to multi-standard applications. This is not due to flexibility in changing the operating frequency. In [Parikh et al] an example is given at 5.8 GHz for a 64QAM modulation scheme and 10 MHz bandwidth WiMAX signal, which is very stringent. The resulting spectrum re-growths are less than -50 dBc/Hz.

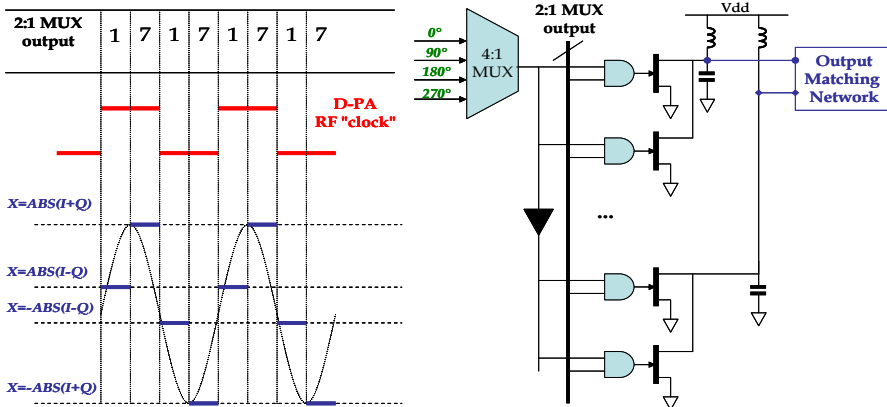


Fig. 24. Signal amplitude and phase estimation (left), DDPA (right). [Parikh et al]

To conclude, digital RF architectures for WB-CW are providing several advantages such as flexibility and size optimization. Their design is subject to the improvement in power consumption of RF blocks. This research topic is currently popular for mobile and connectivity high data rate standards.

5. IR-UWB Architectures

This section is dedicated to IR based and all the UWB systems. Why to consider the UWB systems in addition to already existing technologies? Currently, interfaces have different characteristics in terms of throughput, coverage, access efficiency, quality of service (QoS) and energy consumption. Some of these interfaces offer a QoS for multimedia applications, particularly through a given performance of transmission in a period of time. Other technologies offer specific services such as distance measuring (ranging) like the some UWB systems. UWB technologies are not made for metropolitan networks. The potential of UWB lies in the use of interconnections of a wireless pico-network.

5.1 UWB communications, goals and aspects

There are systems that transmit and receive waves whose relative bandwidth (BW) is greater than or equal to 0.25, see the definition in Fig. 25. The first definition has been amended and replaced by a new one proposed by the FCC. Under this new definition, a UWB signal is a signal whose “-10dB bandwidth” exceed 500 MHz and 20% of center frequency. The main UWB frequency band is between 3.1 and 10.6 GHz. This bandwidth of about 7 GHz could be divided into 14 sub-bands of 500 MHz. A system using the full bandwidth or a set of sub-bands will be considered a UWB system. Today, we can classify UWB into two main categories of applications: UWB Low Data Rate (UWB LDR) and UWB High Data Rate (UWB HDR). They are attached to the IEEE 802.15.4a and 802.15.3a.

Low data rate systems are generally characterized by data rates lower than to 2 Mbps, by ranges of up to 300 meters and finally by a low consumption. They may allow positioning and location functionalities. The high data rate systems are characterized by data rates exceeding 100 Mbps with short ranges (up to a few tens of meters). Their fields of application are the computer data transmission and multimedia systems. Fig. 25 shows the frequency masks used in the USA.

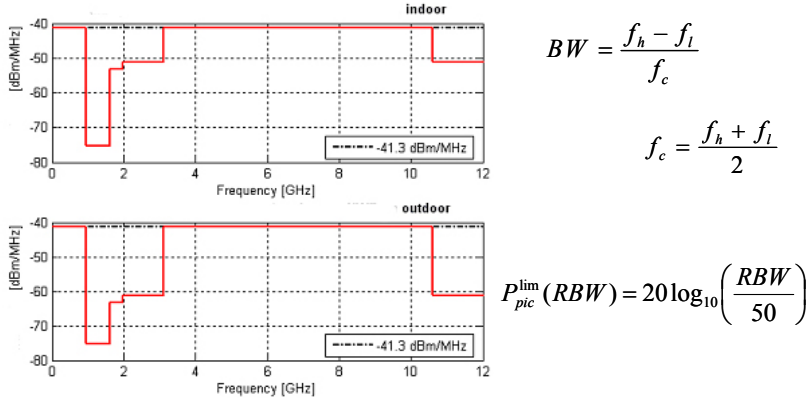


Fig. 25. Frequency masks, relative bandwidth and peak radiated power (RBW in MHz).

The -41.3dBm/MHz limit correspond to a measurement of electromagnetic field equal to 500mVm^{-1} , in any sub-band of 1 MHz, at a distance of 3 meters from the antenna. This level is also named the “Part 15 limit”. In the FCC report, the peak power is also limited. It is measured around the frequency for which the radiation is at its maximum, and is defined in Fig. 25. For $RBW = 50$ MHz, the peak power must not exceed 0 dBm (1 mW). Several transmission techniques approaches have been studied and can be divided into:

- UWB mono-band: impulse approach (IR-UWB, or “classic UWB”)
- UWB Multi-bands: MB-OFDM approach and MB-OOK approach. These systems presents high similarities with WB-CW ones.

5.2 IR-UWB transceiver architectures

The communication is based on short pulses transmission (a few hundred picoseconds), occupying all or part of the UWB spectrum, and repeated with a period of a few

nanoseconds. The signal is transmitted at the baseband frequency. It is interesting even in presence of multi-paths channel. By using a pulsed mode, it allows the measurement of the propagation delay time between the transmitter and the receiver and, consequently, the location. Informations are encoded in the shape of the impulse (amplitude or phase), by the position in time of the pulse or by random sequences of pulses.

Possibilities for modes transmission of the UWB pulses are: coherent, differentially coherent, and non coherent. In the case of coherent transmission, the modulations used are PSK, PAM, and PPM. The receiver will be of a correlator type, e.g. Rake. In the case of differentially coherent transmission, the modulation used is DPSK. Moreover, synchronization becomes un-necessary. For the last case of non coherent transmission, modulations used are OOK or PPM. The receiver will be based on energy detection. The transceiver architectures for IR-UWB are simpler than those for the MB-OFDM or any other WB-CW. The functional blocks of the architecture are given in Fig. 26.

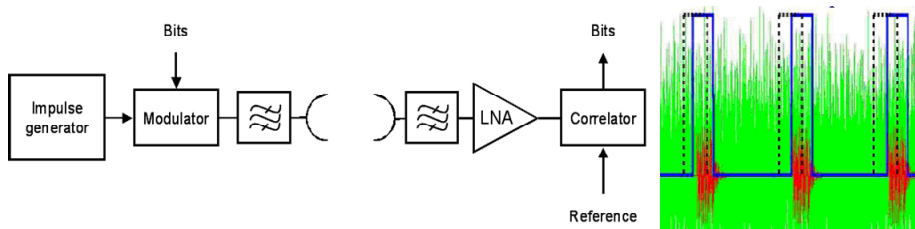


Fig. 26. Functional blocks of IR-UWB transceivers and receiver windowing.

Demodulation should be made on windows slots containing a pulse, as illustrated in Fig. 26. Architectures for impulse radio are divided into two types: Coherent Reception (Co-Rx), requiring knowledge of the phase of the received signal, and Non-Coherent Reception (NCo-Rx), based on detection of energy when pulses are transmitted. In the case of Co-Rx, the receiver uses the recognition of the phase signal. It is generally based on a sliding correlation scheme which produces almost the optimal performance. Therefore the complexity leads to a limitation on the implementation in low consumption equipment. Care should be taken to the synchronization at the acquisition that requires a very fine time to get the precision needed for coherent demodulation.

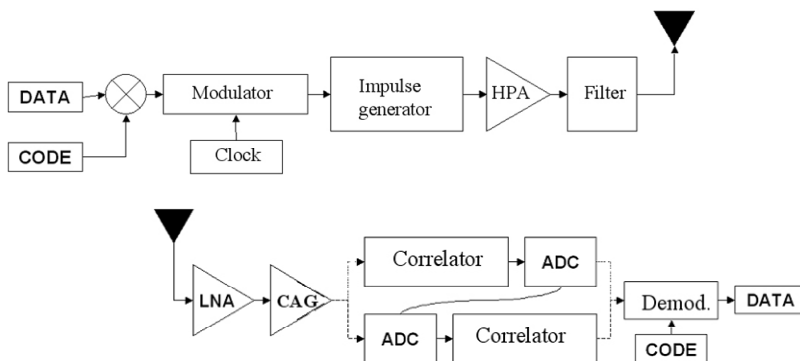


Fig. 27. Examples of Co-Rx transmitter and receiver

The two diagrams in Fig. 27 show the blocks of the transmitter and receiver of a coherent system. Two configurations are possible, the first is mixed analog/digital circuits, the second one is all-digital. Key points of the architecture are: the co-design antenna/LNA, ADC converter performance and the design of the correlator.

For the NCo-Rx, several architectures are possible but the principle is still based on energy detection. This technique is less complex than the previous one but its performance is not as good. The problem is related to the difficulty in designing a receiver able to detect the pulse, when the signal to noise ratio is very poor.

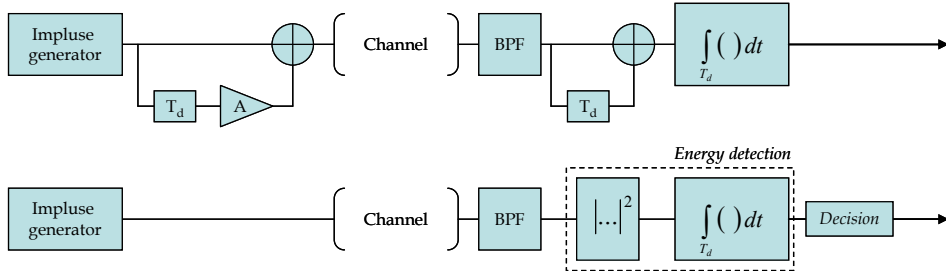


Fig. 28. Examples of NCo-Rx transmitter and receiver

In the case of a non-coherent system, architectures are based on energy detection whose principles are given in Fig. 28. Different types of pulses are used to perform the IR-UWB link: the Gaussian monocycle and its derivatives, the Hermite pulse (Hermite polynomials) or a sinusoid signal windowed by a Gaussian shape. Only the Gaussian pulse or the sinusoid signal windowed by a Gaussian shape are interesting because they can be implanted easily in practice [Marchaland et al., 2007].

5.3 MB-OFDM UWB transceiver architectures

As for CW systems, the MB-OFDM is characterized by a continuous transmission. It uses frequency hopping on at least three bands, 128 sub-carrier band of 528 GHz (in the case of 3.1 - 10.6 GHz band), and the QPSK modulation scheme. It allows a great flexibility in shaping the spectrum. This parallel multi-carrier operation minimizes inter-symbol interference, and the recovery of the energy available can be optimized. The MB-OFDM approach is based on an IFFT at emission and FFT at reception as shown in Fig. 29.

The overall topology of the receiver is very complex. The MB-OFDM technique is well suited to high speeds data transfer in indoor environments. It has good resistance to multi-paths channel, and, allows flexible shaping of the spectrum. Modulation is complex to implement and requires circuitry to perform an FFT in real time, so the digital part of the architecture is quite complex. It needs synchronization. This technique requires front-end with high linearity and low noise (RF elements). Additionally, the consumption of the analog area may be important.

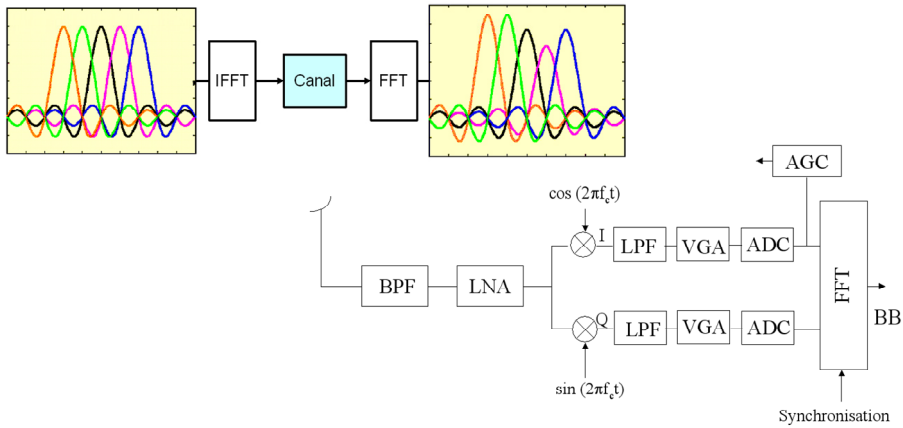


Fig. 29. Considerations for MB-OFDM UWB transceivers

6. Conclusion

The goal of this chapter is to demonstrate the absolute need of matching the architecture design and the signal carrying the information.

In section 1 and 2, considerations about current users' needs helped to identify three types of carrier signals in the context of radio-communications: NB-CW, WB-CW and IR. A separation between CW and IR signals is unavoidable because it drives us to a different technological design. Basic blocks of transmitter architecture are optimized in function with the targeted performance over the bandwidth, as was discussed in section 3. Conclusions were given about actual trends in this research topic. The goal of section 4 was to illustrate the high degree of complexity for actual WB-CW systems, which represents high data rate applications. Section 5 was an overview of IR based architectures.

Whatever the system is, the design of the transceiver architecture has to fulfill challenges such as co-existence (for the transmitter part) and immunity (for the receiver part). This implies a careful reduction of the spectral emission (spectral re-growths) for a transmitter. It also implies a high selectivity for the receiver and, of course, with the lowest sensitivity possible. RF architectures are becoming more and more complex. Especially, in the cases of RF transmitters, high efficient power amplification often results in a combination of different linearization techniques. The ever-increasing performance of the digital part gives us the opportunity to provide more flexibility in the architecture when the dynamic control and the power delivered to the load are satisfying standard requirements. This can be done thanks to the digitalization of some functions such as constant amplitude envelope coding, RF converters and/or PA (D-PA).

The concept of multi-radio points out that the future of RF standards lies in favoring cooperation and flexibility in the managing of the system resources (bandwidths, power and time partitioning). Actual state-of-the-art results are that architectures are merging to reconfigurable RF blocks and when possible RF-digital blocks.

7. References

- Andia, L. et al. Specification of a Polar Sigma Delta Architecture for Mobile Multi-Radio Transmitter - Validation on IEEE 802.16e. *Proc. of IEEE Radio and Wireless Symposium*, pp. 159-162, Orlando, USA, Jan. 2008.
- Baudoin, G. et al. Radiocommunications Numériques : Principes, Modélisation et Simulation. *Dunod, EEA/Electronique*, 672 pages, 2ème édition 2007.
- Baudoin, G. et al. Influence of time and processing mismatches between phase and envelope signals in linearization systems using EER, application to hiperlan 2. *Proc. Conf. IEEE - MTT'2003 Microwave Theory and Technique*, Philadelphia USA, June 2003.
- Berland, C. et al. A transmitter architecture for Non-constant Envelope Modulation. *IEEE Trans. Circuits and Systems II: Express Briefs*, vol. 53, no. 1, pp. 13-17, January 2006
- Choi, J. et al. A $\Sigma\Delta$ digitized polar RF transmitter. *IEEE Trans. on Microwave Theory and Techniques*, Vol. 52, n°12, 2007, pp 2679-2690.
- Cox, D. Linear amplification with non-linear components, LINC method. *IEEE transactions on Communications*, Vol COM-23, pp 1942-1945, December 1974.
- Diet, A. et al. EER architecture specifications for OFDM transmitter using a class E power amplifier. *IEEE Microwave and Wireless Components Letters (MTT-S)*, ISSN 1531-1309, V-14 I-8, August 2004, pp 389-391.
- Diet, A. et al. Flexibility of Class E HPA for Cognitive Radio. *IEEE 19th symposium on Personal Indoor and Mobile Radio Communications, PIMRC 2008*, 15-18 september, Cannes, France. CD-ROM ISBN 978-1-4244-2644-7.
- Diet, A., Baudoin, G., Villegas, M. Influence of the EER/polar Transmitter Architecture on IQ Impairments for an OFDM Signal. *International Review of Electrical Engineering, IREE Praise Worthy Prize*, ISSN 1827-6660, V-3 N-2, March-April 2008, pp 410-417.
- Diet, A., Villegas, M., Baudoin, G. EER-LINC RF transmitter architecture for high PAPR signals using switched Power Amplifiers. *Physical Communication, ELSEVIER*, ISSN: 1874-4907, V-1 I-4, December 2008, pp. 248-254.
- Dürdodt, D. et al. A low-IF Rx two-point $\Sigma\Delta$ -modulation Tx CMOS single-chip bluetooth solution. *IEEE Trans. Microw. Theory Tech.*, vol. 51, no. 9, pp. 1531-1537, Sep. 2001.
- Eloranta, P., Seppinen, P., Parssinen, A. Direct-digital RF-modulator: a multi-function architecture for a system-independent radio transmitter. *Com. Magazine, IEEE V46*, I4, 2008, pp 144-151.
- Groe, J. A Multimode Cellular Radio. *IEEE Trans. On circuits and systems – II: Express briefs*, Vol. 55, No. 3, March 2008, pp. 269-273.
- Groe, J. Polar Transmitters for Wireless Communications. *IEEE Communications Magazine* September 2007, pp. 58-63.
- Hibon, I. et al. Linear transmitter architecture using a 1-bit $\Delta\Sigma$. *European Microwave Week 2005, Proc. Conf. ECWT*, pp. 321-324, Octobre 2005.
- Jardin, P., Baudoin, G. Filter Lookup Table Method for Power Amplifier Linearization. *IEEE Trans. on Vehicular Technology*, N° 3, Vol. 56, pp. 1076-1087, IEEE, Mai 2007.
- Jeong, J., Wang, Y. A Polar Delta-Sigma Modulation (PSDM) Scheme for High Efficiency Wireless Transmitters. *IEEE MTT-S Int. Microwave Symp. Dig.* June 2007.
- Kahn, L. Single Sideband Transmission by Envelope Elimination and Restoration. *Proc. of the I.R.E.*, 1952, pp. 803-806.
- Marchaland, D., Badets, F. Générateur d'impulsions ULB doté d'une fonction intégrée d'émulation numérique. FR0700683, le 31 Janvier 2007.

- McCune, E. Polar Modulation and Bipolar RF Power Devices. *IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, October 2005.
- McCune, E. High efficiency, multimode, multiband terminal power amplifiers. *IEEE Microwave Magazine*, March 2005, Volume: 6, Issue: 1, pp: 44- 55.
- Murmann, B. Digitally Assisted Analog Circuits – A Motivational Overview. *IEEE International Solid-State Circuits Conf.: Special Topic Evening Session*, February 2007.
- Nielsen, M., Larsen, T. Transmitter Architecture Based on $\Delta\Sigma$ Modulation and SW Power Amplification. *IEEE Trans. on Circuits and Syst. II*, 2007, vol. 54, no. 8, pp. 735-739.
- Parikh, V., Balsara, P., Eliezer, O. A fully digital architecture for wideband wireless transmitters. *IEEE RVWS, Radio and Wireless Symposium*, 2008.
- Pozsgay, A. et al. A Fully Digital 65nm CMOS Transmitter for the 2.4-to-2.7GHz WiFi/WiMAX Bands using 5.4GHz RF DACs. *IEEE ISSCC 2008*, pp 360-619.
- Raab, F. et al. RF and Microwave PA and Transmitter Technologies. *High Frequency Electronics*, May-November 2003, pp 22-49.
- Robert, F. et al. Study of a polar $\Delta\Sigma$ transmitter associated to a high efficiency switched mode amplifier for mobile Wimax. *10th annual IEEE Wireless and Microwave Technology Conference, WAMICON*, april 2009, Clearwater, FL, USA.
- Rode, J., Hinrichs, J., Asbeck, P. Transmitter architecture using digital generation of RF signals. *IEEE Radio and Wireless Conf.*, pp. 245-248, August 2003.
- Schantz, H. Art and Science of UWB Antennas. *Artech House*, 2005.
- Sokal, N., Sokal, A. Class E, A new Class of high efficiency Tuned single ended switching PAs. *IEEE journal of Solid State Circuits*, Vol. 10, No. 3, Juin 1975, pp 168-176.
- Sowlati, T. et al. Quad band GSM/GSM/GPRS polar loop transmitter. *IEEE Journal of Solid-State Circuits*, Volume 39, Issue 12, Dec. 2004 Page(s): 2179 – 2189.
- Staszewski, R et al. All digital PLL and transmitter for mobile phones. *IEEE Journal of Solid-State Circuits*, Volume 40, Issue 12, Dec. 2005 Page(s): 2469 – 2482.
- Suarez Penaloza, M. et al. "Study of a Modified Polar Sigma-Delta Transmitter Architecture for Multi-Radio Applications", *EuMW*, 27-31 Octobre 2008, Amsterdam.
- Villegas, M. et al. Radiocommunications Numériques : Conception de circuits intégrés RF et micro-ondes. *Dunod, EEA/Electronique*, 464 pages, 2ème édition 2007.
- Wendell, B. et al. Polar modulator for multi-mode cell phones. *Proceedings of the IEEE 2003 Custom Integrated Circuits Conference*, Sept. 2003, pp: 439 – 44.

Analytical SIR for Cross Layer Channel Model

Abdurazak Mudesir

*Jacobs University Bremen, School of Engineering and Science, Research 1,
Campus Ring 12. 28759 Bremen
Germany*

Harald Haas

*The University of Edinburgh, Institute for Digital Communications,
The Kings Buildings Edinburgh EH9 3JL
UK*

1. Introduction

In a wireless communication environment characterized by dynamically-varying channels, high influence of interference, bandwidth shortage and strong demand for quality of service (QoS) support, the challenge for achieving optimum spectral efficiency and high data rate is unprecedented. One of the bottlenecks in achieving these goals is modeling of the propagation environments Yun & Iskander (2004).

The commonly used radio propagation models only account for the large scale path loss Rapaport (2001) or only multipath propagations Alouini & Goldsmith (1997), which are incomplete for studying realistic system deployment scenarios. The authors in Alouini & Goldsmith (1997) calculate the capacity of Nakagami multipath fading (NMF) channels assuming that the carrier-to-noise ratio (CNR) is gamma distributed. This assumption neglects the effects of shadowing and large scale path loss. This chapter presents an "exact" pdf derived from a model which is more closely related to a realistic deployment scenario. With the results provided here, it is possible to calculate more precise capacity figures. Moreover since the new path loss model takes into consideration the interaction of the system or data link layer with the physical layer (PHY), it is particularly important in studying the cross-layer interaction between MAC and PHY layer.

The general approach of cross-layer design is to maintain the already established layered architecture, capture the important information that influences other layers, exchange the information between layers and implement adaptive protocols and algorithms at each layer to optimize the system performance Xia & Liang (2005).

Adhering to the idea of cross-layer solutions, many researchers are working towards finding viable cross layer designs spanning the different layers of the OSI model. Yeh and Cohen proposed a generalized analytical framework for cross-layer design focusing on resource allocation in the presence of multi-access fading channels Yeh & Cohen (2003). Ahn Ahn et al. (2002) uses the information from the medium access control (MAC) layer to do rate control at the network layer.

Cross-layer design relies on channel quality information. The channel quality can be captured by a single parameter, namely the received SIR. The SIR between two communicating nodes

will typically decrease as the distance between the nodes increases, and will also depend on the signal propagation and interference environment. Moreover, the SIR varies randomly over time due to the propagation environment and interference characteristics. Therefore modeling the SIR on the assumption of the cellular structure and the well known path loss model that ignores the small scale fading would not be applicable to self-configuring cross-layer design. Therefore analytical derivation of the pdf of SIR is a crucial step in constructing efficient cross-layer design.

Tellambura in Tellambura (1999) uses a characteristic function method to calculate the probability that the SIR drops below some predefined threshold (probability of outage) under the assumption of Nakagami fading. Zhan Zhang (1996) also uses a similar characteristic function approach to derive outage probability for multiple interference scenario. These papers give a significant advantage in reducing the computational complexity involved in solving multiple integrals in SIR computation. But, a major shortcoming of these and other similar papers Zorzi (1997) is that, only the small scale fading (physical layer) or large scale fading (data link layer) is considered in analytically deriving the SIR statistics.

The rest of this chapter is organized as follows. In Section II the system model considered is presented and in Section III the analytical derivation is described in detail. Section IV provides the numerical and the simulation results. Section V concludes the chapter.

2. System Model and Problem Formulation

For simplicity the cell layout used to derive the pdf of the SIR assumes circular cells, as shown in Fig. 1, with maximum cell radius R_c instead of hexagonal cells. The cells are randomly positioned resulting in potentially overlapping cells. Randomly positioned cells model an important network scenario, which lacks any frequency planning as a result of self-configuring and self-organising networks, cognitive radio, multihop *ad hoc* communication and cross layer system design. A receiver experiences interference from transmitters within its accessibility radius, R_{ac} . Due to propagation path loss, a transmitter outside the accessibility region incurs only a negligible interference. Since the aim is to model a realistic interference limited environment, the receiver accessibility radius is taken to be much greater than the cell radius i.e $R_{ac} \gg R_c$. The dashed line in Fig. 1 represents the interference link between transmitter, Tx y , and receiver, Rx z while the solid line shows the desired link between transmitter Tx x and receiver Rx z and vice versa. Throughout the derivation omni-directional antennas with unity gains are considered. The pdf is calculated assuming one interfering user. The results obtained can be extended to multiple interfering users by using laguerre polynomials to approximate the multiple integration resulting from the multiple interfering users. The analytical derivation of SIR for multiple interference is under study.

3. Analytical derivation of the pdf of the sir

In an interference limited environment, the received signal quality at a receiver is typically measured by means of achieved SIR, which is the ratio of the power of the wanted signal to the total residue power of the unwanted signals. let P_t and P_r denote the transmit and received power respectively. Let G denote the path gain and G_{yz} is the link gain between the interfering transmitter y and the receiver z . For the purpose of clarity, unless otherwise stated, a single subscript x , y or z specifies the node, and a double subscript such as xz specifies the link between node x and node z . A node is any entity, mobile station(MS) or base station(BS)

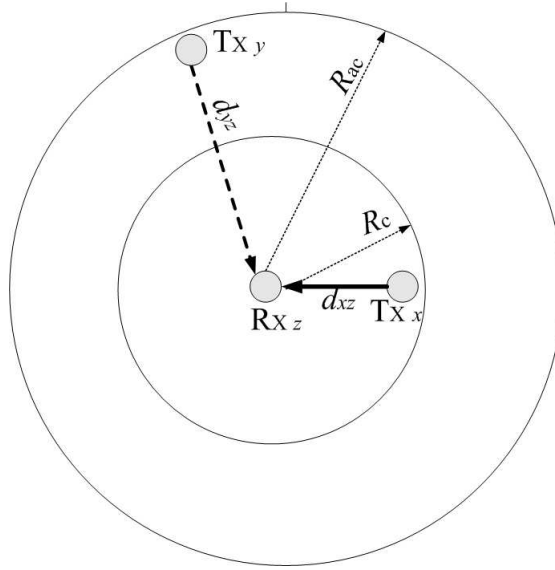


Fig. 1. Model to drive the pdf of the SIR from a single neighboring cell

that is capable of communicating. For a single interfering user y depicted in Fig. 1:

$$\text{SIR}_z = \frac{P_{tx} G_{xz}}{P_{ty} G_{yz}} \quad (1)$$

Assuming fixed and constant transmit powers, $P_{tx} = P_{ty} = \text{const}$, (1) simplifies to:

$$\text{SIR}_z = \frac{G_{xz}}{G_{yz}} \quad (2)$$

$$L = \frac{1}{G} \Rightarrow \text{SIR}_z = \frac{L_{yz}}{L_{xz}} \quad (3)$$

where L_{xz} , L_{yz} are the path losses between transmitter $\text{Tx } x$ and receiver $\text{Rx } z$ and $\text{Tx } y$ and $\text{Rx } z$ respectively.

Like the gain parameter G , the loss parameter L incorporates effects such as propagation loss, shadowing and multipath fading.

The generalized path loss model for the cross layer environment is given by:

$$L = \underbrace{C \left(\frac{d}{d_0} \right)^\gamma e^{(\beta\zeta)}}_{\text{large scale path loss}} \cdot \underbrace{\frac{1}{|H(f)|^2}}_{\text{small scale path loss}} \quad (4)$$

Where $C = \frac{\hat{C}}{\bar{C}}$ is an environment specific constant, \hat{C} the constant corresponding to the desired link while \bar{C} corresponds to the interference link. The distance d_0 is a constant and d is a

random variable, γ is the path-loss exponent, ξ is the random component due to shadowing, $\beta = \ln(10)/10$ and $|H(f)|$ is a random variable modeling the channel envelop.

The commonly used path loss equation Rappaport (2001) only accounts for the large scale path loss with regular cell deployment scenarios, which is incomplete for studying self-organizing networks. The new path loss model proposed here takes into consideration the interaction of the large scale path loss as well as the small scale fading. This model is particularly important in studying the performance of self-organizing self-configuring networks.

For the interference scenario described in the system model, the path loss for the desired path and the path loss between the interfering transmitter y and the receiver z (interfering link) are:

$$L_{xz} = \tilde{C} d_{xz}^{\gamma_{xz}} e^{(\beta \xi_{xz})} \frac{1}{|H_{xz}|^2} \quad (5)$$

$$L_{yz} = \hat{C} d_{yz}^{\gamma_{yz}} e^{(\beta \xi_{yz})} \frac{1}{|H_{yz}|^2} \quad (6)$$

where L_{xz} is the path loss model for the desired link and L_{yz} is the path loss model for the interfering link. d_{yz} models the distance between the interference causing transmitter, x , and the victim receiver y . γ_{yz} and γ_{xz} are the path loss exponents, ξ_{xz} and ξ_{yz} are Gaussian distributed random variables modeling the shadow fading with each zero mean and variances v_{xz}^2 and v_{yz}^2 respectively, and $|H_{xz}|$ and $|H_{yz}|$ are the channel envelope modeling the channel fading. For the purpose of clarity, the time and frequency dependencies are not shown. The channel envelope is assumed to follow the Nakagami- m distribution. Nakagami distribution is a general statistical model which encompasses Rayleigh distribution as a special case, when the fading parameter $m = 1$, and also approximates the Rician distribution very well. In addition, Nakagami- m distribution will also provide the flexibility of choosing different distributions for the desired link and interfering link, such as the Rayleigh for the channel envelope of the desired link, and Rician for the interfering link, or vice versa.

Using equations (3) and (5), the SIR can be given as:

$$\text{SIR} = \frac{C d_{yz}^{\gamma_{yz}} e^{(\beta \xi_{yz})} |H_{xz}|^2}{\tilde{d}_{xz}^{\gamma_{xz}} e^{(\beta \xi_{xz})} |H_{yz}|^2} \quad (7)$$

From (7), the SIR has six random variable components, $\Phi_{xz} = d_{xz}^{\gamma_{xz}}$, $\Phi_{yz} = d_{yz}^{\gamma_{yz}}$, $\Lambda_{xz} = e^{(\beta \xi_{xz})}$, $\Lambda_{yz} = e^{(\beta \xi_{yz})}$, $|H_{xz}|^2$ and $|H_{yz}|^2$. In order to analytically derive the pdf of the SIR, the pdf of the individual components and also their ratios and products need to be determined first.

The following two formulas provide the basic framework for the analysis and will be used throughout the derivation. Given two independent random variables X and Y the pdf of their product $f_Z(z)$ where $Z = XY$ is

$$f_Z(z) = \int f_X(z/x) f_Y(x) (1/|x|) dx \quad (8)$$

Given two independent random variables Y and X the pdf of their ratio $f_Z(z)$ where $Z = \frac{Y}{X}$ is

$$f_Z(z) = \int f_X(x) f_Y(zx) |x| dx \quad (9)$$

3.1 Pdf of the ratio of the propagation loss

It is assumed that the distance between the interfering transmitter and the receiver, d_{yz} , is uniformly distributed up to a maximum distance of R_{ac} , and that the distance between an interfering transmitter and intended receiver, d_{xz} , is uniformly distributed up to a maximum distance of R_c . Therefore Φ_{xz} and Φ_{yz} are both functions of random variables, and their pdfs can be derived using the following random variable transformation Papoulis (1991).

$$p(\theta) = \left. \frac{p(\delta)}{\left| \frac{d(\theta)}{d(\delta)} \right|} \right|_{\delta=F^{-1}(\theta)} \quad (10)$$

Where θ and δ are random variables with pdfs $p(\theta)$ and $p(\delta)$ respectively, and where θ is a function of $F(\delta)$, $d(\theta)$ and $d(\delta)$ are the first derivatives of θ and δ respectively.

The mathematical representation of the pdfs of d_{xz} and d_{yz} are

$$f_{D_{xz}}(d_{xz}) = \frac{2d_{xz}}{R_c^2} \quad 0 < d_{xz} \leq R_c \quad (11)$$

$$f_{D_{yz}}(d_{yz}) = \frac{2d_{yz}}{R_{ac}^2} \quad 0 < d_{yz} \leq R_{ac} \quad (12)$$

Let $f_{\Phi_{xz}}(\phi_{xz})$ and $f_{\Phi_{yz}}(\phi_{yz})$ denote the pdfs of Φ_{xz} and Φ_{yz} . Then employing the transformation (10), $f_{\Phi_{xz}}(\phi_{xz})$ and $f_{\Phi_{yz}}(\phi_{yz})$ are derived as

$$f_{\Phi_{xz}}(\phi_{xz}) = \frac{2\phi_{xz}^{2/\gamma_{xz}-1}}{R_c^2 \gamma_{xz}} \quad 0 < \phi_{xz} \leq R_c^{\gamma_{xz}} \quad (13)$$

$$f_{\Phi_{yz}}(\phi_{yz}) = \frac{2\phi_{yz}^{2/\gamma_{yz}-1}}{R_{ac}^2 \gamma_{yz}} \quad 0 < \phi_{yz} \leq R_{ac}^{\gamma_{yz}} \quad (14)$$

Using (9), the pdf of the ratio of the propagation loss, $\Phi = \frac{\Phi_{yz}}{\Phi_{xz}}$, is found to be,

$$f_{\Phi}(\phi) = \begin{cases} Y\phi^{2/\gamma_{yz}-1} & \text{for } 0 < \phi \leq \varsigma \\ \tilde{Y}\phi^{-2/\gamma_{xz}-1} & \text{for } \varsigma < \phi < \infty \end{cases} \quad (15)$$

where $\varsigma = \frac{R_{ac}^{\gamma_{yz}}}{R_c^{\gamma_{xz}}}$, $Y = \frac{2R_c^{2/\gamma_{yz}}}{R_{ac}^2(\gamma_{yz} + \gamma_{xz})}$ and $\tilde{Y} = \frac{2R_{ac}^{2/\gamma_{xz}}}{R_c^2(\gamma_{yz} + \gamma_{xz})}$

The next step to derive the pdf of the SIR is find the pdf of the ratio of the lognormal shadowing.

3.2 Pdf of the ratio of the lognormal shadowing

Given a normally distributed random variable X with mean μ and variance σ^2 , and a real constant c , the product cX is known to follow a normal distribution with mean $c\mu$ and a variance $c^2\sigma^2$ and e^X has a log-normal distribution. Since ξ_{xz} is normally distributed with mean μ and variance σ^2 , $\Lambda_{xz} = e^{(\beta\xi_{xz})}$ is a lognormal distributed random variable with mean μ_{xz} and variance $v_{xz}^2 = \beta^2\sigma_{xz}^2$ expressed in terms of the normally distributed ξ_{xz} , while the mean and variance of $\Lambda_{yz} = e^{(\beta\xi_{yz})}$ are μ_{yz} and $v_{yz}^2 = \beta^2\sigma_{yz}^2$ respectively.

$$f_{\Lambda_{xz}}(\lambda_{xz}) = \frac{e^{-1/2 \frac{(\ln(\lambda_{xz}) - \mu_{xz})^2}{v_{xz}^2}}}{\lambda_{xz} v_{xz} \sqrt{2\pi}}, \quad 0 \leq \lambda_{xz} < \infty \quad (16)$$

$$f_{\Lambda_{yz}}(\lambda_{yz}) = \frac{e^{-1/2 \frac{(\ln(\lambda_{yz}) - \mu_{yz})^2}{v_{yz}^2}}}{\lambda_{yz} v_{yz} \sqrt{2\pi}}, \quad 0 \leq \lambda_{yz} < \infty \quad (17)$$

Since the ratio of two independent lognormal random variables is itself a lognormal distributed random variable. Therefore the pdf of $\Lambda = \frac{\Lambda_{yz}}{\Lambda_{xz}}$ is:

$$f_{\Lambda}(\lambda) = \frac{e^{-1/2 \frac{(\ln(\lambda) - \mu)^2}{\sigma^2}}}{\lambda \sigma \sqrt{2\pi}}, \quad 0 \leq \lambda < \infty \quad (18)$$

where

$$\sigma = \beta \sqrt{v_{xz} + v_{yz}}, \quad \mu = 0;$$

The last components remaining from (7) are the random variables modeling the channel envelop and their ratios.

3.3 Pdf of the ratio of the channel envelope

In order to accommodate different channel fading distributions, Nakagami- m distribution was used to model the channel envelope. Nakagami- m distribution is the most general of all distribution known until now Nakagami (1960).

The Nakagami- m pdf is given by:

$$f_{|H_{xz}|}(h_{xz}) = \frac{2}{\Gamma(m_{xz})} \left(\frac{m_{xz}}{\Omega_{xz}} \right)^{m_{xz}} h_{xz}^{2m_{xz}-1} e^{-\frac{m_{xz} h_{xz}^2}{\Omega_{xz}}}, \quad 0 \leq h_{xz} < \infty \quad (19)$$

$$f_{|H_{yz}|}(h_{yz}) = \frac{2}{\Gamma(m_{yz})} \left(\frac{m_{yz}}{\Omega_{yz}} \right)^{m_{yz}} h_{yz}^{2m_{yz}-1} e^{-\frac{m_{yz} h_{yz}^2}{\Omega_{yz}}}, \quad 0 \leq h_{yz} < \infty \quad (20)$$

where $m \geq 1/2$ represents the fading figure, $\Omega = E(x^2)$ is the average received power and $\Gamma(\cdot)$ is the gamma function given as

$$\Gamma(m) = \int_0^{\infty} x^{m-1} e^{-x} dx.$$

The pdfs of the received instantaneous power, $H_{xz} = |H_{xz}|^2$ are modeled by a gamma distribution. For the desired user the pdf of the receive signal power, $f_{H_{xz}}(h_{xz})$, is given as

$$f_{H_{xz}}(h_{xz}) = \frac{h_{xz}^{m_{xz}-1}}{\Gamma(m_{xz})} \left(\frac{m_{xz}}{\Omega_{xz}} \right)^{m_{xz}} e^{-\frac{m_{xz} h_{xz}}{\Omega_{xz}}}, \quad 0 \leq h_{xz} < \infty \quad (21)$$

and for the interfering user the PDF, $f_{|H_{yz}|}(h_{yz})$, is

$$f_{H_{yz}}(h_{yz}) = \frac{h_{yz}^{m_{yz}-1}}{\Gamma(m_{yz})} \left(\frac{m_{yz}}{\Omega_{yz}} \right)^{m_{yz}} e^{-\frac{m_{yz} h_{yz}}{\Omega_{yz}}}, \quad 0 \leq h_{yz} < \infty \quad (22)$$

Using (8) and (9) the pdf of the ratio of gamma distribution, $\Psi = \frac{H_{xz}}{H_{yz}}$ is:

$$f_{\Psi}(\psi) = M \frac{\psi^{m_{xz}-1}}{\left(\frac{m_{yz}}{\Omega_{yz}} + \frac{m_{xz}}{\Omega_{xz}} \psi\right)^{(m_{yz}+m_{xz})}} \quad 0 \leq \psi < \infty \quad (23)$$

where

$$M = \frac{\Gamma(m_{yz})\Gamma(m_{xz})}{\Gamma(m_{yz}+m_{xz})} \left(\frac{m_{yz}}{\Omega_{yz}}\right)^{m_{yz}} \left(\frac{m_{xz}}{\Omega_{xz}}\right)^{m_{xz}}. \quad (24)$$

Using the beta function, also called the Euler integral of the first kind, M can be re-written as

$$M = \frac{\left(\frac{m_{yz}}{\Omega_{yz}}\right)^{m_{yz}} \left(\frac{m_{xz}}{\Omega_{xz}}\right)^{m_{xz}}}{B(m_{xz}, m_{yz})}, \quad (25)$$

where

$$B(m_{xz}, m_{yz}) = \int_0^1 t^{m_{xz}-1} (1-t)^{m_{yz}-1} dt = \frac{\Gamma(m_{yz}+m_{xz})}{\Gamma(m_{yz})\Gamma(m_{xz})}.$$

The final step in the derivation of the pdf of the SIR is deriving the product of the above obtained pdfs.

3.4 Pdf of the SIR

As shown in (7) the pdf of the SIR is the product of the three individual random variables, Φ , Λ and Ψ . Using the equations presented so far, the final pdf of the SIR is presented in (26)

$$f_{\text{SIR}}(\zeta) = M_{\zeta}^{2m_{xz}-1} \times \quad (26)$$

$$\frac{\int_0^{\infty} \left(A_1 \chi^{q_1} \left(\operatorname{erf} \left(\frac{\frac{\frac{2}{\gamma_{yz}} \sigma^2 + \ln \left(\frac{\chi}{\left(\frac{R_{ac}}{R_{T_{xz}}} \right)} \right)}{\sqrt{2\sigma}}} \right) - 1 \right) + B_1 \chi^{q_2} \left(-1 - \operatorname{erf} \left(\frac{\frac{\frac{-2}{\gamma_{xz}} \sigma^2 + \ln \left(\frac{\chi}{\left(\frac{R_{ac}}{R_{T_{xz}}} \right)} \right)}{\sqrt{2\sigma}}} \right) \right) \right)}{\left(\frac{m_{yz}}{\Omega_{xz}} + \frac{m_{xz}}{\Omega_{yz}} \left(\frac{\zeta}{\chi} \right)^2 \right)^{(m_{yz}+m_{xz})}} d\chi$$

where $q_1 = \frac{2}{\gamma_{yz}} - 2m_{yz} - 1$, $q_2 = \frac{-2}{\gamma_{xz}} - 2m_{yz} - 1$,

$$A_1 = \frac{-\frac{2R_c^2 \gamma_{xz}}{R_{ac}^2 (\gamma_{yz} + \gamma_{xz})}}{2} e^{\left(\frac{2}{\gamma_{yz}}\right) \sigma^2}$$

and

$$B_1 = \frac{-\frac{2R_c^2 \gamma_{yz}}{R_{ac}^2 (\gamma_{yz} + \gamma_{xz})}}{2} e^{\left(\frac{2}{\gamma_{xz}}\right) \sigma^2}$$

The final equation does not have a closed form solution but it is possible to solve the integration using numerical methods.

4. Signal to Interference and Noise Ratio

In case of an environment that is not interference limited, the SINR (signal to interference and noise ratio) is required to fully describe the communication channel. SINR can easily be found by modifying the SIR equation given in (1):

$$\text{SINR}_z = \frac{G_{xz}}{G_{yz} + N} \quad (27)$$

where N is the random variable modeling the Gaussian noise with mean $m_N = 0$ and a standard deviation of σ_N . By applying the generalized path loss equation in (4), SINR at the receiver Rx z is given by:

$$\text{SINR}_z = \frac{\frac{|H_{xz}|^2}{d_{xz}^{\gamma_{xz}} e^{(\beta_{\xi,xz}^2)}}}{\frac{|H_{yz}|^2}{d_{yz}^{\gamma_{yz}} e^{(\beta_{\xi,yz}^2)}} + N} \quad (28)$$

where the pdfs of the individual random variables are given in the previous section. let $\Theta_{xz} = \Phi_{xz} \Lambda_{xz} = d_{xz}^{\gamma_{xz}} e^{(\beta_{\xi,xz}^2)}$ which are derived in the previous section. The pdf of Θ_{xz} , $f_{\Theta_{xz}}(\theta_{xz})$, is given as:

$$f_{\Theta}(\theta_{xz}) = \int f_{\Phi}(\theta_{xz}/\lambda_{xz}) f_{\Lambda_{xz}}(\lambda_{xz}) (1/|\lambda_{xz}|) d\lambda_{xz} \quad (29)$$

$$f_{\Theta}(\theta_{xz}) = \int_{\frac{\theta_{xz}}{R_c^2}}^{\infty} \frac{2(\frac{\theta_{xz}}{\lambda_{xz}})^{2/\gamma_{xz}-1}}{R_c^2 \gamma_{xz}} \frac{e^{-1/2 \frac{\beta(\ln(\lambda_{xz})-\mu)^2}{v_{xz}^2}}}{\lambda_{xz} v_{xz} \sqrt{2\pi}} \frac{1}{\lambda_{xz}} d\lambda_{xz} \quad (30)$$

$$f_{\Theta}(\theta_{xz}) = D \left(1 - \frac{\text{erf}(2v_{xz}^2 - \gamma_{xz}m_{xz} + \gamma_{xz} \log \frac{\theta_{xz}}{R_c^{\gamma_{xz}}})}{\sqrt{(2)}\gamma_{xz}v_{xz}} \right) \quad (31)$$

$$\text{where } D = \frac{e^{\frac{2v_{xz}^2 - 2\gamma_{xz}m_{xz}}{\gamma_{xz}^2}}}{R_c^2 \gamma_{xz}} \theta_{xz}^{\frac{2}{\gamma_{xz}-1}}$$

The next step in the derivation is to find the pdf of the path loss of the desired link by utilizing (9) and (19). Let $S = \frac{H_{xz}}{d_{xz}^{\gamma_{xz}} e^{(\beta_{\xi,xz}^2)}}$ be the random variable denoting the path loss of the desired link. The pdf of S is given as:

$$f_S(s) = K \int_0^{\infty} h_{xz}^{2m_{xz}} e^{-\frac{m_{xz} h_{xz}^2}{\Omega_{xz}}} \left(1 - \frac{\text{erf}(2v_{xz}^2 - \gamma_{xz}m_{xz} + \gamma_{xz} \log \frac{sh_{xz}}{R_c^{\gamma_{xz}}})}{\sqrt{(2)}\gamma_{xz}v_{xz}} \right) dh_{xz} \quad (32)$$

$$\text{where } K = \frac{2}{\Gamma(m_{xz})} \frac{m_{xz}^2}{\Omega_{xz}} D$$

The pdf of the path loss of the interference path denoted by the random variable $I = \frac{|H_{yz}|}{d_{yz}^{\gamma_{yz}} e^{(\beta_{\xi,yz}^2)}}$ is give as:

$$f_I(i) = \hat{K} \int_0^{\infty} h_{yz}^{2m_{yz}} e^{-\frac{m_{yz} h_{yz}^2}{\Omega_{yz}}} \left(1 - \frac{\text{erf}(2v_{yz}^2 - \gamma_{yz}m_{yz} + \gamma_{yz} \log \frac{ih_{yz}}{R_{rv}^{\gamma_{yz}}})}{\sqrt{(2)}\gamma_{yz}v_{yz}} \right) dh_{yz} \quad (33)$$

$$\text{where } \hat{K} = \frac{2}{\Gamma(m_{yz})} \frac{m_{yz}^2}{\Omega_{yz}} D$$

In order to find the pdf of the interference plus noise, $I + N$, it is assumed that interference is independent of noise. The pdf of the sum of two independent random variables U and V , each of which has a probability density function, is the convolution of their individual density functions:

$$f_{U+V}(z) = \int f_U(z-x)f_V(x)dx \quad (34)$$

therefore the pdf of $I + N$, $f_{I+N}(z)$ is given by:

$$f_{I+N}(z) = \int_0^\infty f_I(z-x)f_N(x)dx \quad (35)$$

where $f_N(x) = \frac{e^{-\frac{1}{2}(\frac{x}{\sigma_N})^2}}{\sqrt{2\pi}\sigma_N}$. Utilizing (9), the pdf of the SINR is given by :

$$f_{\text{SINR}}(\nu) = \int_0^\infty f_{I+N}(z)f_S(\nu z)zdz \quad (36)$$

For the special case where the noise approaches zero, the pdf of the noise is represented as delta function or also known as, a unit impulse function, around zero. Therefore (35) can be rewritten as:

$$f_{I+N}(z) = \int_0^\infty f_I(z-x)f_N(x)dx = \int_0^\infty f_I(z-x)\delta(x)dx = f_I(z) \quad (37)$$

Thus

$$f_{\text{SINR}}(\nu) = \int_0^\infty f_{I+N}(z)f_S(\nu z)zdz = \int_0^\infty f_I(z)f_S(\nu z)zdz \quad (38)$$

by the definition given in (9), the $f_{\text{SINR}}(\nu)$ given in (38) is the pdf of the SIR $\frac{S}{I}$. Therefore, when the noise approaches to zero, the pdf of the SINR given in (36) reduces to the pdf of SIR given in (26).

This sub-section has presented the pdf of the SINR as an extension to the pdf of the SIR. To validate the analytically derived SINR pdf, it is important to show that the core derivation, SIR derivation, is valid. The next sub-section validates the derivation through comparative numerical simulations of the SIR. The results presented were obtained using the adaptive Simpson quadrature numerical integration of the SIR.

5. Results and discussion

Monte Carlo simulations are carried out in order to validate the analytically derived pdf results. Fig. 2 and 3 show plots of the pdf of the SIR $f_{\text{SIR}}(\zeta)$ for different scenarios. The results presented in Figs. 2 - 4 show that the analytical pdf is in good agreement with the Monte Carlo simulation. The parameters used for the shadow fading, channel standard deviation and path loss exponents reflect a realistic deployment scenario for users moving at a speed of 25 to 40 km/h Eltahir (2007). All simulations assume a channel envelope with a Nakagami-m distribution with different m parameter, which corresponds to different fading scenario. These parameters are summarized in Tables 1- 3.

Fig. 2 depicts three different plots depending on the $\frac{R_{\text{ac}}}{R_c}$. As the cell radius R_c increases there is a significant cell overlap leading to high mean value of interference which in turn leads to lower SIR mean value. Therefore, as the ratio of the cell radius to the accessibility radius approaches to one, the pdf is skewed towards smaller SIR. These plots show that the node with the lowest cell radius, $R_c = 100$ m, has the highest SIR mean.

Fig. 3 shows the effect of different environments on the pdf of the SIR. The figure presents plots from an *ad hoc* free space outdoor deployment with line of sight scenario on the desired link, $\gamma = 2$ and $m = 3$, to the most severe non-line-of-sight scenario of obstructed indoor (in building) environment, $\gamma = 4$ and $m = 0.5$. The radius of the cell, R_c , has been set to 100 m, which is considered a good configuration example for *ad hoc* networks. The accessibility radius, R_{ac} is assumed to be 500 m. The results illustrate that the node with the best line-of-sight (LOS) link, $\gamma_{xz} = 2$ and $\gamma_{yz} = 4$, has the highest mean SIR value and the biggest variance or spread. While the node with the most obstructed inbuilding environment, exhibits the lowest mean and the smallest variance or spread of all. These can be attributed to the higher interference contribution of interfering node in NLOS link than those in LOS condition.

Fig. 4 presents the cumulative density function of the SIR. The simulation parameters are summarized in table 3. From Fig. 4 it can be observed that for a target SIR of 25 dB, being a reasonable assumption for 64-QAM modulation, the probability that the SIR exceeds the target SIR in the most severe non-line-of-sight scenario of obstructed indoor (in building) environment is about 10% resulting in a high outage probability enforcing the use of lower order modulation schemes. On the other hand, for the link with best LOS condition of outdoor free space environment the probability that the SIR exceeds the target SIR is 85% allowing the use of higher order modulation. Therefore from the results in Fig. 4, it can be deduced that the analytical work presented in this chapter can be used in determining the boundaries for varying the modulation order. A similar work of determining the boundaries for adaptive modulation was presented by Goldsmith *et al.* Alouini & Goldsmith (2000) assuming Nakagami distribution thus ignoring the shadowing effect, the pdf presented here can be used to extend the results presented in Alouini & Goldsmith (2000).

6. Conclusion

The main contribution of this chapter is the derivation of the pdf of the SIR for cross layer design without recourse to Monte Carlo simulations. The derivation was carried out using a generalized path loss model that accounts for both large and small scale path loss. The use of Nakagami- m distribution for the fading channel gives the flexibility to use Rayleigh or different channel fading models for the desired and interfering links. The results obtained show excellent agreement with the Monte Carlo based results. The SIR derivation was in turn used to derive the pdf of the SINR. The SINR derivation is important in non-interference limited environment. These derivations can be further used in applications where the knowledge of SIR is necessary, such as link adaptation algorithms and cognitive radio design. The analytical derivation of the pdf from a single interferer in this chapter lays a solid foundation to calculate the statistics from multiple interferers.

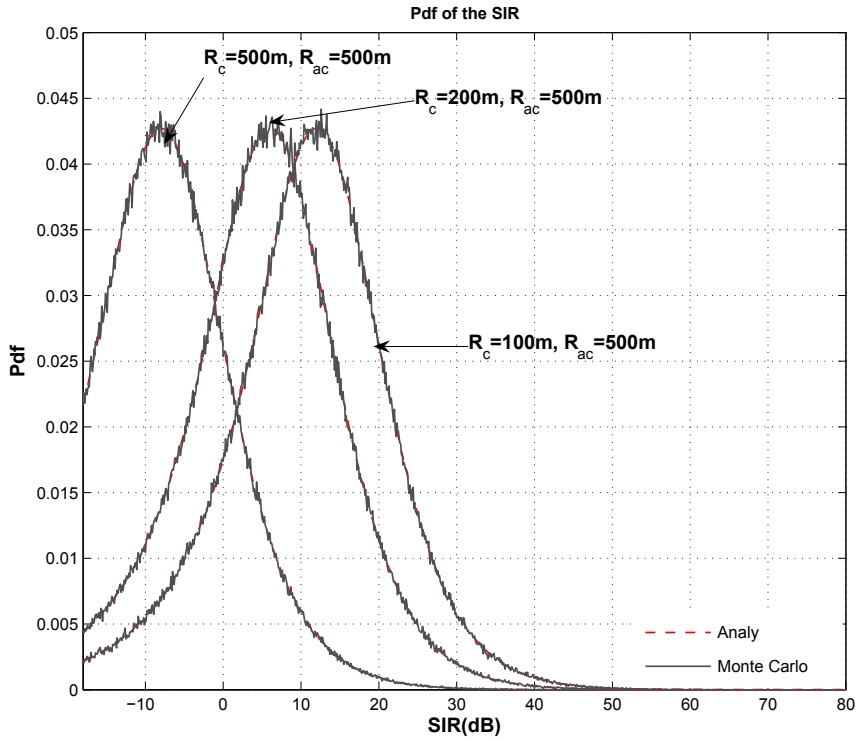


Fig. 2. Plots of the pdf of the SIR for different values of cell radius

Parameter	Values
R_c	100 m
R_{ac}	500 m
v_{xz}	6 dB
v_{yz}	10 dB
γ_{xz}	2
γ_{yz}	4
m_{xz}	5
m_{yz}	0.5
Ω_{xz}	4 dB
Ω_{yz}	6 dB

Table 1. System parameters for Fig. 2(varying cell and accessibility radius)

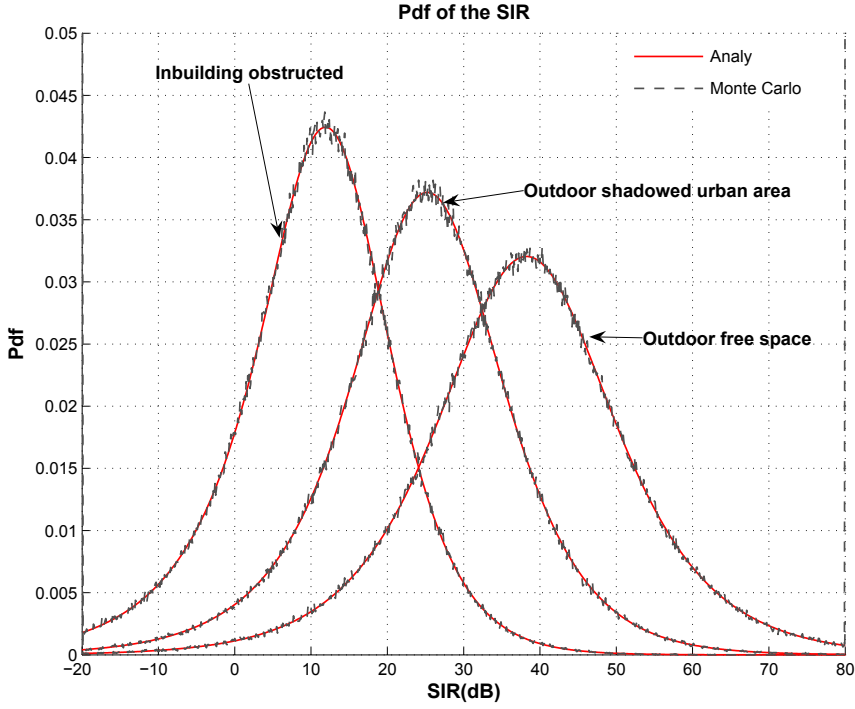


Fig. 3. Plots of the pdf of the SIR for different environments

Parameter	Inbuilding obstructed	Outdoor shadowed urban	Outdoor free space
R_c	100 m	100 m	100 m
R_{ac}	500 m	500 m	500 m
v_{xz}	10 dB	8 dB	10 dB
v_{yz}	10 dB	10 dB	10dB
γ_{xz}	4	3	4
γ_{yz}	4	4	4
m_{xz}	3	1	0.5
m_{yz}	0.5	0.5	0.5
Ω_{xz}	4 dB	4 dB	4 dB
Ω_{yz}	6 dB	4 dB	6 dB

Table 2. System parameters for Fig. 3

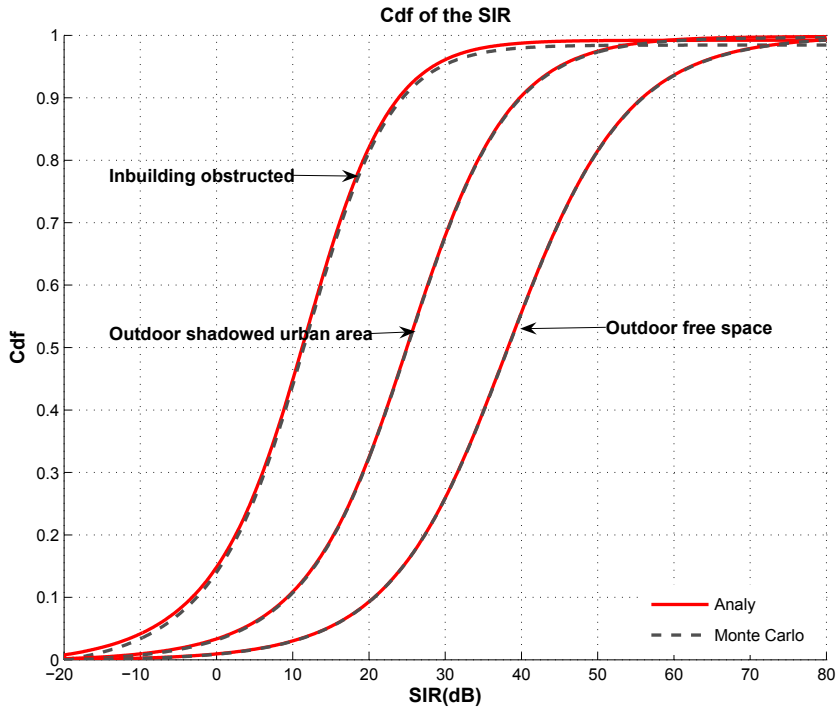


Fig. 4. Plots of the pdf of the SIR for different environments

Parameter	Inbuilding obstructed	Outdoor shadowed urban	Outdoor free space
R_c	100 m	100 m	100 m
R_{ac}	500 m	500 m	500 m
v_{xz}	10 dB	8 dB	10 dB
v_{yz}	10 dB	10 dB	10 dB
γ_{xz}	4	3	4
γ_{yz}	4	4	4
m_{xz}	3	1	0.5
m_{yz}	0.5	0.5	0.5
Ω_{xz}	4 dB	4 dB	4 dB
Ω_{yz}	6 dB	4 dB	6 dB

Table 3. System parameters for Fig. 4

7. References

- Ahn, G.-S., Campbell, A., Veres, A. & Sun, L.-H. (2002). Support Service Differentiation for Real-Time and Best-Effort Traffic in Stateless Wireless Ad Hoc Networks (SWAN), *IEEE Trans. Mobile Comput.* **1**(3): 192–207.
- Alouini, M.-S. & Goldsmith, A. (1997). Capacity of Nakagami Multipath Fading Channels, *Proc. of the IEEE Vehicular Technology Conference(VTC)*, Vol. 1, Arizona, USA, pp. 358–362.
- Alouini, M.-S. & Goldsmith, A. (2000). Adaptive Modulation over Nakagami Fading Channels, *Kluwer Journal on Wireless Commun.* **13**(1–2): 119–143.
- Eltahir, I. (2007). The Impact of Different Radio Propagation Models for Mobile Ad hoc Networks (MANET) in Urban Area Environment, *Proc. of the International Conference on Wireless Broadband and Ultra Wideband Communications (AusWireless)*, Sydney, Australia, pp. 30–30.
- Nakagami, M. (1960). The m-distribution: A General Formula of Intensity Distribution, *Statistical Methods of Radio Wave Propagation*, W. C. Hoffman, Ed., New York, Pergamon pp. 3–36.
- Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes*, 3 edn, McGraw–Hill.
- Rappaport, T. S. (2001). *Wireless Communications: Principles and Practice*, 2 edn, Prentice Hall PTR.
- Tellambura, C. (1999). Cochannel Interference Computation for Arbitrary Nakagami Fading, *IEEE Trans. Veh. Technol.* **48**(2): 487–489.
- Xia, X. & Liang, Q. (2005). Bottom-up Cross-layer Optimization for Mobile ad hoc Networks, *Proc. of the IEEE Military Communications Conference (MILCOM)*, Vol. 4, Atlantic City, USA, pp. 2624–2630.
- Yeh, E. & Cohen, A. (2003). A Fundamental Cross-layer Approach to Uplink Resource Allocation, *Proc. of the IEEE Military Communications Conference (MILCOM)*, Vol. 1, Monterey, CA, pp. 699–704.
- Yun, Z. & Iskander, M. (2004). Progress in Modeling Challenging Propagation Environments, *Proc. of the IEEE Antennas and Propagation Society International Symposium*, Vol. 4, California, USA, pp. 3637–3640.
- Zhang, Q. (1996). Outage Probability in Cellular Mobile Radio Due to Nakagami Signal and Interferers With Arbitrary Parameters, *IEEE Trans. Veh. Technol.* **45**(2): 364–372.
- Zorzi, M. (1997). On the Analytical Computation of the Interference Statistics with Applications to the Performance Evaluation of Mobile Radio Systems, *IEEE Trans. Commun.* **45**(1): 103–109.

The Impact of Fixed and Moving Scatterers on the Statistics of MIMO Vehicle-to-Vehicle Channels

Ali Chelli and Matthias Pätzold
*University of Agder
 Norway*

1. Introduction

In most countries, the reduction in road casualties is a top priority. The intelligent transportation system (ITS) is a national program in the U.S. aiming to improve road safety. In order to deploy the ITS, vehicle-to-vehicle (V2V) communication techniques are needed. The dedicated short range communication (DSRC) standard (ASTM, 2003) is designed for V2V communications. Several task groups are working on this standard including the IEEE 802.11p (802.11p, 2006) and the IEEE 1609.4 (WAVE, 2005).

The statistical properties of V2V channels are different from the conventional fixed-to-mobile channels. Therefore, new channel models are needed for V2V communications. The geometrical two-ring model (Patel et al., 2003), (Pätzold et al., 2005) has been proposed for V2V communications. Unfortunately, this channel model cannot be used to describe propagation conditions along streets for V2V channels. In fact, in such an environment, the wave-guiding along the street has a dominant effect. It was suggested in (Molisch et al., 1999) that the wave-guiding can be implemented by using geometry-based channel models, where the scatterers are located on straight lines. The geometrical street model introduced in (Chelli & Pätzold, 2007) captures the propagation effects if the communicating vehicles are moving along a straight street. The street model has been extended with respect to multiple clusters of scatterers as well as to frequency selectivity in (Chelli & Pätzold, 2008). In (Zajić et al., 2008), a 3D channel model for V2V communications has been proposed. Measurement results in (Zajić et al., 2008) have shown that for vehicles driving in the middle lanes of highways or in urban environment, the double-bounce rays caused by fixed scatterers are dominant. In contrast to our model presented in (Chelli & Pätzold, 2007) and (Chelli & Pätzold, 2008) where single-bounce scattering is assumed, we assume double-bounce scattering for fixed scatterers. Double-bounce models are fundamentally different from single-bounce models. In fact, for double-bounce models the angles of departure (AoD) and the angles of arrival (AoA) are independent. This is in contrast to single-bounce models where the AoD and the AoA are closely related. Due to this dissimilarity, the statistical properties of double-bounce models and single-bounce models are different. Therefore, double-bounce models should be studied carefully.

Furthermore, the presence of moving scatterers in a highway environment has a big impact on the channel behaviour. For this reason, we study the effect of passing vehicles on the channel statistical properties. Measurement results in (Millott, 1994) have shown that the amplitude of waves scattered from more than one vehicle is small and the practical impact of vehicular

scattering is confined to single-bounce rays. Therefore, double-bounce scattering from moving scatterers has been neglected in our model. When scatterers are moving with a high speed relatively to the transmitter (receiver) the AoD (AoA) becomes time-variant resulting in a non-stationary channel model. However, when vehicles are facing road congestion, the relative speed of the cars in the vicinity of the transmitter or the receiver is low. In such conditions, we can still consider the AoD and the AoA as non-time-variant during a sufficiently large period of time. This assumption can be accepted especially if the scatterers are moving in the same direction as the transmitter and the receiver.

The remainder of the chapter is organized as follows. In Section II, the geometrical street model is presented. Based on this geometrical model, we derive a reference model in Section III. In Section IV, we study the correlation properties of the proposed channel model. Numerical results of the correlation functions are presented in Section V to validate all theoretical results by simulations. Finally, Section VI provides some concluding remarks.

2. The Geometrical Street Model

A typical highway propagation environment for V2V communication is presented in Fig. 1. The highway encompasses three lanes used for traffic in the same direction. We can distinguish between two types of scatterers namely fixed scatterers and moving scatterers. The fixed scatterers are represented by the buildings located on both sides of the street, while the moving scatterers are the vehicles in the vicinity of the transmitter MS_T and the receiver MS_R . In order to be able to develop an appropriate channel model for the propagation scenario presented in Fig. 1, we first need to produce a representative geometrical model for such an environment. Towards this aim, we model each building by a cluster of scatterers located on a straight line on the left or right hand side of the street. A vehicle can be modeled by a cluster of scatterers located on a line as well. The fixed clusters are represented by solid lines while the moving clusters are represented by dashed lines. The geometrical street model encompassing fixed and moving scatterers is illustrated in Fig. 2. The fixed scatterers around the transmitter (receiver) are denoted by S_m^T (S_n^R). The moving scatterers are designated by S_p^M . The propagation environment encompasses \mathcal{C}^T (\mathcal{C}^R) fixed clusters around the transmitter (receiver) and \mathcal{C}^M moving clusters. For fixed clusters, the AoD is referred to as α_m^T , whereas the AoA is denoted by β_n^R . For moving clusters, the symbols α_p^M and β_p^M stand for the AoD and the AoA, respectively. It has to be noted that for fixed clusters the AoD and the AoA are independent since double-bounce scattering is assumed. For moving clusters, the AoD and the AoA are closely related due to the single-bounce scattering assumption. All scatterers belonging to a given moving cluster have the same velocity v_S and the same direction of motion ϕ^S . The transmitter and the receiver are moving with velocity v_T and v_R , respectively. The angle of motion of the transmitter and the receiver w.r.t the x -axis is referred to as ϕ^T and ϕ^R , respectively. Moreover, the transmitter (receiver) is equipped with M_T (M_R) antenna elements. The antenna element spacing at the transmitter and the receiver antenna are denoted by δ_T and δ_R , respectively. The angle γ_T (γ_R) describes the tilt angle of the transmit (receive) antenna array.

3. The Reference Model

Starting from the geometrical model shown in Fig. 2, we derive a reference model for the MIMO V2V channel. First, we consider the case where we have one moving cluster and two fixed clusters: one cluster is near to the transmitter and the other cluster is close to the

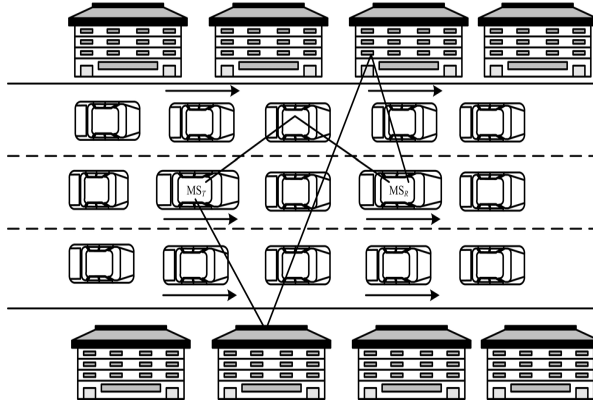


Fig. 1. A highway propagation environment for vehicle-to-vehicle communications under congestion conditions.

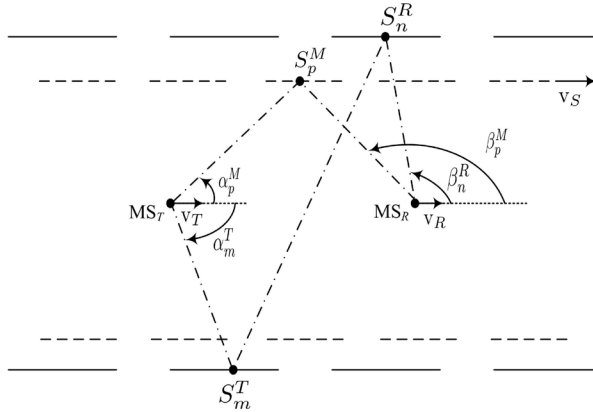


Fig. 2. The geometrical street model encompassing moving and fixed scatterers.

receiver. The total number of fixed scatterers around the transmitter (receiver) is denoted by M (N), while the number of moving scatterers is referred to as P . The complex channel gain $g_{kl}(t)$ describing the link between the l th transmit antenna element A_l^T ($l = 1, 2, \dots, M_T$) and the k th receive antenna element A_k^R ($k = 1, 2, \dots, M_R$) of the underlying $M_T \times M_R$ MIMO V2V channel model can be expressed as $g_{kl}(t) = g_{kl}^F(t) + g_{kl}^M(t)$. The term $g_{kl}^F(t)$ stands for the channel gain due to double-scattering from the fixed clusters. The channel gain caused by the moving cluster is denoted by $g_{kl}^M(t)$. We assume that the line-of-sight component is obstructed. Next, we derive analytical expressions of the channel gains $g_{kl}^F(t)$ and $g_{kl}^M(t)$.

3.1 The Channel Gain Due to Fixed Scatterers

The plane wave emitted from the l th transmit antenna element A_l^T travels over the scatterers S_m^T and S_n^R before impinging on the k th receive antenna element A_k^R . Based on the geometrical model in Fig. 2, the channel gain, due to double-scattering from fixed clusters, $g_{kl}^F(t)$ can be

written as

$$g_{kl}^F(\vec{r}_T, \vec{r}_R) = \sum_{m,n=1}^{M,N} c_{mn} e^{j(\theta_{mn} + \vec{k}_m^T \cdot \vec{r}_T - \vec{k}_n^R \cdot \vec{r}_R - k_0 d_{mn})} \quad (1)$$

where c_{mn} and θ_{mn} stand for the joint gain and the joint phase shift resulting from the interaction with the fixed scatterers S_m^T and S_n^R . The joint channel gain can be written as $c_{mn} = 1/\sqrt{MN}$, while the joint phase shift can be expressed as $\theta_{mn} = (\theta_m + \theta_n) \bmod 2\pi$, where mod stands for the modulo operation. The terms θ_m and θ_n are the phase shifts associated with the scatterers S_m^T and S_n^R , respectively. It has to be noted that θ_m , θ_n , and $\theta_{m,n}$ are independent identically distributed (i.i.d.) random variables uniformly distributed over $[0, 2\pi)$.

The second phase term in (1), $\vec{k}_m^T \cdot \vec{r}_T$, is related to the transmitter movement. The symbol \vec{k}_m^T denotes the wave vector pointing in the propagation direction of the m th transmitted plane wave, and \vec{r}_T is the spatial translation vector of the transmitter. The scalar product $\vec{k}_m^T \cdot \vec{r}_T$ can be expressed as

$$\vec{k}_m^T \cdot \vec{r}_T = 2\pi f_{\max}^T \cos(\alpha_m^T - \phi^T) t \quad (2)$$

where $f_{\max}^T = v_T/\lambda$ stands for the maximum Doppler frequency associated with the mobility of the transmitter. The symbol λ denotes the wavelength.

The third phase term in (1), $\vec{k}_n^R \cdot \vec{r}_R$, is caused by the receiver movement. The symbol \vec{k}_n^R denotes the wave vector pointing in the propagation direction of the n th received plane wave, and \vec{r}_R is the spatial translation vector of the receiver. The scalar product $\vec{k}_n^R \cdot \vec{r}_R$ can be written as

$$\vec{k}_n^R \cdot \vec{r}_R = -2\pi f_{\max}^R \cos(\beta_n^R - \phi^R) t \quad (3)$$

where $f_{\max}^R = v_R/\lambda$ stands for the maximum Doppler frequency due to the receiver movement.

The term $k_0 d_{mn}$ in (1) is associated with the total travelled distance and can be expressed as

$$k_0 d_{mn} = \frac{2\pi}{\lambda} (D_{lm} + D_{mn} + D_{nk}) \quad (4)$$

where D_{lm} denotes the distance from the l th transmit antenna element A_l^T to the scatterer S_m^T . The symbol D_{mn} stands for the distance between the scatterers S_m^T and S_n^R . The term D_{nk} denotes the distance from the scatterer S_n^R to the k th receive antenna element A_k^R . The distances D_{lm} and D_{nk} can be approximated as

$$D_{lm} \approx D_m^T - (M_T - 2l + 1) \frac{\delta_T}{2} \cos(\alpha_m^T - \gamma_T) \quad (5)$$

$$D_{nk} \approx D_n^R - (M_R - 2k + 1) \frac{\delta_R}{2} \cos(\beta_n^R - \gamma_R) \quad (6)$$

where D_m^T denotes the distance from the transmitter to the scatterer S_m^T and D_n^R corresponds to the distance from the receiver to the scatterer S_n^R . After substituting (2)–(6) in (1) the channel gain caused by double-scattering from fixed clusters can be expressed as

$$g_{kl}^F(t) = \sum_{m,n=1}^{M,N} \frac{a_m^T b_n^R c_{mn}^{TR}}{\sqrt{MN}} e^{j(2\pi(f_m^T + f_n^R)t + \theta_{mn})} \quad (7)$$

where

$$a_m^T = e^{j\pi \frac{\delta_T}{\lambda} (M_T - 2l + 1) \cos(\alpha_m^T - \gamma_T)} \quad (8)$$

$$b_n^R = e^{j\pi \frac{\delta_R}{\lambda} (M_R - 2k + 1) \cos(\beta_n^R - \gamma_R)} \quad (9)$$

$$c_{mn}^{TR} = e^{-j\frac{2\pi}{\lambda} (D_m^T + D_{mn} + D_n^R)} \quad (10)$$

$$f_m^T = f_{\max}^T \cos(\alpha_m^T - \phi^T) \quad (11)$$

$$f_n^R = f_{\max}^R \cos(\beta_n^R - \phi^R). \quad (12)$$

It has to be mentioned that the envelope $|g_{kl}^F(t)|$ follows a double Rayleigh distribution since double-bounce scattering is assumed (Salo et al., 2006).

3.2 The Channel Gain Due to Moving Scatterers

The plane wave emitted from the l th transmit antenna element A_l^T travels over the scatterer S_p^M before impinging on the k th receive antenna element A_k^R . Based on the geometrical model in Fig. 2, the channel gain $g_{kl}^M(t)$ of the moving cluster can be expressed as

$$g_{kl}^M(\vec{r}_T, \vec{r}_R, \vec{r}_S) = \sum_{p=1}^P c_p e^{j(\theta_p + \vec{k}_p^T \cdot \vec{r}_T - \vec{k}_p^R \cdot \vec{r}_R - \vec{k}_p^T \cdot \vec{r}_S + \vec{k}_p^R \cdot \vec{r}_S - k_0 d_p)} \quad (13)$$

where c_p and θ_p represent the gain and the phase shift resulting from the interaction with the moving scatterer S_p^M , respectively. The channel gain is given by $c_p = 1/\sqrt{P}$, while the phase shifts θ_p are i.i.d. random variables uniformly distributed over $[0, 2\pi)$. The phase changes $\vec{k}_p^T \cdot \vec{r}_T$ and $\vec{k}_p^R \cdot \vec{r}_R$ are associated with the movement of the receiver and the transmitter, respectively, and can be written as

$$\vec{k}_p^T \cdot \vec{r}_T = 2\pi f_{\max}^T \cos(\alpha_p^M - \phi^T) t \quad (14)$$

$$\vec{k}_p^R \cdot \vec{r}_R = -2\pi f_{\max}^R \cos(\beta_p^M - \phi^R) t. \quad (15)$$

The spatial translation \vec{r}_S of the moving scatterer S_p^M influences the wave emitted from the transmitter resulting in a phase change $\vec{k}_p^T \cdot \vec{r}_S$. Moreover, the scatterer S_p^M interacts with the wave reflected to the receiver resulting in a phase change $\vec{k}_p^R \cdot \vec{r}_S$. These phase changes can be expressed as

$$\vec{k}_p^T \cdot \vec{r}_S = 2\pi f_{\max}^S \cos(\alpha_p^M - \phi^S) t \quad (16)$$

$$\vec{k}_p^R \cdot \vec{r}_S = -2\pi f_{\max}^S \cos(\beta_p^M - \phi^S) t \quad (17)$$

where $f_{\max}^S = v_S/\lambda$ is referred to as the maximum Doppler frequency caused by the moving cluster. Recall that all scatterers S_p^M belonging to the moving cluster have the same speed v_S . The phase change resulting from the total travelled distance d_p can be expressed as

$$k_0 d_p = \frac{2\pi}{\lambda} (D_{lp} + D_{pk}) \quad (18)$$

with

$$D_{lp} \approx D_p^T - (M_T - 2l + 1) \frac{\delta_T}{2} \cos(\alpha_p^M - \gamma_T) \quad (19)$$

$$D_{pk} \approx D_p^R - (M_R - 2k + 1) \frac{\delta_R}{2} \cos(\beta_p^M - \gamma_R) \quad (20)$$

where D_p^T and D_p^R denote the distances from the scatterer S_p^M to the transmitter and the receiver, respectively. After substituting (14)–(20) in (13) the channel gain due to the moving cluster can be written as

$$g_{kl}^M(t) = \sum_{p=1}^P \frac{a_p^M b_p^M c_p^M}{\sqrt{P}} e^{j(2\pi(f_p^{TM} + f_p^{RM} - f_p^{TS} - f_p^{RS})t + \theta_p)} \quad (21)$$

where

$$a_p^M = e^{j\pi \frac{\delta_T}{\lambda} (M_T - 2l + 1) \cos(\alpha_p^M - \gamma_T)} \quad (22)$$

$$b_p^M = e^{j\pi \frac{\delta_R}{\lambda} (M_R - 2k + 1) \cos(\beta_p^M - \gamma_R)} \quad (23)$$

$$c_p^M = e^{-j\frac{2\pi}{\lambda} (D_p^T + D_p^R)} \quad (24)$$

$$f_p^{TM} = f_{\max}^T \cos(\alpha_p^M - \phi^T) \quad (25)$$

$$f_p^{RM} = f_{\max}^R \cos(\beta_p^M - \phi^R) \quad (26)$$

$$f_p^{TS} = f_{\max}^S \cos(\alpha_p^M - \phi^S) \quad (27)$$

$$f_p^{RS} = f_{\max}^S \cos(\beta_p^M - \phi^S). \quad (28)$$

It has to be noted that the AoD α_p^M and the AoA β_p^M are dependent since single-bounce scattering is assumed. The exact relationship between the AoD and the AoA can be found in (Chelli & Pätzold, 2007). The envelope $|g_{kl}^M(t)|$ follows a Rayleigh distribution due to the single-bounce scattering assumption.

3.3 The Multiple-Cluster Channel Gain

The channel gain $g_{kl}(t)$ has been derived assuming a scattering environment with two fixed clusters and one moving cluster. However, in real environment, one can find several buildings and several vehicles near to the mobile transmitter and receiver. Therefore, it is of interest to derive an expression for the channel gain in a multiple-cluster case $z_{kl}(t)$. The environment encompasses \mathcal{C}^T (\mathcal{C}^R) fixed clusters around the transmitter (receiver) and \mathcal{C}^M moving clusters. We added the subscripts $(\cdot)_{c^T}$, $(\cdot)_{c^R}$, and $(\cdot)_{c^M}$ to all affected symbols to distinguish between the fixed clusters around the transmitter, the fixed clusters around receiver, and the moving clusters, respectively. The fixed cluster c^T has a limited length L_{c^T} , it follows that the AoDs α_{m,c^T}^T are restricted to the interval $[\alpha_{\min,c^T}^T, \alpha_{\max,c^T}^T]$. Analogously, the AoDs β_{n,c^R}^R , α_{p,c^M}^M , and β_{p,c^M}^M are confined to the intervals $[\beta_{\min,c^R}^R, \beta_{\max,c^R}^R]$, $[\alpha_{\min,c^M}^M, \alpha_{\max,c^M}^M]$, and $[\beta_{\min,c^M}^M, \beta_{\max,c^M}^M]$, respectively. Moreover, all the AoDs α_{m,c^T}^T ($m = 1, 2, \dots$) have the same distribution and will be noted henceforth by $\alpha_{c^T}^T$. The same statement holds for the angles β_{n,c^R}^R , α_{p,c^M}^M , and β_{p,c^M}^M which will be denoted by $\beta_{c^R}^R$, $\alpha_{c^M}^M$, and $\beta_{c^M}^M$, respectively.

In a multiple-cluster scenario the channel gain describing the link $A_l^T - A_k^R$ can be expressed as

$$z_{kl}(t) = z_{kl}^F(t) + z_{kl}^M(t) \quad (29)$$

where $z_{kl}^F(t)$ is the received diffuse component due to double-scattering from all fixed clusters. The term $z_{kl}^M(t)$ is the received diffuse component caused by single-scattering from all moving clusters. The channel gains $z_{kl}^F(t)$ and $z_{kl}^M(t)$ can be written as

$$z_{kl}^F(t) = \sum_{c^T, c^R=1}^{C^T, C^R} w_{c^T} w_{c^R} g_{kl, c^T, c^R}^F(t) \quad (30)$$

$$z_{kl}^M(t) = \sum_{c^M=1}^{C^M} w_{c^M} g_{kl, c^M}^M(t) \quad (31)$$

where w_{c^T} , w_{c^R} , and w_{c^M} are positive constants representing the weighting factors of the clusters c^T , c^R , and c^M , respectively. We impose the boundary condition $\sum_{c^T, c^R=1}^{C^T, C^R} w_{c^T}^2 w_{c^R}^2 + \sum_{c^M=1}^{C^M} w_{c^M}^2 = 1$, to normalize the mean power of $z_{kl}(t)$ to unity.

4. Correlation Properties

In this section, we derive analytical expressions for the correlation functions of the proposed MIMO V2V channel model, such as the 3D space-time CCF, the temporal ACF, and the 2D space CCF. The 3D space-time CCF $\rho_{kl, k'l'}(\delta_T, \delta_R, \tau)$ can be expressed as

$$\begin{aligned} \rho_{kl, k'l'}(\delta_T, \delta_R, \tau) &:= E \left\{ z_{kl}^*(t) z_{k'l'}(t + \tau) \right\} \\ &= \sum_{c^T, c^R=1}^{C^T, C^R} w_{c^T}^2 w_{c^R}^2 \rho_{kl, k'l', c^T, c^R}^F(\delta_T, \delta_R, \tau) \\ &\quad + \sum_{c^M=1}^{C^M} w_{c^M}^2 \rho_{kl, k'l', c^M}^M(\delta_T, \delta_R, \tau) \end{aligned} \quad (32)$$

where $(\cdot)^*$ denotes the complex conjugation and $E\{\cdot\}$ stands for the expectation operator. The term $\rho_{kl, k'l', c^T, c^R}^F(\delta_T, \delta_R, \tau)$ represents the 3D space-time CCF due to double-scattering from the clusters c^T and c^R . This correlation function can be written as

$$\begin{aligned} \rho_{kl, k'l', c^T, c^R}^F(\delta_T, \delta_R, \tau) &:= E \left\{ (g_{kl, c^T, c^R}^F(t))^* g_{k'l', c^T, c^R}^F(t + \tau) \right\} \\ &= \rho_{c^T}^F(\delta_T, \tau) \cdot \rho_{c^R}^F(\delta_R, \tau) \end{aligned} \quad (33)$$

where

$$\rho_{c^T}^F(\delta_T, \tau) = \int_{\alpha_{\min, c^T}^T}^{\alpha_{\max, c^T}^T} c_{ll'}^F(\delta_T, \alpha_{c^T}^T) e^{j2\pi f^T(\alpha_{c^T}^T)\tau} p_{\alpha_{c^T}^T}(\alpha_{c^T}^T) d\alpha_{c^T}^T \quad (34)$$

and

$$\rho_{c^R}^F(\delta_R, \tau) = \int_{\beta_{\min, c^R}^R}^{\beta_{\max, c^R}^R} d_{kk'}^F(\delta_R, \beta_{c^R}^R) e^{j2\pi f^R(\beta_{c^R}^R)\tau} p_{\beta_{c^R}^R}(\beta_{c^R}^R) d\beta_{c^R}^R \quad (35)$$

are the transmit and the receive correlation functions, respectively, and

$$c_{ll'}^F(\delta_T, \alpha_{c^T}^T) = e^{j2\pi \frac{\delta_T}{\lambda} (l-l') \cos(\alpha_{c^T}^T - \gamma_T)} \quad (36)$$

$$d_{kk'}^F(\delta_R, \beta_{c^R}^R) = e^{j2\pi \frac{\delta_R}{\lambda} (k-k') \cos(\beta_{c^R}^R - \gamma_R)} \quad (37)$$

$$f^T(\alpha_{c^T}^T) = f_{\max}^T \cos(\alpha_{c^T}^T - \phi^T) \quad (38)$$

$$f^R(\beta_{c^R}^R) = f_{\max}^R \cos(\beta_{c^R}^R - \phi^R). \quad (39)$$

The distributions of the AoD $\alpha_{c^T}^T$ and the AoA $\beta_{c^R}^R$ are denoted by $p_{\alpha_{c^T}^T}(\alpha_{c^T}^T)$ and $p_{\beta_{c^R}^R}(\beta_{c^R}^R)$, respectively.

In (32), the term $\rho_{kl, k'l', c^M}^M(\delta_T, \delta_R, \tau)$, which represents the 3D space-time CCF of the moving cluster c^M , can be expressed as

$$\begin{aligned} \rho_{kl, k'l', c^M}^M(\delta_T, \delta_R, \tau) &:= E\left\{ (g_{kl, c^M}^M(t))^* g_{k'l', c^M}^M(t)(t + \tau) \right\} \\ &= \int_{\alpha_{\min, c^M}^M}^{\alpha_{\max, c^M}^M} c_{ll'}^M(\delta_T, \alpha_{c^M}^M) d_{kk'}^M(\delta_R, g(\alpha_{c^M}^M)) \\ &\quad e^{j2\pi (f^{TM}(\alpha_{c^M}^M) + f^{RM}(g(\alpha_{c^M}^M)) - f^{TS}(\alpha_{c^M}^M))\tau} \\ &\quad e^{-j2\pi f^{RS}(g(\alpha_{c^M}^M))\tau} p_{\alpha_{c^M}^M}(\alpha_{c^M}^M) d\alpha_{c^M}^M \end{aligned} \quad (40)$$

where

$$c_{ll'}^M(\delta_T, \alpha_{c^M}^M) = e^{j2\pi \frac{\delta_T}{\lambda} (l-l') \cos(\alpha_{c^M}^M - \gamma_T)} \quad (41)$$

$$d_{kk'}^M(\delta_R, g(\alpha_{c^M}^M)) = e^{j2\pi \frac{\delta_R}{\lambda} (k-k') \cos(g(\alpha_{c^M}^M) - \gamma_R)} \quad (42)$$

$$f^{TM}(\alpha_{c^M}^M) = f_{\max}^T \cos(\alpha_{c^M}^M - \phi^T) \quad (43)$$

$$f^{RM}(g(\alpha_{c^M}^M)) = f_{\max}^R \cos(g(\alpha_{c^M}^M) - \phi^R) \quad (44)$$

$$f^{TS}(\alpha_{c^M}^M) = f_{\max}^S \cos(\alpha_{c^M}^M - \phi^S) \quad (45)$$

$$f^{RS}(g(\alpha_{c^M}^M)) = f_{\max}^S \cos(g(\alpha_{c^M}^M) - \phi^S). \quad (46)$$

The function $g(\cdot)$ in (40) expresses the exact relationship between the AoD $\alpha_{c^M}^M$ and the AoA $\beta_{c^M}^M$. An expression for $g(\cdot)$ can be found in (Chelli & Pätzold, 2007).

The temporal ACF $r_{z_{kl}}(\tau)$ of the channel gain $z_{kl}(t)$ is defined as $r_{z_{kl}}(\tau) := E\{z_{kl}^*(t)z_{kl}(t+\tau)\}$ (Papoulis & Pillai, 2002). The temporal ACF $r_{z_{kl}}(\tau)$ can be deduced from the 3D space-time CCF $\rho_{kl,k'l'}(\delta_T, \delta_R, \tau)$ by setting the antenna element spacings δ_T and δ_R to zero, i.e.,

$$\begin{aligned} r_{z_{kl}}(\tau) &= \rho_{kl,k'l'}(0, 0, \tau) \\ &= \sum_{c^T, c^R=1}^{C^T, C^R} w_{c^T}^2 w_{c^R}^2 \rho_{kl,k'l',c^T,c^R}^F(0, 0, \tau) \\ &\quad + \sum_{c^M=1}^{C^M} w_{c^M}^2 \rho_{kl,k'l',c^M}^M(0, 0, \tau). \end{aligned} \quad (47)$$

The 2D space CCF $\rho_{kl,k'l'}(\delta_T, \delta_R)$ is defined as $\rho_{kl,k'l'}(\delta_T, \delta_R) := E\{z_{kl}^*(t)z_{k'l'}(t)\}$. Alternatively, the 2D space CCF $\rho_{kl,k'l'}(\delta_T, \delta_R)$ can be derived from the 3D space-time CCF $\rho_{kl,k'l'}(\delta_T, \delta_R, \tau)$ by setting τ to zero, i.e.,

$$\begin{aligned} \rho_{kl,k'l'}(\delta_T, \delta_R) &= \rho_{kl,k'l'}(\delta_T, \delta_R, 0) \\ &= \sum_{c^T, c^R=1}^{C^T, C^R} w_{c^T}^2 w_{c^R}^2 \rho_{kl,k'l',c^T,c^R}^F(\delta_T, \delta_R, 0) \\ &\quad + \sum_{c^M=1}^{C^M} w_{c^M}^2 \rho_{kl,k'l',c^M}^M(\delta_T, \delta_R, 0). \end{aligned} \quad (48)$$

5. Numerical Results

In this section, we confirm the validity of the analytical expressions presented in the previous section by simulations making use of the sum-of-cisoids method. The simulation models for moving and fixed scatterers are designed using the modified method of equal area (MMEA) proposed in (Gutiérrez & Pätzold, 2007). In order to model all fixed clusters, 50 cisoids are used for the simulation model. The same number of cisoids is used to model all moving clusters. The propagation environment contains six moving clusters: three clusters are located on the right side of the transmitter and the receiver, while the remaining clusters lie on the left side. Each moving cluster has a length of 5 m and is separated by a distance of 45 m from its neighbour clusters. The distance between the transmitter and the moving scatterers located on the left and the right side is set to 3 m. For the fixed clusters, we consider a propagation environment encompassing three clusters on each side of the transmitter. Each cluster has a length of 2 m and is separated by a distance of 34 m from its neighbour clusters. The same number of fixed clusters is considered around the receiver. The distance between the transmitter and the fixed scatterers on the left and right side is set to 300 m. The distance between the transmitter and the receiver is equal to 100 m. The transmitter and the receiver have a speed of 50 km/h and equal angles of motion $\phi^T = \phi^R = 0$. The transmitter and the receiver antenna tilts γ_T and γ_R are set to $\pi/2$. We consider the case of non-isotropic scattering conditions. The AoDs $\alpha_{c^T}^T$ and $\alpha_{c^M}^M$ are uniformly distributed over the intervals $[\alpha_{\min,c^T}^T, \alpha_{\max,c^T}^T]$ and $[\alpha_{\min,c^M}^M, \alpha_{\max,c^M}^M]$, respectively. The uniform distribution is also assumed for the AoAs $\beta_{c^R}^R$ over the interval $[\beta_{\min,c^R}^R, \beta_{\max,c^R}^R]$.

We present some illustrative examples for the temporal ACF $r_{z_{kl}}^M(\tau)$ of the moving clusters in Fig. 3. We study the influence of the speed of the vehicles in the vicinity of the transmitter and

the receiver on the channel behaviour. The term v_S in Fig. 3 denotes the speed of the vehicles on the left and right side. From Fig. 3, we can notice that as the speed v_S decreases, the coherence time of the channel increases. It is well known that the coherence time indicates whether we are facing a fast or a slow fading. As the speed of the vehicles relatively to the transmitter and the receiver decreases the channel changes more slowly. A good fitting between the simulation results and the theoretical results can be observed in Fig. 3. In Fig. 4, we show the

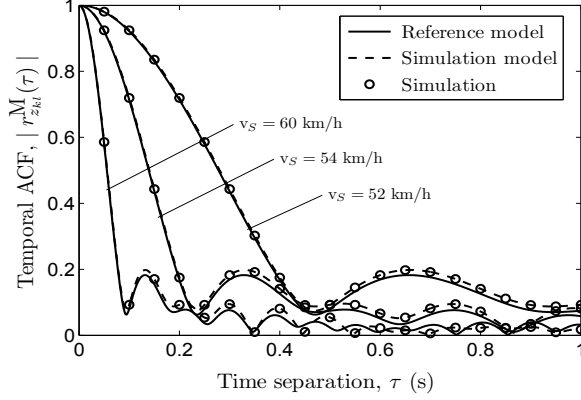


Fig. 3. The absolute value of the temporal ACF $r_{z_{kl}}^M(\tau)$ associated with moving scatterers for various values of the velocity v_S of the surrounding vehicles.

numerical results obtained for the 2D space CCF $\rho_{11,22}^M(\delta_T, \delta_R)$ caused by moving scatterers. It could be seen from Fig. 4 that the cross-correlation function decreases as we increase the antenna spacings δ_T and δ_R . However, the decay of the 2D space CCF $\rho_{11,22}^M(\delta_T, \delta_R)$ is faster along the δ_R direction. Hence, a small antenna spacing at the receiver side guarantees a diversity gain, but at the receiver side, we need a larger spacing to get non-correlated channels. In

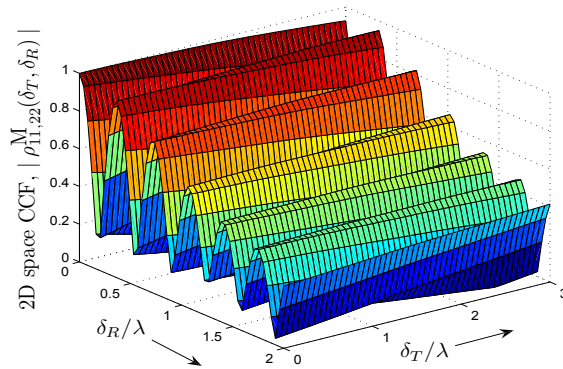


Fig. 4. The absolute value of the 2D space CCF $\rho_{11,22}^M(\delta_T, \delta_R)$ of the reference model caused by moving scatterers.

Fig. 5, we illustrate the numerical results for the transmit correlation function $\rho_T^F(\delta_T, \tau)$ of the reference model resulting from fixed scatterers. The obtained results are confirmed by simulation in Fig. 6. Similar results have been found for the receive correlation function $\rho_R^F(\delta_R, \tau)$ associated with fixed scatterers since the same setting for the scatterers is considered around the transmitter and the receiver. Limited space prevents us from including these results.

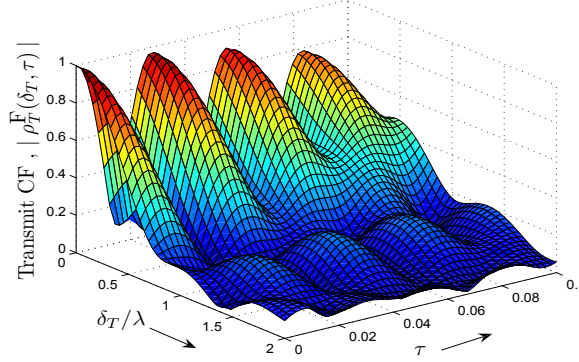


Fig. 5. The transmit correlation function $\rho_T^F(\delta_T, \tau)$ of the reference model due to fixed scatterers.

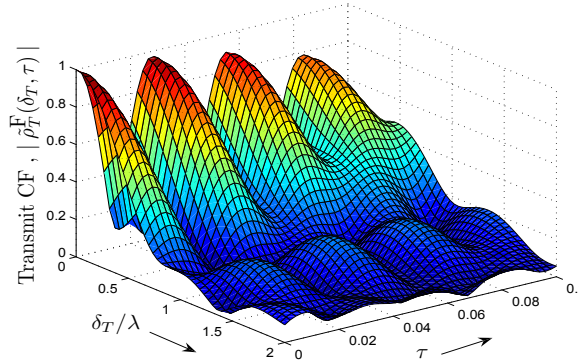


Fig. 6. The transmit correlation function $\tilde{\rho}_T^F(\delta_T, \tau)$ of the simulation model related to fixed scatterers.

6. Conclusion

In this chapter, we have presented a narrowband MIMO V2V channel model, where both the impact of fixed and moving scatterers were taken into account. Double-bounce scattering is assumed for the fixed scatterers, while single-bounce scattering is considered for the moving

scatterers. For reasons of brevity, we have restricted our investigations to non-line-of-sight situations. A reference model has been derived starting from the geometrical street model. The statistical properties of the proposed channel model have been studied. We have provided analytical expressions for the 3D space-time CCF, the temporal ACF, and the 2D space CCF. Supported by our analysis, we are convinced that the effect of moving scatterers on the statistics of V2V MIMO channels cannot be neglected. The investigation of the impact of moving scatterers have revealed that as the speed of the vehicles in the vicinity of the transmitter and the receiver decreases, the channel coherence time increases. The proposed channel model is suitable for a highway environment under congestion conditions. In such conditions, the low relative speed of the vehicles in the vicinity of the transmitter and the receiver allows us to consider that the AoD and the AoA seen from moving clusters are non-time-variant during a sufficiently large period of time. Actually, if the AoD and the AoA are time-variant, the channel model becomes non-stationary. The latter aspect will be investigated in future work.

7. References

- 802.11p (2006). Wireless access in vehicular environment (WAVE) in standard 802.11 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, *Technical Report p/D1.0*, IEEE 802.11.
- ASTM (2003). Standard specification for telecommunications and information exchange between roadside and vehicle systems - 5 ghz band dedicated short range communications (DSRC) medium access control (MAC) and physical layer (PHY) specifications. ASTM E2213-03.
- Chelli, A. & Pätzold, M. (2007). A MIMO mobile-to-mobile channel model derived from a geometric street scattering model, *Proc. 4th IEEE International Symposium on Wireless Communication Systems, ISWCS'07*, Trondheim, Norway, pp. 792–797.
- Chelli, A. & Pätzold, M. (2008). A wideband multiple-cluster MIMO mobile-to-mobile channel model based on the geometrical street model, *Proc. 19th IEEE Int. Symp. on Personal, Indoor and Mobile Radio Communications, IEEE PIMRC 2008*, Cannes, France.
- Gutiérrez, C. A. & Pätzold, M. (2007). Sum-of-sinusoids-based simulation of flat fading wireless propagation channels under non-isotropic scattering conditions, *Proc. 50th IEEE Global Telecommunications Conference, GLOBECOM 2007*, Washington, DC, USA, pp. 3842–3846.
- Millott, L. (1994). The impact of vehicular scattering on mobile communications, *Proc. 44th IEEE Veh. Technol. Conf., VTC'1994*, Stockholm, Sweden, pp. 1733–1736.
- Molisch, A., Kuchar, A., Laurila, J., Hugl, K. & Bonek, E. (1999). Efficient implementation of a geometry-based directional model for mobile radio channels, *Proc. 50th IEEE Veh. Technol. Conf., VTC'1999-Fall*, Amsterdam, Netherlands, pp. 1449–1453.
- Papoulis, A. & Pillai, S. U. (2002). *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York.
- Patel, C. S., Stüber, G. L. & Pratt, T. G. (2003). Simulation of Rayleigh faded mobile-to-mobile communication channels, *Proc. 58th IEEE Veh. Technol. Conf., VTC'03*, Orlando, FL, USA, pp. 163–167.
- Pätzold, M., Hogstad, B. O., Youssef, N. & Kim, D. (2005). A MIMO mobile-to-mobile channel model: Part I - The reference model, *Proc. 16th IEEE Int. Symp. on Personal, Indoor and Mobile Radio Communications, PIMRC 2005*, Berlin, Germany, pp. 573–578.

- Salo, J., El-Sallabi, H. & Vainikainen, P. (2006). Impact of double-Rayleigh fading on system performance, *Proc. 1st IEEE Int. Symp. on Wireless Pervasive Computing, ISWPC 2006*, Phuket, Thailand.
- WAVE (2005). Wireless access in vehicular environments (WAVE) channel coordination. IEEE P1609.4/D05.
- Zajić, A. G., Stüber, G. L., Pratt, T. G. & Nguyen, S. (2008). Statistical modeling and experimental verification of wideband MIMO mobile-to-mobile channels in highway environments, *Proc. 44th IEEE Veh. Technol. Conf., VTC'1994*, Cannes, France.

Planar Antenna Array Hybrid Beamforming for SDMA in Millimeter Wave WPAN

Sau-Hsuan Wu, Lin-Kai Chiu, Ko-Yen Lin and Ming-Chen Chiang
Institute of Communication Engineering
Department of Electrical Engineering
National Chiao Tung University
Hsinchu, Taiwan

1. Introduction

The increasing demands on bandwidth for personal and indoor wireless multimedia applications have driven the research and development for a new generation of broadband wireless personal area network (WPAN) Alliance (n.d.); *IEEE 802.11 WLAN Very High Throughput in 60GHz* (n.d.); *IEEE 802.15 WPAN Millimeter Wave Alternative PHY Task Group (TG3c)* (n.d.); *WirelessHD* (n.d.). This new WPAN is intended to support data rate up to 5Gbps or more and allows for wireless interconnection among devices, such as laptops, camcorders, monitors, DVD players and cable boxes, etc. Moreover, it could also serve as a wireless alternative to the High-Definition Multimedia Interface (HDMI).

In addition to its very large bandwidth, the new WPAN also demands for short-range and secure wireless connections. The characteristics of broad unlicensed bandwidth FCC (2004), high penetration loss Smulders (1995; 2002) and significant oxygen absorption Anderson & Rappaport (2004) at 60GHz radio make it an ideal wireless interface for the next generation WPAN. Furthermore, the millimeter wavelength of 60GHz radio also makes it possible to use tens of tiny antennas to steer radio signals with high directivity to the intended receivers. This feature of high-directivity beam pattern not only improves the wireless link quality Balanis & Ioannides (2007) but also increases the spatial reuse factor, allowing for multiple users to gain access to the wireless channel at the same frequency and time. In view of the great potential of 60GHz radio on WPAN and the advantages of beamforming (BF) for millimeter wave (mmWave) applications, we study in this article a simple hybrid beamforming (HBF) technique for spatial division multiple access (SDMA) using planar antenna arrays (PAAs).

Digital BF has been used to compensate the rather fixed radiation patterns of switch-beam or beam-selection antennas to create more flexible hybrid beam patterns Celik et al. (2006); Rezk et al. (2005); Zhang et al. (2003). In conjunction with the phase shifters of the element antennas of PAAs, a hybrid type of BF is considered in Smolders & Kant (2000) which exploits the advantage of BF both in the baseband and the radio-frequency (RF) ranges. Motivated by the above results and taking into account the practical limitation and implementation cost of the full digital BF, we study herein a special type of baseband and RF HBF for SDMA that only requires four digital processing paths to support HBF on a 8×8 PAA illustrated in Fig. 1. The entire PAA of Fig. 1 is partitioned into four blocks of patch antennas. Each block is driven by a digital BF weight, while each element patch antenna in a block is equipped with

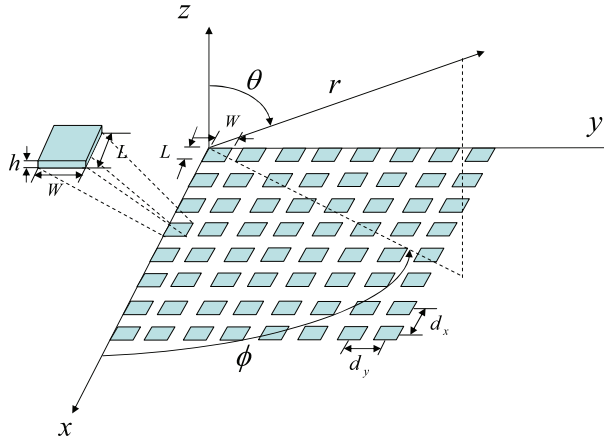


Fig. 1. Antenna arrays of 8×8 planar antennas.

an individual phase shifter. With this configuration, we study the design of HBF for two-user SDMA and compare its performance with that of RF BF only.

The content of this chapter is organized in the following order. First, some concept and the configuration about PAA will be reviewed in Section II and applied to SDMA using reconfigurable PAA. In Section III, the baseband-and-RF hybrid BF (HBF) will be introduced for SDMA, and the linear constrain minimum power (LCMP) digital BF technique will be reapplied to this new setting of PAA HBF over mmWave radio. The signal to interference-plus-noise ratios (SINRs) of users with the aforementioned BF scheme will be investigated in Section IV and the BER simulations will also be conducted on an OFDM-based WPAN to verify the performance of HBF for SDMA over mmWave radio. Conclusions will be drawn in Section V.

2. The configurations of planar antenna arrays

We specify in this section the configurations of planar antenna arrays (PAA) for hybrid BF. Sixty-four identical patch antennas are aligned to form an 8×8 antenna matrix as shown in Fig. 1. Each element antenna is equipped with a phase shifter to maneuver the phase of the signal radiating through it. Given the large number of antennas available for 60GHz applications, it is beneficial to use the antennas to serve multiple users in addition to increasing the received signal to noise ratio (SNR) of a single user. Taking into account the practical limitations of circuit implementations, the arrays of antennas are partitioned into blocks, and each of which is driven by a baseband signal processing path. In other words, the baseband BF weights are applied to the antennas on a block basis. Antennas within the same block are applied the same baseband BF weight. In this section, we characterize the beam pattern of this hybrid type of radio-frequency (RF) and baseband (BB) BF.

To facilitate the analysis and highlight the performance of hybrid BF (HBF), the coupling effects among element antennas are neglected in the sequel. As a result, the total beam pattern of a block of a partition can be expressed as the product of the electric field of a single antenna and the array factor corresponding to the block Balanis (1997).

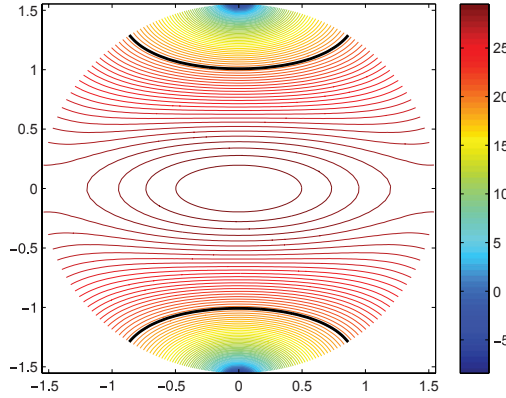


Fig. 2. The contour plot of the antenna pattern when $P=30\text{dB}$.

The far-zone electric field of a single element antenna is given by

$$E(\phi, \theta) = E_\theta \vec{a}_\theta + E_\phi \vec{a}_\phi + E_r \vec{a}_r \quad (1)$$

where

$$E_\theta = j \frac{hWkE_0 e^{-jkr}}{\pi r} \left[\cos \phi \cos X \left(\frac{\sin Y}{Y} \right) \left(\frac{\sin Z}{Z} \right) \right] \quad (2)$$

$$E_\phi = j \frac{hWkE_0 e^{-jkr}}{\pi r} \left[\cos \theta \sin \phi \cos X \left(\frac{\sin Y}{Y} \right) \left(\frac{\sin Z}{Z} \right) \right] \quad (3)$$

and $E_r \cong 0$ as $r \gg \frac{2LW}{\lambda}$ (see Balanis (1997) for the far field definition). The physical meaning of some of the parameters are illustrated in Fig. 1, and E_0 is a constant. For convenience of expression, we also define $X \triangleq \frac{kL}{2} \sin \theta \cos \phi$, $Y \triangleq \frac{kW}{2} \sin \theta \sin \phi$, $Z \triangleq \frac{kh}{2} \cos \theta$ and $k \triangleq \frac{2\pi}{\lambda}$ with λ being the radio wavelength.

The contour plot of the electric field is shown in Fig. 2 as $P|E(\phi, \theta)|^2$ in dB in the cylindrical coordinate, with $P = 30\text{dB}$. The dimensions of the element patch antenna used in the simulation are $L = W = 1\text{mm}$, $h = 0.1\text{mm}$ and the distances between adjacent antennas are set to $dx = dy = 2.5\text{mm}$. The radial coordinate is mapped to the elevation angle θ and the angular coordinate is to the azimuth angle ϕ of the antenna pattern. The vertical coordinate displays the antenna gain in decibel. We note that the antenna pattern is not symmetric with respect to the azimuth angle, ϕ . The pattern is narrower in the direction of $\phi \pm \pi/2$.

The array factor of each block depends on its relative position in the PAA. For the partition shown in Fig. 3, we define an index pair, $(p, q) \in \{0, 1\}^2$, for each block of the PAA. Given the index pair (p, q) of a block, the corresponding array factor follows

$$\begin{aligned} A_{(p,q)}(\phi, \theta) &= e^{jpN(\Psi_x + \beta_{x,(p,q)})} \sum_{n=1}^N e^{j(n-1)(\Psi_x + (-1)^p \beta_{x,(p,q)})} \\ &\quad \times e^{jqM(\Psi_y + \beta_{y,(p,q)})} \sum_{m=1}^M e^{j(m-1)(\Psi_y + (-1)^q \beta_{y,(p,q)})} \end{aligned} \quad (4)$$

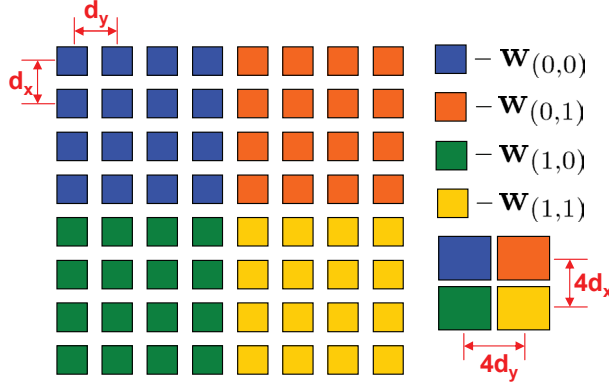


Fig. 3. Partitions of the planar antenna arrays. Patch antennas in different color belong to different block.

where $\Psi_x \triangleq kd_x \cos \phi \sin \theta$ and $\Psi_y \triangleq kd_y \sin \phi \sin \theta$. The distances in the x and y directions between adjacent patch antennas are denoted by d_x and d_y , respectively. And $\beta_{x,(p,q)}$ and $\beta_{y,(p,q)}$ are the corresponding phase differences in the x and y directions between adjacent patch antennas. The number of antennas in the x direction of a block is M and the number of antennas in the y direction is N . Given the desired direction $(\phi_{d,(p,q)}, \theta_{d,(p,q)})$ set for the block (p, q) , $\beta_{x,(p,q)}$ and $\beta_{y,(p,q)}$ are equal to

$$\beta_{x,(p,q)} = (-1)^{p+1} kd_x \sin \theta_{d,(p,q)} \cos \phi_{d,(p,q)} \pm 2c_1 \pi \quad (5)$$

$$\beta_{y,(p,q)} = (-1)^{q+1} kd_y \sin \theta_{d,(p,q)} \sin \phi_{d,(p,q)} \pm 2c_2 \pi \quad (6)$$

where c_1 and c_2 are any integers. It is noted that the array factor (4) is a periodic function with a period of 2π and is zero whenever (ϕ, θ) satisfy either one of the following two conditions

$$\Psi_x + (-1)^p \beta_{x,(p,q)} = 2c_3 \pi / N \quad (7)$$

$$\Psi_y + (-1)^q \beta_{y,(p,q)} = 2c_4 \pi / M \quad (8)$$

when c_3 and c_4 are integers not equal to the multiples of N and M , respectively.

2.1 SDMA using reconfigurable PAA

Adjusting the antenna phases of a block based on (4) ~ (8), the main beams of different blocks can be tuned towards the directions of different users, offering spatial division to support multiple access (SDMA) of users. Despite the array gain provided by (4), the SINR of SDMA also depends on the antenna pattern of (1) and the beam patterns of adjacent users.

Suppose that all patch antennas in Fig. 3 are used to support a single user. The maximum achievable array gain in this case is $20 \log_{10}(64) \doteq 36$ dB when

$$\Psi_x + (-1)^p \beta_{x,(p,q)} = 2c_5 \pi \quad (9)$$

$$\Psi_y + (-1)^q \beta_{y,(p,q)} = 2c_6 \pi, \{c_5, c_6\} \in \mathbb{Z} \quad (10)$$

where \mathbb{Z} stands for the integer number. Fig. 4 shows the contour plot of the corresponding array pattern when $\theta_{d,(0,0)} = 0$. It is clear that the maximum array gain of 36 dB is achieved

at $\theta = \theta_{d,(0,0)} = 0$. Scaling the power by $1/MN$ for each element antenna of the $M \times N$ PAA, the maximum effective array gain is $10 \log_{10}(64) \doteq 18$ dB.

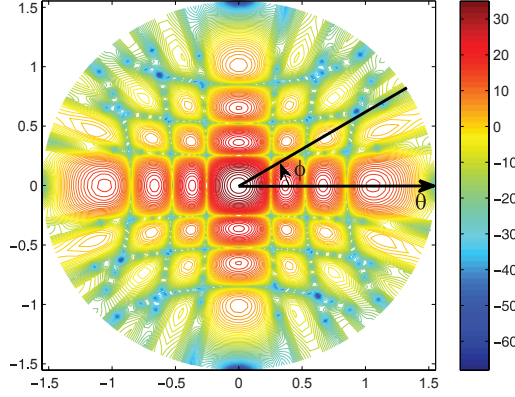


Fig. 4. Contour plot of the array pattern for one-user case when $\theta_{d,(0,0)} = 0$.

On the other hand, if each block in Fig. 3 serves a user with an 4×4 antenna array, the maximal array gain now reduces to a smaller value of $20 \log_{10}(16) \doteq 24$ dB. Except for the smaller array gain, each user's signal is also interfered by the signals of adjacent users. Fig. 5 shows the array factors for the 4-user SDMA based on the partition in Fig. 3. The four beams are pointed toward the elevation angle of $\frac{\pi}{4}$ and the azimuth angles of $\frac{i\pi}{4}$, $i = 1, \dots, 4$, respectively. As can be seen from the figures, the side beams of adjacent users overlap with the main beam of the user of interest, making it difficult to maintain the SINR in practice. To provide a better control for the SINRs of users supported with PAA, a hybrid approach of BF (HBF) is introduced in the next section to take the advantage of baseband BF techniques.

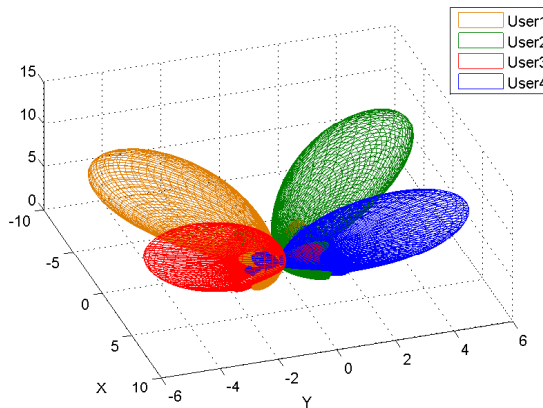


Fig. 5. The array factors of the 4-user SDMA based on the partition in Fig. 3.

3. SDMA using hybrid beamforming

The SDMA method introduced in Section 2.1 is based on phase adjustment with the phase shifter of each element antenna. However, adjusting only the phases of the radio signals sometimes may not be able to achieve the desired SINR for the user of interest, as the beam direction of the user might be severely jammed by the side beams of other users. To overcome this difficulty, baseband BF techniques can be used to jointly steer the beam patterns and suppress the interference for all users. More specifically, in addition to steering the main beam towards the direction of interest, the baseband array factor can be nulled as well in the directions of other users' main beams.

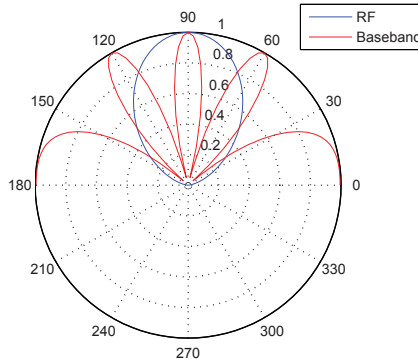


Fig. 6. The polar plot of the HBF pattern with the partition in Fig. 3 when $\theta_d = \pi/2$.

However, it is impractical to apply a baseband BF weight for each element antenna of the 8×8 PAA. Taking into account the implementation cost, each partition of PAA is driven by a common baseband BF weight, while each antenna is still equipped with an individual phase shifter. To distinguish the array factor $B(\phi, \theta)$ formed with the baseband BF weights of a user from the array factor $A(\phi, \theta)$ obtained by tuning the phase of the radiated wave of each antenna, we refer to $B(\phi, \theta)$ as the baseband array factor (BAF) in contrast to the array factor $A(\phi, \theta)$ tuned in the radio-frequency (RF) band.

Now we consider this hybrid type of baseband and RF BF for the simple partition shown in Fig. 3. Suppose that the RF array factor (RAF) for different blocks of a user are the same and pointed to the desired direction of interest, the composite beam pattern of HBF is given by

$$H(\phi, \theta) \triangleq B(\phi, \theta)A(\phi, \theta)E(\phi, \theta) \quad (11)$$

where $A(\phi, \theta)$ is the array factor of the 4×4 antenna arrays. In the extreme case of Fig. 3 that the entire PAA is used to support a single user, the BAF is given by

$$B(\phi, \theta) = \sum_{r=0}^1 \sum_{s=0}^1 w_{(r,s)} e^{j4r\Psi_x} e^{j4s\Psi_y}. \quad (12)$$

where $\Psi_x \triangleq kd_x \cos \phi \sin \theta$ and $\Psi_y \triangleq kd_y \sin \phi \sin \theta$. The enlarged distances between the adjacent effective antennas make $4kd_x = 4kd_y = 4\pi$ in (12) as $dx = dy = \lambda/2$, which in turn

results in the periodic baseband beam pattern of $B(\phi, \theta)$ shown in Fig. 6. The angular coordinate corresponds to the elevation angle θ and the radial coordinate represents the normalized BF gain. Due to the periodic pattern, the product of $B(\phi, \theta)$ and $A(\phi, \theta)$ will yield significant sidelobes on both sides of the main beam. For clarity, the RAF $A(\phi, \theta)$ of the 4×4 block is also shown in Fig. 6. Since the patch antenna has a fixed radiation pattern, its pattern is not shown in the figure.

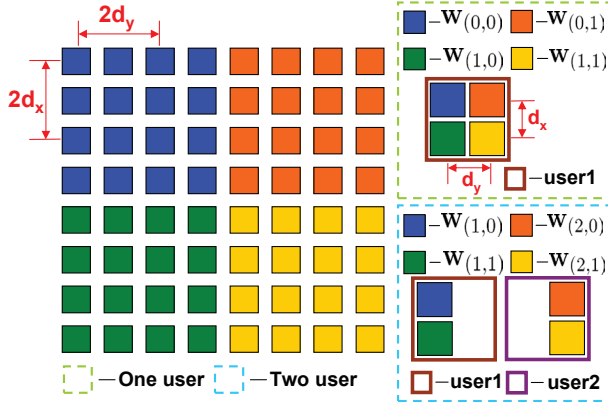


Fig. 7. The partition of the PAA for two-user SDMA, where patch antennas of the same color belong to the same block.

3.1 HBF based on the MD beamforming

According to the configuration of Fig. 3, we now implement HBF for two-user SDMA based on the partition in Fig. 7. The antennas in blue and green colors of Fig. 7 belong to user one, and the antennas in orange and yellow belong to user two. Namely, two BF weights are employed for each user. The resultant BAF for user one and two are given by

$$B_1(\phi, \theta) = w_{(1,0)} + w_{(1,1)} e^{j\Psi_x} \quad (13)$$

$$B_2(\phi, \theta) = w_{(2,0)} e^{j\Psi_y} + w_{(2,1)} e^{j\Psi_x + j\Psi_y}. \quad (14)$$

Now to steer the main beam towards the direction of the user of interest and, in the mean time, to suppress the interference in the direction of the other user, a typical method is the so-called maximum directivity (MD) BF Kuhwald & Boche (1999).

The MD BF basically constructs the baseband BF weights by superposition of the steering vectors

$$\mathbf{s}_1(\phi, \theta) \triangleq [1 \ e^{j\Psi_x}]^T \quad (15)$$

$$\mathbf{s}_2(\phi, \theta) \triangleq [e^{j\Psi_y} e^{j(\Psi_x + \Psi_y)}]^T \quad (16)$$

of user one and two in (13) and (14), respectively. Specifically, the BAFs are expressed as

$$B_1(\phi, \theta) \triangleq \sum_{i=1}^2 b_i [1 \ e^{-j\psi_{xi}}] \mathbf{s}_1. \quad (17)$$

$$B_2(\phi, \theta) \triangleq \sum_{i=1}^2 c_i [e^{-j\Psi_{yi}} e^{-j(\Psi_{xi} + \Psi_{yi})}]^H \mathbf{s}_2 \quad (18)$$

where $\Psi_{xi} \triangleq 4kd_x \cos \phi_i \sin \theta_i$ and $\Psi_{yi} \triangleq 4kd_y \sin \phi_i \sin \theta_i$, and $\{\phi_i, \theta_i\}$ is the desired beam direction of user i . Substituting the constraints of

$$B_m(\phi_i, \theta_i) = \begin{cases} 1, & i = m \\ 0, & i \neq m \end{cases}, \quad i, m \in \{1, 2\}. \quad (19)$$

back into (17) and (18) yields the coefficients b_i and c_i . Furthermore, equating the $B_m(\phi, \theta)$ respectively for $m \in \{1, 2\}$ in (13) ~ (18) results in the baseband BF weights

$$w_{(1,r)} = \sum_{i=1}^2 b_i e^{-jr\Psi_{xi}}, \quad (20)$$

$$w_{(2,r)} = \sum_{i=1}^2 c_i e^{-j\Psi_{xi}} e^{-jr\Psi_{yi}} \quad r \in \{0, 1\}. \quad (21)$$

3.2 HBF based on the linear constrained minimization of power

Though simple and straightforward, the MD BF does not take into account the power consumption in HBF design. A widely used approach for power minimization is the linear constrained minimum power (LCMP) method Trees (2002). To minimize the power consumption of BF and, in the mean time, null the interference in the beam direction of the user of interest, we apply the LCMP subject to (s.t.) constraints similar to that of the MD BF in (19).

Let $u_i(t), i \in \{1, 2\}$ be the transmitted signal of user i , with $E[|u_i(t)|^2] = 1$. The baseband transmitted signal for the two-user SDMA can be modeled as

$$\mathbf{x}(t) = \mathbf{s}_1 u_1(t) + \mathbf{s}_2 u_2(t). \quad (22)$$

where the steering vectors \mathbf{s}_1 and \mathbf{s}_2 are defined in (15) and (16), respectively. To design the BF weight vector \mathbf{w}_i for user i such that the output power and the interference to the beam direction of the other user are both minimized, the LCMP is formulated as

$$\arg \min_{\mathbf{w}_i} \mathbf{w}_i^H \mathbf{S}_x \mathbf{w}_i \quad (23)$$

$$\text{s.t. } \mathbf{w}_i^H \mathbf{C} = \mathbf{e}_i \quad (24)$$

where $\mathbf{S}_x \triangleq E\{\mathbf{x}^2(t)\}$, $\mathbf{C} \triangleq [\mathbf{s}_1(\phi_1, \theta_1), \mathbf{s}_2(\phi_2, \theta_2)]$ with $\{\phi_i, \theta_i\}$ being the desired beam direction of user i , and \mathbf{e}_i is a 1×2 basis vector with 1 in the i th position and the others zero.

The above optimization problem can be easily solved by making use of the Lagrange multiplier as below

$$J = \mathbf{w}_i^H \mathbf{S}_x \mathbf{w}_i + [\mathbf{w}_i^H \mathbf{C} - \mathbf{e}_i^H] \lambda + \lambda^H [\mathbf{C}^H \mathbf{w}_i - \mathbf{e}_i] \quad (25)$$

with the Lagrange multiplier $\lambda \triangleq [\lambda_1, \lambda_2]^T$. Taking the complex gradient of J respect to \mathbf{w}_i^H and setting it to zero yields

$$\mathbf{S}_x \mathbf{w} + \mathbf{C} \lambda = 0 \implies \mathbf{w} = -\mathbf{S}_x^{-1} \mathbf{C} \lambda. \quad (26)$$

Substituting (26) into (24) gives

$$-\lambda^H \mathbf{C}^H \mathbf{S}_x^{-1} \mathbf{C} = \mathbf{e}_i^H \implies \lambda^H = -\mathbf{e}_i^H [\mathbf{C}^H \mathbf{S}_x^{-1} \mathbf{C}]. \quad (27)$$

Furthermore, substituting (27) back into (26), the resultant optimal BF weight vector for user i is given by

$$\mathbf{w}_i^H = \mathbf{e}_i^H [\mathbf{C}^H \mathbf{S}_x^{-1} \mathbf{C}]^{-1} \mathbf{C}^H \mathbf{S}_x^{-1}. \quad (28)$$

4. Computer Simulations

We demonstrate simulation results for the HBF schemes studied in the previous section for SDMA. The transmit SNR in the following simulations is set to 30 dB for each user if no specific description.

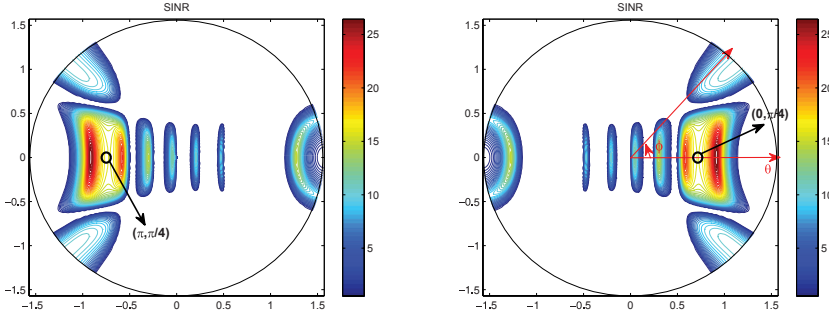


Fig. 8. The contour plots of SINR for user 1 and 2 using the RF BF. The left plot corresponds to user 2 and the right plot to user 1, with the desired directions of user one and two set at $(\phi_1, \theta_1) = (0, \pi/4)$ and $(\phi_2, \theta_2) = (\pi, \pi/4)$, respectively.

Fig. 8 presents the contour plots of the SINRs of user one and two when using the RF BF method of (4). The contour plots are shown in the cylindrical coordinate. The radial coordinate maps to the elevation angle θ and the angular coordinate maps to the azimuth angle ϕ . The desired directions of user one and two are set at $(\phi_1, \theta_1) = (0, \pi/4)$ and $(\phi_2, \theta_2) = (\pi, \pi/4)$, respectively. The resultant SINR in the desired direction of each user is equal to 16.18dB.

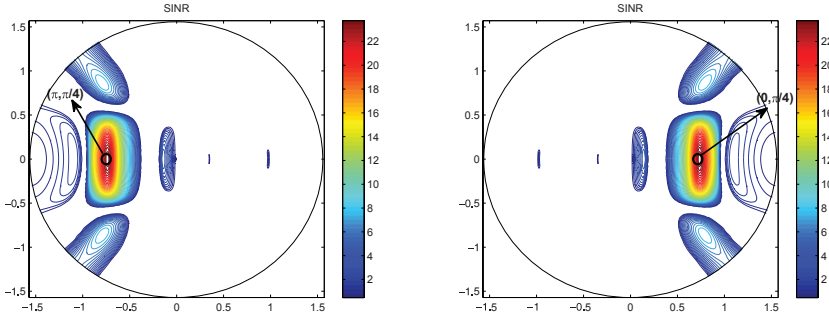


Fig. 9. The contour plots of SINR for user 1 and 2 using the HBF of LCMP. The left plot corresponds to user 2 and the right plot to user 1, with the desired directions of user one and two set at $(\phi_1, \theta_1) = (0, \pi/4)$ and $(\phi_2, \theta_2) = (\pi, \pi/4)$, respectively.

With the same simulation setting for RF BF, the contour plots of SINRs with HBF of LCMP are shown in Fig. 9. In comparison with the results in Fig. 8, we can see that the main beam here for each user is much more stronger and narrower, while the side beams are relative fewer and weaker. In addition, the resultant SINR in the desired direction of each user is now increased

to 24.09dB in Fig. 9 as oppose to the 16.18dB in Fig. 8 for the RF BF. This demonstrates that the interference can be suppressed effectively in the desired directions of users with the HBF of LCMP method. Besides, the results of HBF with MD BF are also similar to those of LCMP and, hence, are not shown here.

In addition to SINR, directivity is also an important performance measure to characterize the effectiveness of BF. To reflect the interference due to the multiple access in SDMA, the original definition for directivity in Balanis (1997) is modified into

$$D_i = \frac{4\pi \text{SINR}_i(\phi_i, \theta_i)}{\int_0^{2\pi} \int_0^{\frac{\pi}{2}} \text{SINR}_i(\phi, \theta) \sin(\theta) d\theta d\phi}. \quad (29)$$

This new definition for directivity automatically refers to the traditional notion of directivity in the single-user system.

According to this definition of directivity, Table 1 summarizes the results of SINRs, directivities and radiation powers for the RF BF and HBF schemes of MD and LCMP, respectively, when the desired directions of users are set at $(\phi_1, \theta_1) = (0, \pi/4)$ and $(\phi_2, \theta_2) = (\pi, \pi/4)$. The radiation power for each user is evaluated with

$$\int_0^{2\pi} \int_0^{\frac{\pi}{2}} |H_i(\phi, \theta)|^2 \sin(\theta) d\theta d\phi \quad (30)$$

and is denoted by P_T in the table.

It is clear from the table that HBF not only achieves higher SINRs and directivities in this case, but also uses less power for BF. This demonstrates its great potential for SDMA in mmWave applications.

User1 User2 (ϕ, θ)	User1 (0, $\pi/4$)			User2 ($\pi, \pi/4$)		
	RF	MD	LCMP	RF	MD	LCMP
SINR (dB)	16.18	24.09	24.09	16.18	24.09	24.09
Directivity	13.33	156.3	156.3	13.55	159.4	159.4
P_T	0.1388	0.0775	0.0775	0.1358	0.0764	0.0764

Table 1. The SINRs, directivities and the radiation power of the RF BF and the HBF schemes of MD and LCMP.

In addition to typical measures for the evaluation of a BF scheme, we also investigate the effectiveness of HBF from the perspective of wireless communications systems. To this end, we study the performance of a two-user SDMA with PAA in orthogonal frequency division multiplexing (OFDM) wireless communications systems. The OFDM system simulates the physical layer proposal of IEEE802.11 task group AD for very high throughput in 60GHz Channel Models for 60 GHz WLAN Systems (n.d.). Specifically, we consider an indoor simulation setting as shown in Fig. 10 for a two-user SDMA system operating in the 60GHz wireless channel model proposed for IEEE802.11ad Channel Models for 60 GHz WLAN Systems (n.d.). The system bandwidth is assumed to be 1GHz and the length of fast fourier transform (FFT) is 1024.

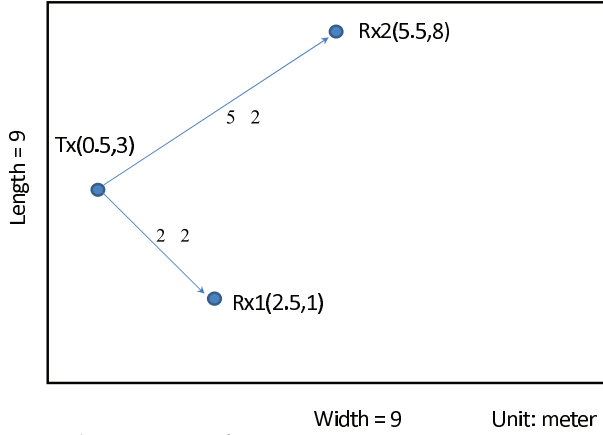


Fig. 10. The indoor simulation setting for a two-user SDMA system.

The simulated transmitter is placed at the position of (0.5, 3) from the lower left corner of the room and transmits signals simultaneously to the two users locating at (2.5, 1) and (5.5, 8), respectively, using either the RF BF setting shown in Fig. 8 or the HBF setting of LCMP in Fig. 9. The users' data are first modulated with QPSK and then transformed with the 1024-point inverse fast fourier transform (IFFT). The transformed OFDM symbol is added with a cyclic prefix of 1/8 of the OFDM symbol and then transmitted with the corresponding BF techniques to the intended users simultaneously. The transmitted signals propagate through the 60GHz channel modeled according to Channel Models for 60 GHz WLAN Systems (n.d.) to come to the two receivers of Fig. 10. At the receivers, we assume that the same RF BF with a 8×4 PAA is used by each receiver to enhance the signal qualities, supposing that the beam pattern of each user is pointed to its desired angle of arrival.

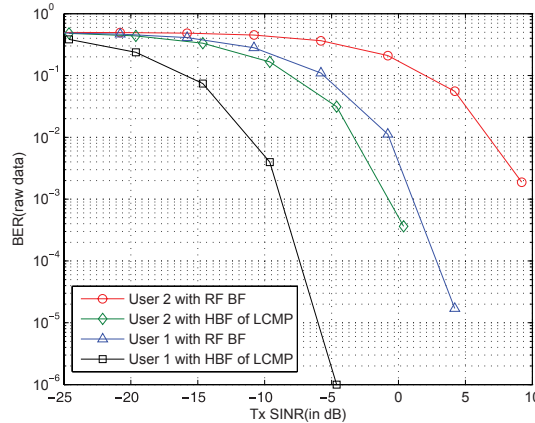


Fig. 11. The BERs for a two-user SDMA system in an indoor environment of Fig. 10.

To recover the data, the receivers first remove the cyclic prefixes of each and then perform FFT followed by channel equalization for symbol detection. The bit error rate (BER) for each user is shown in Fig. 11 with respect to the transmitted SINR. As can be seen from the figure, the SNR advantages for users using the HBF of LCMP are around 10dB against the users using the RF BF only, even though the HBF gain of LCMP is only 7.91dB higher than that of RF BF.

5. Conclusions

We presented HBF techniques for SDMA in 60GHz radio using reconfigurable PAA. According to our simulation studies, the BF gain of HBF can be as much as 7.9dB higher than that of RF BF. Furthermore, the BER of using HBF in a simulated OFDM environment can have an even higher SNR advantage of 10dB against that of using the RF BF only. These results demonstrate the potential of HBF with PAA for SDMA in mmWave applications.

6. References

- Alliance, W. (n.d.). <http://www.wimedia.org/>.
- Anderson, C. R. & Rappaport, T. S. (2004). In-building wideband partition loss measurements at 2.5 and 60 GHz, *IEEE Trans. on Communications* 3(3): 922–928.
- Balanis, C. A. (1997). *Antenna Theory*, 2 edn, John Wiley & Sons.
- Balanis, C. A. & Ioannides, P. I. (2007). *Introduction to Smart Antennas*, Morgan and Claypool.
- Celik, N., Kim, W., Demirkol, M., Iskander, M. & Emrick, R. (2006). Implementation and experimental verification of hybrid smart-antenna beamforming algorithm, *IEEE Antennas and Wireless Propagation Letters* 5: 280–283.
- Channel Models for 60 GHz WLAN Systems (n.d.). 11-09-0334-02-00ad-channel-models-for-60-ghz-wlan-systems.doc. available at http://www.ieee802.org/11/Reports/tgad_update.htm.
- FCC (2004). Code of federal regulation, title 47 telecommunication, chapter 1, part 15.255.
- IEEE 802.11 WLAN Very High Throughput in 60GHz (n.d.). available at <http://www.ieee802.org/11/>.
- IEEE 802.15 WPAN Millimeter Wave Alternative PHY Task Group (TG3c) (n.d.). available at <http://www.ieee802.org/15/pub/TG3c.html>.
- Kuhwald, T. & Boche, H. (1999). A constrained beam forming algorithm for 2D planar antenna arrays, *Proc. IEEE VTC-Fall*, Amsterdam, The Netherlands.
- Rezk, M., Kim, W., Yun, Z. & Iskander, M. (2005). Performance comparison of a novel hybrid smart antenna system versus the fully adaptive and switched beam antenna arrays, *IEEE Antennas and Wireless Propagation Letters* 4: 285–288.
- Smolders, A. B. & Kant, G. W. (2000). THousand Element Array (THEA), *Proc. IEEE Antennas and Propagation Society International Symposium*, Salt Lake City, UT.
- Smulders, P. (1995). Broadband wireless LANs: a feasibility study, *Ph. D. thesis*, Eindhoven, Univ. of Tech., The Netherlands, ISBN 90-386-0100-X. available at <http://alexandria.tue.nl/extra3/proefschrift/PRF11B/9505571.pdf>.
- Smulders, P. (2002). Exploiting the 60 GHz band for local wireless multimedia access: prospects and future directions, *IEEE Communications Magazine* 2(1): 140–147.
- Trees, H. L. V. (2002). *Optimum array processing. Part. IV of detection, estimation and modulation theory.*, John Wiley & Sons.
- WirelessHD (n.d.). <http://www.wirelesshd.org/index.html>.
- Zhang, Z., Iskander, M., Yun, Z. & Host-Madsen, A. (2003). Hybrid smart antenna system using directional elements - performance analysis in flat Rayleigh fading, *IEEE Trans. on Antennas and Propagation* 51(10): 2926–2935.

A Distributed Multilayer Software Architecture for MIMO Testbeds

José A. García-Naya, M. González-López and L. Castedo
Universidade da Coruña
Spain

1. Introduction

The use of multiple antennas at both transmission and reception, also known as Multiple Input Multiple Output (MIMO) transmission systems, has received a lot of interest from the wireless communications industry during the last years. Communications in wireless channels using MIMO technologies exhibits a superior performance in terms of spectral efficiency, reliability and data rate when compared to conventional single antenna technologies (Foschini & Gans, 1998; Telatar, 1999). Existing and emerging standards for wireless communications such as IEEE 802.11 (WiFi), IEEE 802.16 (WiMAX) and Long Term Evolution (LTE), support multi-antenna transmission in their highest performance profiles.

In spite of their potential performance-enhancing capabilities, most of the research on MIMO technologies up to the moment is based on theoretical studies. Typically, the expected gains of MIMO technologies are only shown under ideal conditions since most analysis rely on simulations. Experiments in real-world scenarios by means of hardware implementations are necessary to measure the actual performance of multi-antenna transmission methods. Hardware implementations not only take into account the real multipath propagation in wireless channels but also the implementation impairments so often ignored during the simulations. Hardware implementations can be split into three groups (Rupp et al., 2006). The first one is constituted by *demonstrators* which are frequently designed having in mind a particular standard or specification. Demonstrators usually exhibit good technical features for real-time implementations but they are extremely expensive and present poor flexibility and modularity. The second group is formed by *prototypes* of a final product. A prototype is a real-time implementation of a system specifically developed to support an industrial need. Prototypes often constitute a preliminary stage where the system is implemented and debugged and later on implemented as a consumer product. Finally, the third set is formed by *testbeds* that support real-time transmission capabilities while data is generated and post-processed off-line.

In addition, hybrid solutions can be devised. As an example, a testbed can carry out some operations in real-time with the purpose of speeding up the measurement process. Usually, candidate signal processing operations to be implemented in real-time are those that operate at sample level and/or do common tasks for all experiments, i.e. I/Q modulation, up-sampling, pulse-shaped filtering, etc. Throughout this chapter we will focus on testbeds because they use open designs and are more often found in public research centres and academia.

Various MIMO testbeds have been reported in the literature (Borkowski et al., 2006; Caban et al., 2006; Fabregas et al., 2006; Haustein et al., 2006; Nieto et al., 2006; Ramírez et al., 2008;

Rao et al., 2004; Wilzeck et al., 2006; Zhu & Fitz, 2005). Some of them have been constructed to evaluate a particular standard or specification while others have been designed for general purpose. Flexibility, development time consumption, throughput or costs are important features when comparing existing testbeds. Also, it should be noticed the educational possibilities of testbeds that open the door to many teaching opportunities. In the literature, however, there is a lack of up-to-date guides and tutorials useful for the construction of a new testbed from the scratch. Indeed, there exists few contributions (Caban et al., 2006; García-Naya et al., 2008a; Rao et al., 2004; Rupp et al., 2007) that contain a detailed description of the constructed MIMO testbed and, except for (García-Naya et al., 2008a; Rupp et al., 2007), the information contained on them is already outdated.

Based on the previous experience acquired by the research group of the authors in building and setting up MIMO testbeds, as well as performing indoor measurements (Pérez-Iglesias et al., 2008; Ramírez et al., 2008), we can assert that once the testbed hardware is available and properly configured, accessing the testbed becomes the main and also frequently ignored issue. When a research team decides to start the process of acquiring and/or constructing a new testbed, they need to take into account numerous aspects related both with the hardware and its technical features, and the extensibility possibilities for the future (García-Naya et al., 2008a; Rupp et al., 2007). Usually, most of the efforts are devoted to the testbed setup, which results in equipment that is hardly usable by people not involved in its design and later configuration. This makes extremely difficult to access to the testbed.

As a result, the migration of an algorithm from a simulation environment to a testbed involves cumbersome low-level programming to access the hardware as well as a very detailed knowledge of the hardware. Additionally, hardware implementation problems frequently ignored by simulations arise, such as time and frequency synchronization, I/Q imbalances, non linear distortions caused by the power amplifiers, etc. All these issues make difficult to assess new MIMO transmission methods in a testbed. For this reason, it is desirable to make the testbed accessible to final users at a reasonable abstraction level. This goal represents an important challenge due to the large amount of heterogeneous technologies and development environments that have to be integrated together. However, if this challenge is accomplished, the final result is a very attractive product for the user, who can focus on the development of new transmission techniques that can be easily translated to the testbed and later evaluated in realistic scenarios.

In this chapter we describe how to solve all previously mentioned limitations by using a distributed multilayer software architecture. This architecture enables the testbed to be easily accessible by the researchers and to be integrated in the development environment they are using. Although all designs and results herein presented are particularized for our MIMO testbed (García-Naya et al., 2008b; Ramírez et al., 2008), they are easily adaptable to most of the existing testbeds, making even possible the integration of heterogeneous testbed nodes in order to build a multi-terminal testbed.

The proposed software architecture consists of three different layers:

1. **The middleware layer (MWL)** is the lowest-level layer that interacts with the testbed hardware. It makes the testbed accessible through standard TCP socket connections.
2. **The signal processing layer (SPL)** performs the necessary operations required to convert the discrete-time sequences provided by the final user into discrete-time signals suitable to be transmitted by the hardware. At the receiver, it is usual to perform signal processing operations like time and frequency synchronization in case they are not carried out by the testbed hardware. The tasks compounding the signal processing layer

can also be executed in real-time by the testbed hardware. For example, digital up and down converters are frequently available in the libraries of programmable hardware modules.

3. **The testbed interface layer (TIL)** is the highest-level layer and presents the testbed to the user at an adequate abstraction level. The TIL has to be designed and implemented for a specific development environment. For example, if the final user makes use of Matlab, then a specific implementation of the TIL for Matlab has to be developed. The main purpose of the TIL is to provide a simplified interface to access the testbed. This does not prevent from providing mechanisms to control the hardware in detail from the TIL. It is very important to emphasize that there is no logic in the TIL except that necessary for adapting the data format from the specific environment used by the SPL and vice versa. The only requirement for the TIL to be implemented is the availability of a standard TCP socket library.

The whole design and implementation of the proposed software architecture is done under the following premises:

1. **Layers have to be as decoupled as possible.** This is a fundamental idea in modern software design and structured programming. The key idea is not to replicate functionalities that are already present in another layer, which would be a symptom of a bad design. The basic premise is to design everything to be fully decoupled and reusable; and later on introduce some small violations of this principle only if they are strictly needed to increase the overall system performance.
2. **Each layer can be extended and/or customized** to be able to adapt them to future specifications and/or heterogeneous hardware environments. It is important that this layer upgrade be done without needing a complete remake of the software, which is frequently the only solution in case of monolithic non-layered systems.
3. Finally, **layers should be distributed**, which means that they use remote connections to interact among them. On the one hand, this helps to ensure the decoupling and independence principles whereas, on the other hand, allows the testbed to be remotely accessible from the user Personal Computer (PC).

The remaining of this chapter is structured as follows. Section 2 and Section 3 provide a global description of the testbed hardware and software possibilities, respectively. Section 4 describes the software technologies that were used to develop the proposed distributed multilayer software architecture. In Section 5 the software architecture is presented and described. The interaction among the different layers and components of the architecture is studied in Section 6 and Section 7. Finally, Section 8 is devoted to the conclusions.

2. Basic Testbed Hardware

Testbeds are often used to verify if a new signalling technique (or even a complete standard system) that has been proven useful by simulations is also valid in realistic wireless scenarios. Testbeds allow the dimensioning of the hardware needs for real-time implementations, not only for the digital signal processing modules (mainly, DSP and FPGA) but also for the analogue Radio Frequency (RF) front-ends. Testbeds allow capturing the requisites for the hardware used in the final real-time implementation. Consequently, evaluating performance with testbeds allows knowing whether a given technique is feasible from the real-time hardware and implementation requirements.

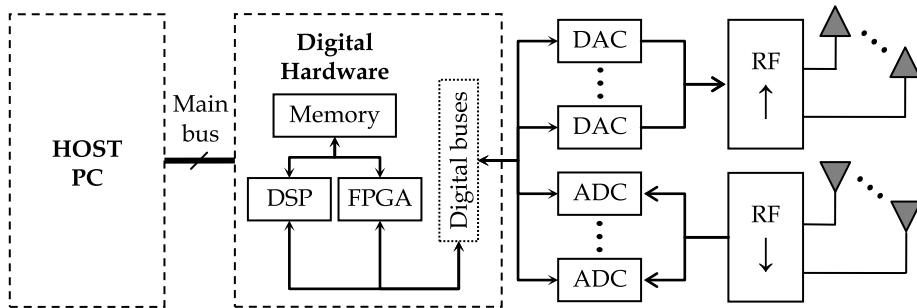


Fig. 1. Block diagram of the basic hardware configuration of a testbed.

Among the three different types of hardware implementations described above (demonstrators, prototypes and testbeds), testbeds present the following advantages:

- **Flexibility.** Testbed hardware is meant to be used for off-line processing. Only the signals are sent and acquired in real-time. This implies that testbed hardware is not subject to the real-time restrictions even though the hardware can include some sort of real-time capabilities.
- **Modularity.** Usually, the minimum modularity found in a testbed is given by the separation between the digital hardware (up to the D/A and A/D converters) and the RF hardware. Sometimes, the digital hardware can be split in different modules performing specific operations: D/A and A/D conversion, digital up and down conversion, signal buffering, etc. For the RF section it is possible to find self-made solutions or commercial products. Lastly, commercial RF front-ends also permit some degree of flexibility, for example allowing dual-band operation, RF carrier selection for each band or adjusting the gains at the transmitter and receiver amplifiers.
- **High-level language development.** Having in mind that most of the processing tasks are carried out off-line, general-purpose processors are the most adequate for processing the generated/acquired signals. This allows the utilization of high level programming environments (e.g. Matlab, C/C++, Java) and the simplification of the implementation stages both in time and complexity. This feature also provides an additional flexibility degree, because it is easy to change the implementation on the fly.
- **Floating-point versus fixed-point precision.** As a consequence of the off-line processing and the usage of high-level programming environments, the operations are carried out in the floating-point domain instead of using fixed-point operations available for real-time devices. This permits the researcher to focus on the implementation rather than considering some other problems like arithmetic precision of the operations.

Testbed hardware components can be classified into three groups according to their functionality, as shown in Fig. 1. The first one is the host system, usually a PC and consequently referred to as host PC. It is the equipment that allocates one or more boards containing the digital section of the hardware testbed (including D/A and A/D converters). The second group is constituted by the digital hardware components and, finally, the third one is formed by the RF front-ends. The D/A and A/D converters generally constitute the frontier between

the digital and analogue hardware. In the digital section, a bus termed main bus allows transferring the data from the host to the testbed hardware and the other way around. Next, a set or just one digital bus interconnects the digital hardware (DSPs, FPGA, memory buffers, etc.) with the D/A and A/D. Finally, for the interconnection among the D/A and A/D converters and the RF front-ends coaxial cables are used.

With this basic configuration, it is possible to send samples directly coming from the main bus, convert them into the analog domain using the D/A converters and up convert them to the desired carrier RF using the front-ends. At the receiver side, the signals are down converted by the RF front-ends, digitally converted by the A/D converters and then sent to the host through the main bus. Note that the main bottleneck in this scheme is the maximum data rate provided by the main bus, especially if there are no digital up and down converters available. In the most recent boards, by using PCI express or similar solutions is possible to use the host memory as a buffer while samples are transferred in real-time through the main bus.

A first improvement of this basic scheme consists in incorporating a digital up converter (DUC) before each D/A converter at the transmitter and a digital down converter (DDC) after each A/D converter at the receiver. These devices can be dedicated elements or can be implemented in a FPGA. Incorporating DUC and DDC into the MIMO testbed design allows transferring complex signals from/to the host, reducing both the transfer rate at the main bus and the software complexity. For MIMO operation, DDCs and DUCs must be fully synchronized. Additionally, another improvement consists in implementing some of the time-consuming sample-level tasks in FPGAs (e.g. time and frequency synchronization).

2.1 Baseband Components

Baseband components are the hardware elements necessary to deal with baseband and/or Intermediate Frequency (IF) signals. Frequently, such baseband components are allocated on carrier boards installed on conventional PCs. Current testbeds use carrier boards equipped with standardized buses to access testbed hardware such as USB, PCI, cPCI, PCI express or similar. A manufacturer compliant with the PXI alliance (PXI, 2009) produces carrier elements that are compatible with the most typical buses present in host equipments. When available, important hardware components in a testbed are the storage buffers, especially when the main bus does not support the necessary rate demanded by the D/A and A/D converters. Such buffers allow performing signal operations off-line while data is sent and acquired in real-time.

The interconnection among the previously described elements is carried out using buses that must be capable of transmitting the data fast enough. Finally, external circuitry such as clock distribution and/or triggering is needed. Sometimes, the most advanced hardware manufacturers include such circuitry as part of the commercial boards, simplifying the later setup. When MIMO processing is required, fully synchronization among the different devices is mandatory. Sometimes manufacturers announce a MIMO system but just a scaled SISO solution is offered.

2.2 RF Front-Ends

RF front-ends constitute one of the major hindrances in the testbed building process. They are responsible of up converting IF or baseband signals to RF. The most common RF bands are the unlicensed ISM located at 2.4 and 5.8 GHz. It is easier to find components for such bands but also unpredictable interference is present in such spectrum portions. High linearity and flexibility in terms of supported carried frequencies and bandwidths are desirable features for

the RF front-ends. However, the most advanced front-ends commercially available are only dual-band and have fixed maximum bandwidth. A fundamental feature needed to be able to carry out experiments is the possibility of modifying the transmit power. Also, the majority of the front-ends are designed for SISO operation, making difficult to adapt them to a MIMO system. Moreover, once the front-end has been acquired, extensive test on it is needed and expensive measurement equipment is required.

It is important to emphasize that RF hardware is expensive compared to the cost of the digital hardware and does not follow Moore's law (Moore, 1998).

3. Basic Testbed Software

Methodologies that cover the entire development process, from source code suitable for simulations and executed off-line, to real-time implementation in the testbed, as well as software tools according to these methodologies, are extremely scarce. The most popular methodology for testbed development is rapid prototyping (Kaiser et al., 2004; Rupp et al., 2003; 2006). A typical approach used to develop real-time algorithms to be run in a testbed consists in starting with a simulation implementation. Next, the simulation is migrated to the testbed and, finally, a real-time implementation is obtained. It is also convenient to split the real-time implementation into several steps: firstly, a fixed-point code is produced; next a DSP implementation is obtained; and, afterwards, the software modules that do not meet time requirements are migrated to FPGA, making use of high level tools when possible. Nowadays, except for the Mathworks tool suite (Mathworks, 2009) combined with some VHDL code generators, there is a lack of high level tools and development environments suitable to fit the previous steps.

Also, some standardisation efforts have just started in order to make compatible hardware components and software modules from different manufactures and developers. The PXI alliance (PXI, 2009) plus the Software Defined Radio (SDR) Forum (SDR Forum, 2009) and initiatives such as the Software Communications Architecture (SCA) (SCA, 2009) constitute the starting point of a new generation of modular, flexible and standard radio interfaces suitable for research. Nowadays, however, the previously mentioned initiatives have not already produced as a result the availability of high level tools allowing final users and researchers not involved in the hardware development to access the testbed hardware at a reasonable abstraction level.

All above reasons serve as a motivation for our work, which aims at bridging the existing gap among the hardware, the software elements provided by the manufacturers and the abstraction level required by the final users. Moreover, the proposed solution allows embedding real-time modules in the processing chain, as if they were part of the testbed hardware. For example, real-time frequency and time synchronization algorithms can be run at the receiver side and integrated into the digital hardware. Additionally, given that the proposed solution is designed to be easily and quickly integrated in any other kind of system, it becomes a very useful tool to help in testing real-time applications, especially during the first stages of the development. A good example is the ability to develop in parallel different parts of the transmitter and the receiver in real-time. While the real-time transmitter is still under development, the team developing the real-time receiver can make use of the testbed to generate the transmit signals and feed the real-time receiver with them.

4. Software Technology Behind the Distributed Multilayer Architecture

The field of computer science has come across problems associated with complexity since its constitution. The software architecture discipline is centred on the idea of reducing complexity through abstraction and separation. There are different kinds of software architectures. Among them, the most interesting for our work are: the client-server architecture, which serves as the basis for most of the available architectures; the three-tier model; and finally, the most recent and advanced model-view-controller architecture.

4.1 Client-Server Software Architecture (Two-Tier Software Architecture)

Client-server describes the relationship between two computer processes in which one process (the *client*) makes a service request to another process (the *server*). Applications following the client-server architecture represent an evolution with respect to those called "monolithic".

The basic operation mode for client-server applications consists in that each instance of the client process sends requests to one or more connected servers. In turn, servers accept these requests, process them, and return the requested information to the client. The basic client-server architecture considers only two types of hosts: clients and servers. Consequently, it is also known as two-tier model. A special case of the two-tier software architecture arises when an instance simultaneously acts as a client or as a server, resulting in the peer-to-peer architecture.

The two-tier software architecture presents the following advantages with respect to monolithic systems:

- The responsibilities of the whole system are now split between the client and the server, which brings the opportunity to decouple them.
- Servers can control the access to the resources to guarantee that only those clients with appropriate permissions access and change the data.
- Different client/server types can work with different kinds of servers/clients, which helps in the integration of heterogeneous systems.

As a main disadvantage, the mechanism used to interconnect clients and servers (frequently standard socket connections) may become both a bottleneck and a weak point of the system, because a failure in the interconnection mechanism will stop the whole system.

4.2 Multi-Tier Software Architecture

The multi-tier software architecture represents a natural evolution of the client-server model towards a higher number of levels. That is the reason why it is also known as the *n-tier software architecture*. Basically, the multi-tier software architecture is a client-server architecture consisting of more than two tiers, and where the inner tiers act simultaneously as client for one tier and as server for the other. The main difference with respect to the peer-to-peer model is that the tier order (its position or level in the multi-tier structure) does matter. An inner layer is also termed middleware because acts as an intermediary between two adjacent tiers.

Note that the multi-tier software architecture, as well as the two-tier one, is a mechanism to design the physical structure of the system. Given one of the models, several degrees of freedom are still possible to get the logic structure of the system.

4.3 Tier Interconnection Mechanisms

Up to now, we were using a certain abstraction level to define the previous architectures. Usually, each tier is mapped to a set of processes obtained from the same master process. This

can be easily understood with an example. Imagine a two-tier system. One tier is the client while the other is the server and they are mapped into two different processes interconnected by means of sockets. It does not matter if both processes are in the same physical machine or not. What is important is how they interact with each other. The server will be waiting for a request from one of the clients. As soon as the request arrives, the server starts to serve the petition and, simultaneously, starts waiting for another request. When the server process is going to serve a new request, it clones itself to serve the petition. Frequently, this process cloning mechanism is omitted and it is considered as an implementation detail. Therefore, to describe the interaction between different tiers, it is assumed that each tier is mapped into a process that interchanges messages with other processes (tiers). Although this approach is not completely rigorous, it is enough for our purposes. In the literature such interactions are described by means of sequence diagrams as we will do.

But there is still one open question: How are messages sent from one tier to another? In principle, data transfer between two tiers is part of the architecture design, but in practice only the message types and the data are defined by using sequence diagrams, which drives to the definition of a part of the system termed *protocol*. A protocol defines the interaction among all tiers, specifying the type of messages to be sent from one tier to another or to several other tiers, as well as the type of data sent in each message. Messages are transferred using plain socket connections or more elaborated mechanisms (e.g. SNMP, CORBA, Java RMI or even Web Services). The different approaches have their own advantages and disadvantages and there is still no perfect system for message interchange. However, what is really important is to use standardized mechanisms for message interchanging, thus guaranteeing independence among tiers.

4.4 The Model-View-Controller Architectural Pattern

The Model-View-Controller (MVC) is a pattern used in software engineering derived from the multi-tier software architecture. Successful use of the pattern drives to the isolation of the business logic from the user interface, allowing one to be freely modified without affecting the other. The controller collects user inputs, the model manipulates (and usually stores) application data, and the view presents results to the user. The MVC was first described in (Trygve, 1978) and can be used as an architectural or design pattern. As an architectural pattern, it splits an application into three independent layers that can be run in different computers: the presentation or user interface layer, termed the view; the business logic, called the model; and the controller, an intermediate layer adapting the view to the model and vice versa.

4.5 Applying Software Engineering to MIMO Testbeds

In the first sections of this chapter we present the typical hardware architecture available in MIMO testbeds as well as the typical abstraction level offered by the software included with the hardware. We stressed that such abstraction level is by far not enough and, consequently, a final user or a researcher not involved in the testbed development and later setup cannot directly access it. Along this Section, we introduced architectural software models, starting with the well-known client-server model and ending with the recently proposed model-view-controller, widely used in web applications as well as distributed applications. These architectural models serve as an inspiration to propose a novel software architecture useful to bridge the existing gap between the abstraction level offered by testbeds and the one demanded by researchers and final users. This new architecture is explained in the following Section.

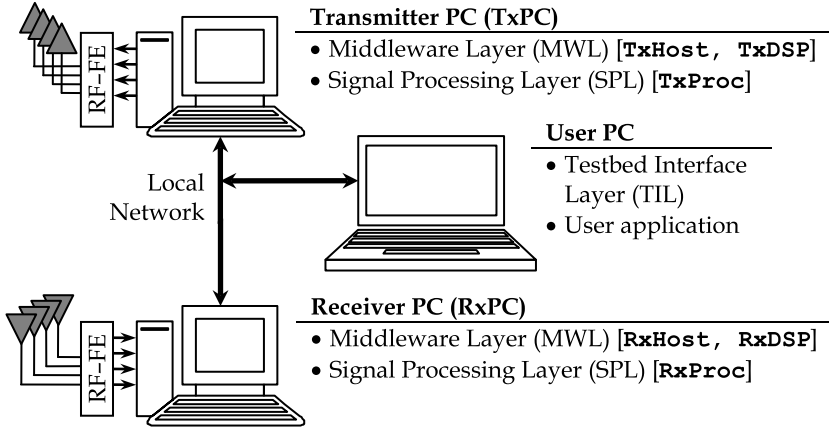


Fig. 2. General scheme of the MIMO Testbed showing the three different layers: middleware (MWL), signal processing (SPL) and testbed interface (TIL). The corresponding process name for each layer and each PC is shown between brackets.

5. Distributed Multilayer Software Architecture for MIMO Testbeds

Fig. 2 shows the proposed testbed hardware organization, which can be extrapolated to most of the testbeds. It is constituted by two ordinary PCs hosting the testbed hardware, one for the transmitter (referred to as TxPC) and other for the receiver (named RxPC). The digital section of the testbed hardware is installed inside the PCs while the RF front-ends (RF-FE) remain outside the PC case. The two PCs are attached to the network, something that is very useful because this way the PC desktops can be remotely accessed. The remote user PC is also attached to the network. In the rest of the Section we are assuming the following:

1. The testbed consists of the transmitter and the receiver, i.e., two nodes. Although our designs can be easily extended to an arbitrary number of nodes, it is better to start with the simplest case of two nodes to keep things simple.
2. It does not matter whether the testbed operates outdoor, indoor, outdoor to indoor or vice versa. We assume that a network connection can always be established among the testbed nodes and the user PCs.
3. The testbed PCs will use a standard operating system supporting remote desktop or remote operation from a PC attached to the network. This is a fundamental feature because it enormously simplifies software deployment as well as day-to-day maintenance of the systems. Otherwise, a replacement for such tools should be provided.

5.1 From the MIMO Testbed to the Multilayer Software Architecture

Fig. 2 shows, close to the PC drawings, the names of the corresponding three layers of the proposed architecture that are, from the lowest to the highest level: the middleware layer (MWL), the signal processing layer (SPL) and the testbed interface layer (TIL). The MWL and the SPL are split into transmit and receive parts, while the TIL has just one instance running on the user PC. Actually, the TIL is just an Application Program Interface (API) that has to be

included with the user application. In our standard deployment, the MWL and the SPL are allocated in the same PCs containing the rest of the testbed hardware. Although the MWL is the only layer required to be installed in the hardware PCs, the rest of the software can be deployed in any machine attached to the network.

This is not an arbitrary proposal but a design inspired on the multi-tier and the model-view-controller software architectures. First of all, let us identify the use cases of our application. Basically, there is just one use case or action carried out by the final user: transmitting a set of discrete-time sequences through the testbed and, consequently, through the wireless channel. Therefore, the layer corresponding to the view is the interface that allows accessing the testbed, sending the sequences and getting back the acquired signals. This is what we term testbed interface layer, and is the topmost layer of our software architecture.

The controller plays a very important role in a system. It is responsible of getting the service requests from the view, adapting and sending them to the model where the business logic resides. The gathered results are then sent back to the view to be presented to the user. The controller plays the role of the middleware concept presented in the multi-tier software architecture. In the testbed system the controller is called middleware and its functionality is clear: to configure and control the hardware in order to be able to take the requests (discrete-time sequences) from the TIL; to adapt and send them to the hardware to be transmitted by the antennas; and, finally, to carry out the reciprocal operations at the receiver side. At the end of the process the TIL returns the acquired discrete-time signals.

However, our architecture presents three strong differences with respect to the model-view-controller and the multi-tier architecture:

- Adaptation of the discrete-time sequences provided by the TIL in the requests, both at the transmitter and the receiver, consists in performing different signal processing operations that sometimes can even be executed using different kind of processors. For instance, a general-purpose processor (GPP), a graphic processor unit (GPU) or real-time devices like an FPGA or a DSP. It is thus necessary to put all these operations together and, consequently, the SPL concept comes out. Therefore, the SPL is in charge of carrying out most of the signal processing operations needed between the MWL and the TIL.
- One of the reasons to use a multi-tier architecture jointly with the model-view-controller is the ability of the resulting application to be distributed among different machines. Thus, different instances of the tiers run on different machines. However, in the testbed system there are two kinds of nodes: the transmitter and the receiver. If a multi-node testbed is available, different transmitter and receiver pairs will be distributed among different nodes. As a result, the MWL (and consequently the SPL) is split into two sides: the transmitter side and the receiver side. They will be referred to as Tx MWL, Rx MWL, Tx SPL and Rx SPL, respectively. Additionally, they are also identified because the processes mapping the architecture design are also split into two sets, one for the transmitter and the other for the receiver.
- Finally, another particularity present when adapting the model-view-controller and the multi-tier software architectures to testbed software architecture is that the hardware is required to be attached to the host system by means of a bus (e.g. PCI, PCIX, PCI Express, etc.). Consequently, it is not possible to sustain standard network connections through such buses, and the standard tools and techniques available for interconnecting layers are not applicable (except when a custom implementation is provided or when

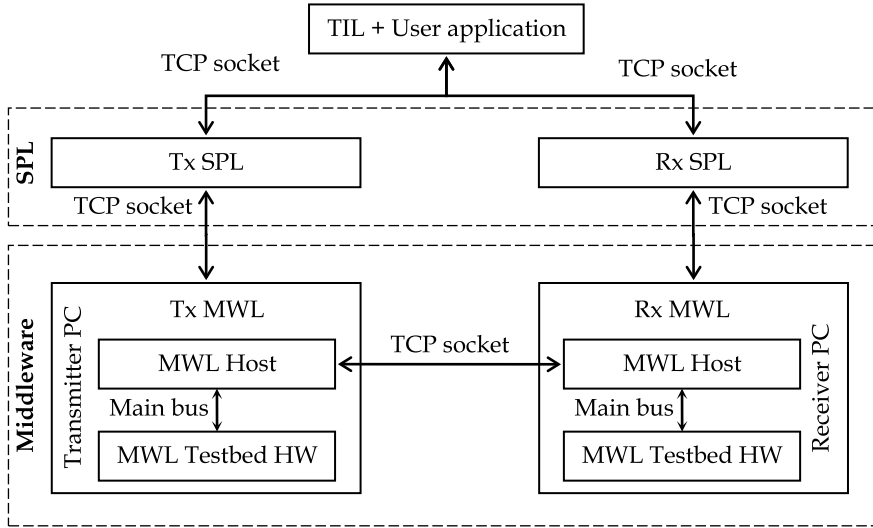


Fig. 3. Basic structure of the distributed multilayer software architecture for MIMO testbeds. The three layers are shown: MWL, SPL and TIL. Additionally, the testbed-hardware sub-layer in the middleware and the different interconnection mechanisms are included.

the hardware is attached through a network connection). These reasons motivate another division in the MWL, generating a *Testbed-hardware* sub-layer (with its respective parts for the transmitter and the receiver sides) responsible of dealing with hardware aspects as well as solving the bus connection issues with the host part of the MWL. Having in mind that in our testbed this sub-layer runs on the DSPs available at the transmitter and the receiver, the corresponding processes are referred to as TxDSP and RxDSP.

5.2 Logical and Physical Designs of the Distributed Software Architecture

In Fig. 3, the basic structure of the proposed architecture is shown. There are two sides, transmitter and receiver, joined at the TIL, which has to hold connections with both sides of the SPL. All links between the different layer elements are implemented by using standard socket connections. The exceptions are the links between the sub-layers of the MWL that use a proprietary protocol over the existing main bus interconnecting the hardware and the host. There are strong reasons for using standard socket connections instead of any other higher-level mechanism like, for example, web services. However, this does not imply that in some situations other connection types can better solve some problems. It is also possible to use different link types among different layers. For example, the TIL and the SPL can be connected using web services, thus allowing total independence of the platform and full remote access. However, the ability of such high level techniques to sustain high data rates and low latency connections is not the best. For these reasons, sockets offer enough flexibility while providing fast connections. In the case of bus connections the bus type obviously limits the latency and the data rates.

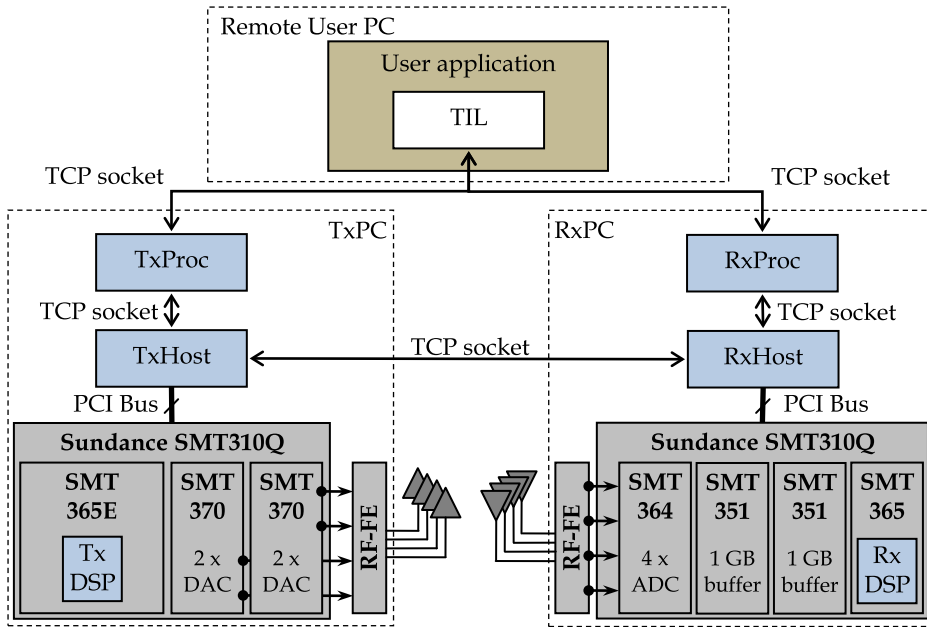


Fig. 4. Testbed scheme containing the hardware and the software architecture deployment as well as the links among the components. The different processes are shown in blue, and are located in the usual place for a typical architecture deployment. The digital hardware sections of the testbed, as well as the RF front-ends, are shown in gray. Finally, the user application is in dark green, containing the TIL in white. Note that TIL appears in white and not in blue because it is not a process but an user application.

While in Fig. 2 the general structure of the testbed is depicted, showing the correspondence between the hardware elements and the software layers, Fig. 4 illustrates the block diagram of the entire system. Three main parts can be distinguished: the testbed hardware that allows us to transmit discrete-time signals over multiple antennas; the multilayer software architecture that makes the hardware accessible to end researchers at a high abstraction level; and, finally, the user application implemented using the testbed capabilities and the architecture facilities. The lowest software level (i.e. the MWL) is required to be installed in the same PCs as the testbed hardware because it uses the system buses to communicate with the hardware. Otherwise, this restriction would not be applicable. The other two layers can be installed in any other available PC. However, in the standard deployment, the SPL is included with the MWL in the testbed PCs. Finally, the TIL is installed in the remote user PC.

5.3 Short Description of the MIMO Testbed Hardware

At the bottom of Fig. 4 a very basic diagram of our testbed hardware is shown. A first release of the testbed hardware was presented in (Ramírez et al., 2008), where the baseband modules were from Sundance Multiprocessor Ltd. and the RF front-ends were developed at the Uni-

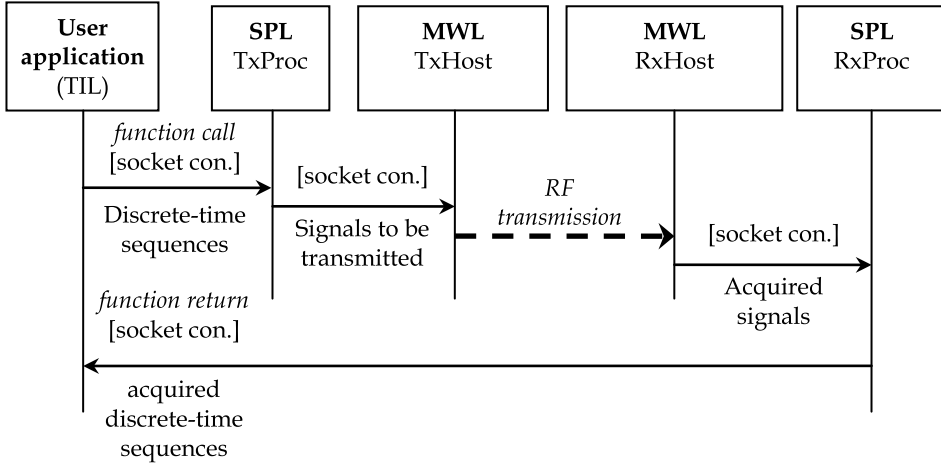


Fig. 5. Frame transmission example. It shows how a single frame transmission is carried out from the discrete-time sequences. For the sake of simplicity, the MWL is considered as a whole, even when it is actually split in two sub-layers.

versity of Cantabria. This hardware has been updated and new RF front-end modules from Lyrtech have been included. Also, minor module reorganisations have been done. Currently, the hardware of the testbed is based on a Sundance Multiprocessor (Sundance, 2009) SMT310Q PCI carrier board and a basic processing module: the Sundance SMT365 equipped with a Xilinx Virtex-II FPGA and a Texas Instruments C6416 DSP at 600 MHz. The processing module has two buses that can transfer 32-bit words up to 400 MB/s, allowing the connection with the Sundance SMT370 module, which contains a dual AD9777 D/A converter and two AD6645 A/D converters. The SMT370 module also has a 2 Msamples per-channel memory that is used to load the frames to be transmitted. At the receiver side, the data acquired by the A/D converters is stored in real-time in two SMT351G memory modules offering 256 Msamples per A/D converter. Finally, in an off-line task, data is passed to the middleware through the PCI bus. At the transmitter side the Sundance SMT365E is used instead of the SMT365. It contains a larger FPGA as well as a larger amount of memory but the remaining features are the same.

5.4 Simple Transmission Example

In order to shed more light into the system architecture, let us explain, step by step, the transmission of a single frame (see Fig. 5). After the discrete-time sequences to be transmitted have been generated at the user application, a function from the TIL is called passing to it the corresponding symbol vectors (one vector per transmitting antenna). These symbols are then sent to the SPL (TxProc) where they are converted (if necessary) to discrete-time signals ready to be transmitted by the hardware. Finally, such signals are subsequently sent to the middleware (TxHost). When both the Tx and Rx PCs are ready to complete a transmission, the signals are passed to the testbed hardware through the corresponding TxDSP process to be transmitted by the antennas. At the receiver side, the MWL (RxHost) acquire the signals stored into the hardware buffers by RxDSP. Next, they are forwarded to the receiver SPL (RxProc). At this

moment, the Tx-Rx PCs are ready to process another frame while the acquired signals are converted to discrete-time sequences in the same format required by the end user. Finally, the discrete-time sequences are forwarded to the user application through the TIL, completing the entire process.

The previous example clearly shows the advantages derived from the use of a multilayer architecture instead of a monolithic system: the hardware is better exploited and each layer can be customized or extended according to specific capabilities of the hardware or particular needs demanded by users, all without developing a completely new monolithic system each time the specifications change. The last advantage turns especially valuable in a research environment where specifications change very fast.

5.5 Testbed Interface Layer

After describing the complete system architecture in the previous sections, this and the following sub-sections focus on the layers that compound the architecture designed for our MIMO testbed (Ramírez et al., 2008). In this subsection we deal with the topmost testbed interface layer (TIL), leaving the description of the other two lower-level layers for the next sub-sections.

The TIL interacts with the researcher or the final user by calling a simple function implemented and specifically adapted to the development environment under consideration, with the only requirement of supporting standard TCP socket connections. At the transmitter side, its main task consists in sending to the SPL the discrete-time sequences to be transmitted plus the necessary parameters. In the same way, the TIL receives the acquired signals, being noticed if any error occurs.

The main goal of the TIL is making the rest of the layers accessible to the final users by taking into account the type of development environment they use. For this reason, the user layer is jointly executed with the user application (see Fig. 3 and Fig. 4), being also integrated in the same process (or set of processes) forming the user application (i.e. a simulation or a demonstration software showing the rest of the testbed capabilities). Therefore, a different TIL implementation satisfying the particular user development environment requirements can be made available (Matlab, Simulink, C/C++, Java, etc).

5.6 Signal Processing Layer

The signal processing layer (SPL) is network-connected to both the TIL and the MWL (see Fig. 3 and Fig. 4). It provides remote access and makes the other two layers platform-independent with respect to each other. This fact permits the SPL to be run in a PC cluster, allowing intensive computing to generate the data to be transmitted and to process the acquired sequences. This layer consists of two different processes that carry out the signal processing operations needed to link the TIL and the MWL. The first process (TxProc) receives the discrete sequences to be transmitted from the TIL and performs all necessary operations such as up sampling, pulse-shape filtering, I/Q modulation and frame assembly, in order to generate the discrete-time signals that will be sent to the middleware (TxHost). Similarly, the second process (Rx-Proc) waits for the acquired signals from the MWL and performs the time and frequency synchronization operations followed by the demodulation, filtering and down sampling. The resulting vectors are sent to the TIL. It is important to emphasize that the SPL is designed and implemented by clearly separating its two main tasks: interconnection with the adjacent layers (TIL and MWL) and execution of the required signal processing tasks. Depending on the

type of sequences given to TxProc, not all operations will be required either at the transmitter or the receiver side.

The SPL can also incorporate advanced features such as a limited feedback channel for the evaluation of precoded MIMO systems. This feedback channel is easily implemented using the network connection available between the transmitter and the receiver.

A SPL detached from the other architecture layers is fundamental to deploy multiuser MIMO scenarios. After extending the current TxProc and RxProc implementations in order to support multiuser MIMO techniques, the set of TxProc processes run at the transmitter users while the RxProc processes do at the receiver users.

In some cases, the researchers may wish to implement most or even all operations present in the SPL. It is still possible to configure the layer as a bypass, making only use of the interconnection capabilities and not performing any signal processing operation at all.

5.7 Middleware Layer

The middleware layer (MWL) fills the gap between the testbed hardware and the signal processing layer, allowing discrete-time signals to be transferred through the system bus and making possible the synchronization between the TxPC and the RxPC using standard TCP network connections. The MWL is split into two different sub-layers (see Fig. 3 and Fig. 4). The topmost sub-layer is responsible of establishing the network connections between the transmitter and the receiver, and with the higher layer (the SPL). The bottom sub-layer corresponds to the testbed hardware configuration and control software. Fig. 3 and Fig. 4 shows the MWL with its two sub-layers plus the connections with adjacent layers.

Four different processes constitute the middleware. The first two, termed TxHost and RxHost, run respectively on the TxPC and RxPC. They are implemented in standard C++ language and use sockets to establish the necessary network connections:

- One connection between the TxHost and RxHost processes. It is used to synchronize the transmitter and the receiver, so the receiver knows when the signal acquisition process has to start.
- Another connection is established between the TxHost and the TxProc processes and it is used to link the transmitter side of both the middleware and the signal processing layers.
- Finally, there is also a connection between the RxHost and the RxProc processes.

The remaining two processes are the transmitter and the receiver processes that run on their respective Digital Signal Processors (DSPs) available in the testbed hardware. Thus, the transmitter DSP process (TxDSP) performs data transfers through the PCI bus jointly with the TxHost process and configures and controls the hardware components at the TxPC. In the same way, the RxHost process and the DSP receiver process (RxDSP) are responsible of transferring the data through the PCI bus and, from the DSP side, controlling and configuring the testbed hardware components.

The MWL serves the maximum number of requests that the hardware supports. It serves a request while the data for the next frame is still being generated by any computer in the network and, when the first frame is passed to the SPL, the MWL is ready to accept a new frame. With this architecture scheme, the MWL simultaneously serves requests from other SPL instances running in different PCs.

For the specific case of our MIMO testbed, in order to release the MWL implementation from dealing with the lowest level hardware details, the Sundance SMT6025 software development

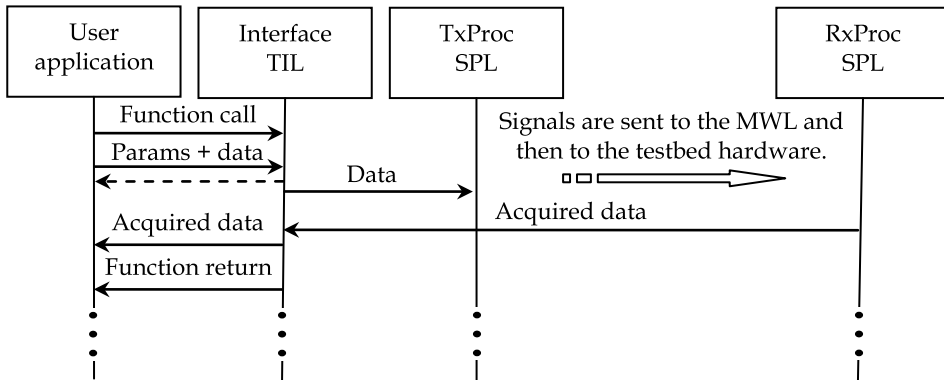


Fig. 6. Sequence diagram describing the interaction among the user application, the testbed interface layer (TIL) and the signal processing layer (SPL).

kit and the Sundance SMT6300 operating system driver are used in the TxHost and the RxHost implementations. The Texas Instruments Code Composer (Texas Instruments, 2009), as well as the 3L Diamond DSP+FPGA (3L Ltd., 2009), are utilized for the TxDSP and RxDSP implementations. If a different hardware type is considered, there should exist different tools avoiding dealing with the lowest-level hardware details.

The middleware concept constitutes a great leap forward in MIMO testbed technology, allowing access to hardware using ordinary network connections and allowing synchronization between the transmitter and the receiver. It also permits to incorporate a wired feedback channel. The MWL design supports extensibility in many senses. For instance, many transmitter and receivers can be used through broadcast network connections. Also, although the DSP processes are hardware dependent, the control logic between the transmitter and the receiver, as well as with the higher layers, is valid for different hardware testbeds. Finally, since the hardware interfaces from the host side and the software related to DSPs are developed as separated software modules, they can be replaced by implementations suitable for any other testbed.

6. Interaction between TIL, SPL and the User Application

Fig. 6 shows a simplified version of the interaction among the user application, the TIL and the SPL. When the user application has generated the discrete-time sequences to be transmitted, the testbed calls some function in the TIL. Both sequences, as well as their type, are included as input parameters. Then, depending on the specified additional input parameters, the TIL offers the possibility to return the control to the user application in order to carry on doing other tasks while the signals are being transmitted and acquired. The TIL establishes a connection with the TxProc at the SPL and sends the transmit data. Next, the interaction described in Fig. 7 takes place and, as a result, the signals are properly transmitted and afterwards acquired by the MWL and sent to the RxProc at the SPL. Finally, the acquired data is transferred to the TIL and sent to the user application. If the user application got the control after calling the TIL, then the acquired sequences are not transferred to the TIL until the user application calls again the TIL.

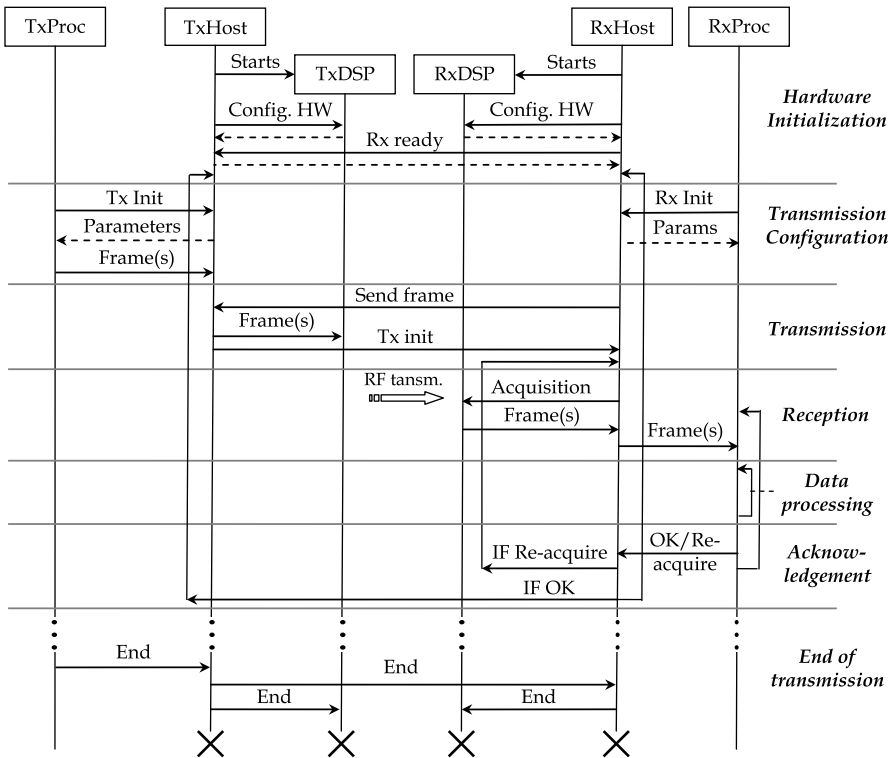


Fig. 7. Basic sequence diagram of one of the variety of protocols implemented in our testbed. It is termed basic operation mode. Dashed arrows represent acknowledgements (ACK). Regular arrows represent messages sent from a source to a destination.

7. Interaction between SPL, MWL and the Hardware

In this section we will explain some functioning modes of the testbed architecture. All of them have been implemented in our MIMO testbed and they allow us different operational modes.

7.1 Basic Mode Operation

Fig. 7 shows the sequence diagram of the "basic operation mode protocol". The protocol fully specifies the behaviour of the six processes of the software architecture. Thus, different protocols allow using the testbed and the software architecture in different scenarios and for distinct purposes. The basic operation protocol consists of the following seven stages:

1. **Hardware initialization.** Both transmitter and receiver hardware must be set up before beginning the actual operation. At the operating system level, two ordinary processes are started: TxHost and RxHost, which then load the TxDSP and RxDSP binaries into the respective DSPs. Once loaded, these processes initialize the baseband hardware modules. Additionally, they allocate the memory buffer needed for communication through the PCI bus. In this stage, the "Starts" message ensures that the communication

through the TIL. The transmitter side will continuously send the frame until a new message "Send Frame" is received from RxHost.

4. **Reception.** After receiving the "Tx init" message from TxHost, RxHost sends the "Acquisition" message to RxDSP. Then, RxDSP captures the signals and sends them back to RxHost through the PCI bus. Then, signals are sent to RxProc, properly processed in this layer, and finally they are sent to the TIL, making them available to the user application.
5. **Data processing.** In our testbed data processing is completely carried out off-line, even though hardware permits real-time processing by means of, for instance, available FPGAs. Basically, during this stage the received data are properly adapted and processed.
6. **Acknowledgement.** After processing the data, the receiver must tell the transmitter what to do next. The receiver can ask for retransmission of the last frame or for a new frame. In the first case, the receiver moves on the transmission stage. In the latter, the receiver needs to prepare a new transmission, so it must go back to the transmission configuration stage.
7. **End of transmission.** TxProc receives a finalization message, which is properly propagated to the rest of the testbed processes with the objective of adequately closing all opened resources and to finish all testbed processes.

7.2 Burst Transmission

This case involves the transmission of a number of data frames with the goal of testing a given transmission system. The sequence diagram is the same as the one depicted in Fig. 7 with the following two exceptions:

- At the beginning of the transmission configuration stage, TxProc notifies the number of frames to be transmitted and sends them to TxHost.
- After finishing each single-frame transmission (probably with a certain number of re-acquisitions carried out by the receiver side) TxHost automatically moves on the next frame to be transmitted.

This mode can be exploited to emulate a real wireless communication system. For instance, a video can be sent in real-time and the users can watch it at the receiver side.

7.3 Performance Measurements

Another case of use is the automation of performance measurements corresponding to a particular algorithm or a specific transmission system. In this case, the protocol works in a quite similar way as in burst transmission: during the data processing stage the measurements are performed and the receiver decides whether it has to keep on doing measurements or it has already collected enough data. This functioning mode was used while doing performance measurements in (Pérez-Iglesias et al., 2008; Ramírez et al., 2008), where different STBC techniques were tested in real environments. The data processing stage can be modified in order to just store the acquired signals at the receiver side to be later on processed off-line.

7.4 Using a Feedback Channel

MIMO systems performance can be improved if the receiver feeds back Channel State Information (CSI) from the receiver to the transmitter, so it can adapt its transmission scheme to the actual channel characteristics. It is straightforward to implement a feedback channel within

the protocol: the CSI information must be sent during the acknowledgement stage to notify the transmitter whether it has to re-configure the transmission parameters or not. To do so, in the sequence diagram shown in Fig. 7 there should be added two "Send CSI" messages: one from the RxHost to the TxHost and another from the TxHost to the TxProc.

If the signal adaptation at the transmitter side takes place at the user application, then no feedback messages are needed in the protocol because everything is available at the user application. It is necessary to call the testbed interface layer twice: one to be able to estimate the channel and the following with the adapted signals.

7.5 Deployment of a Multiuser Environment

The last given case of study describes the functioning of the designed protocol when working in a multiuser environment where several nodes receive information from a central node (broadcast channel). The protocol just needs to be generalized to support the synchronization of several receiver nodes.

Fig. 8 presents the use case for two receiver users. In order to implement this scenario it is necessary to incorporate some modifications:

- After configuring the testbed hardware, TxHost waits for the "ready" message from all receive users.
- TxHost waits for the "Send Frame" message from all users before starting the transmission.
- The frame is cyclically transmitted and each individual user can carry out as many acquisitions as desired. Once all users have acquired all necessary data, they send an "OK" message to the TxHost. When TxHost has received all "OK" messages from all users, the current frame transmission is finished.
- When TxProc receives an end message, it passes it to TxHost which launches a multicast message to all receive users in order to properly finish.

7.6 Using a Simpler Protocol

In some cases it is interesting to provide a simpler protocol in order to delegate most of the control mechanisms to the user application. Now the TIL has two different functions allowing the control of the transmitter and the receiver separately. After the hardware initialization step (see Fig. 7), the user application decides to send a set of discrete-time sequences and then uses the TIL to send them to the SPL. Afterwards, the SPL sends them to the MWL to be transmitted, but RxHost is waiting for RxProc in order to carry out the acquisition. When the user application decides to acquire the data, it calls again the TIL which asks RxProc to acquire the data. Immediately, RxProc asks RxHost to acquire the data and returns it to the user application through the RxProc and the TIL.

The main difference of this functioning mode with respect to the basic operation mode is the lack of messages between TxHost and RxHost, which are now completely autonomous. This presents the advantage of easier integration of the testbed nodes with other nodes and, at the same time, all nodes can be directly controlled from the user application. However, as a disadvantage, the testbed control is moved to the user application. Consequently, the user application has now to deal with some hardware details (e.g. the size of buffers) but properly encapsulated in the TIL. In addition, performance losses arise as a consequence of the testbed being controlled from the user application.

8. Conclusion

In this chapter we propose a new distributed multilayer software architecture for MIMO testbed user access. The architecture fills the gap that currently exists between commercial hardware components and the most common abstraction level used by researchers. The architecture overcomes the limitations of implementing new algorithms directly from the testbed. Instead of using the low-level interfaces typically provided by manufacturers, the architecture supplies a high-level interface access for testbeds. It releases researchers from the necessity of knowing the details of the testbed hardware. For instance, they can easily test new algorithms without developing a completely new source code release specifically for the testbed, thus speeding up the implementation and test tasks.

The key point to the multilayer architecture design is the use of ordinary network connections to link both software modules and hardware components. These connections are also useful for decoupling the software layers. Several advantages are obtained as a consequence of the multilayer software architecture. Also, multiuser scenarios and feedback channels can be constructed extending the proposed architecture. Finally, it is also possible to simplify the presented architecture protocols in order to ease the integration with heterogeneous nodes. The testbed control is moved to the user application instead of keeping it in the inner layers.

9. References

- 3L Ltd (2009), <http://www.3l.com/>
- Borkowski, D.; Brhl, L.; Degen, C.; Keusgen, W.; Alirezai, G.; Geschewski, F.; Oikonomopoulos, C. & Rembold, B. (2006). SABA: A testbed for real-time MIMO systems. *EURASIP Journal on Applied Signal Processing*, Vol. 2006, 2006.
- Caban, S.; Mehlführer, C.; Langwieser, R.; Scholtz, A.L. & Rupp, M. (2005), Vienna MIMO Testbed, *EURASIP Journal on Applied Signal Processing*, Vol. 2006, 2006.
- Fabregas, A.G.; Guillaud, M.; Slock, D.T.M.; Caire, G.; Gosse, G.; Rouquette, S.; Ribeiro Dias, A.; Bernardin, P.; Miet, X.; Conrat, J.M; Toutain, Y.; Peden, A. & Li, Z. (2005), A MIMO-OFDM Testbed for Wireless Local Area Networks, *EURASIP Journal on Applied Signal Processing*, Vol. 2006, 2006.
- Foschini, G. & Gans, M. (1998), On Limits of Wireless Communications in a Fading Environment when Using Multiple Antennas, *Wireless Personal Communications*, Vol. 6, 1998, pp. 311-335.
- García-Naya, José A.; González-López, M. & Castedo, L. (2008), An Overview of MIMO Testbed Technology, *Proceedings of 4th International Symposium on Image/Video Communications over Fixed and Mobile Networks (ISIVC 2008)*, July, 2008, Bilbao
- García-Naya, José A.; Pérez-Iglesias, Héctor J.; Fernández-Caramés, T. M.; González-López, M. & Castedo, L. (2008), A distributed multilayer architecture enabling end-user access to MIMO testbeds, *Proceedings of IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, 2008 (PIMRC 2008)*, September, 2008, Cannes.
- Haustein, T.; Forck, A.; Gäber, H.; Jungnickel, V. & Schiffermüller, S. (2005), Real-Time Signal Processing for Multiantenna Systems: Algorithms, Optimization, and Implementation on an Experimental Test-Bed, *EURASIP Journal on Applied Signal Processing*, Vol. 2006, 2006.

- Kaiser, T.; Wilzeck, A.; Berentsen, M. & Rupp, M. (2004). Prototyping for MIMO Systems An Overview, *Proceedings of the XII European Signal Processing Conference (EUSIPCO 2004)*, September, 2004, Vienna.
- The Mathworks (2009), <http://www.mathworks.com/>
- Moore, G. (1998), Cramming more components onto integrated circuits, *Proceedings of the IEEE*, Vol. 86, No. 1, 1998, pp. 82–85.
- Nieto, X.; Ventura, L.M. & Mollfulleda, A. (2006), GEDOMIS: a broadband wireless MIMO-OFDM testbed, design and implementation, *Proceedings of 2nd International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (TRIDENTCOM 2006)*, 2006, Barcelona.
- PXI: PCI eXtensions for Instrumentation (2009), <http://www.pxisa.org>
- Pérez-Iglesias, H.J.; García-Naya, J.A.; Dapena, A.; Castedo, L. & Zarzoso, V. (2008), Blind channel identification in Alamouti coded systems: a comparative study of eigendecomposition methods in indoor transmissions at 2.4 GHz, *European Transactions on Telecommunications*, Vol. 19, No. 7, November 2008, pp. 751–759.
- Ramírez, D.; Santamaría, I.; Pérez, J.; Vía, J.; García-Naya, J.A.; Fernández-Caramés, T.M.; Pérez-Iglesias, H.J.; González López, M.; Castedo, L. & Torres-Royo, J.M. (2008), A comparative study of STBC transmissions at 2.4 GHz over indoor channels using a 2×2 MIMO testbed, *Wireless Communications and Mobile Computing*, Vol. 8, No. 9, November 2008.
- Rao, R.M.; Wiejun Zhu Lang, S.; Oberli, C.; Browne, D.; Bhatia, J.; Frigon, J.-F.; Jingming Wang Gupta, P.; Heechoon Lee; Liu, D.N.; Wong, S.G.; Fitz, M.; Daneshrad, B. & Takeshita, O. (2004), Multi-antenna testbeds for research and education in wireless communications, *Communications Magazine, IEEE*, Vol. 42, No. 12, December 2004, pp. 72–81, ISSN: 0163-6804.
- Rupp, M.; Burg, A. & Beck, E. (2003). Rapid Prototyping for Wireless Designs: the Five-Ones Approach, *Signal Processing Europe*, Vol. 83, 2003, pp. 1427–1444.
- Rupp, M.; Mehlhrer, C. & Caban, S. (2006), Testbeds and Rapid Prototyping in Wireless System Design, *EURASIP Newsletter*, Vol. 17, No. 3, 2006, pp. 32–50.
- Rupp, M.; Caban, S. & Mehlhrer, C. (2007), Challenges in Building MIMO Testbeds, *Proceedings of European Signal Processing Conference (EUSIPCO 2007)*, 2006, Poznan.
- SDR Forum: Software Defined Radio Forum (2009), <http://www.sdrforum.org>
- SCA: Software Communication Architecture (2009), <http://sca.jpoejtrs.mil>
- Sundance Multiprocessor, Ltd. (2009), <http://www.sundance.com>
- Telatar, I. E. (1999), Capacity of Multi-Antenna Gaussian Channels, *European Transactions on Telecommunications*, Vol. 10, No. 6, 1999, pp. 585–595.
- Texas Instruments (2009), <http://www.ti.com>
- Trygve, M. H. (1978), MVC, XEROX PARC, <http://heim.ifi.uio.no/~trygver/themes/mvc/mvc-index.html>
- Wilzeck, A.; El-Hadidy, M.; Cai, Q.; Amelingmeyer, M. & Kaiser, T. (2006). MIMO Prototyping Testbed with off-the-shelf plug-in RF Hardware, *IEEE Workshop on Smart Antennas (WSA 2006)*, 2006 Ulm.
- Zhu, W.; Browne, D. & Fitz, M. (2005). An Open Access Wideband Multi-Antenna Wireless Testbed with Remote Control Capability, *Proceedings of 2nd International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (TRIDENTCOM 2005)*, Trento.

Recent Developments in Channel Estimation and Detection for MIMO Systems

Seyed Mohammad-Sajad Sadough
*Shahid Beheshti University, Faculty of Electrical and Computer Engineering
 Iran*

Mohammad-Ali Khalighi
*Institut Fresnel, UMR CNRS 6133, École Centrale Marseille
 France*

1. Introduction

It is well known that multiple-input multiple-output (MIMO) systems can provide very high spectral efficiencies in a rich scattering propagation medium Telatar (1999). They are hence a promising solution for high-speed, spectrally efficient, and reliable wireless communication Raoof et al. (2008). When coherent signal detection is to be performed at the receiver, channel state information at the receiver (CSIR) is required, for which a channel estimation step is necessary. Channel estimation plays a critical role in the performance of the receiver. It is a real challenge in practical MIMO systems where the quality of data recovery is as important as attaining a high data throughput.

In order to obtain the CSIR, usually some known training (also called pilot) symbols are sent from the transmitter, based on which the receiver estimates the channel before proceeding to the detection of data symbols. The classical approach consists in time-multiplexing pilot and data symbols, usually referred to as pilot symbol-assisted modulation (PSAM) Cavers (1991). We start the chapter by introducing the PSAM channel estimation for MIMO systems (Section 2). Instead of this classical channel estimation based on pilot symbols only, we can perform *semi-blind* estimation that in addition to pilot symbols, makes use of data symbols in channel estimation. In this way, a considerable performance improvement can be achieved at the price of increased receiver complexity De Carvalho & Slock (1997); Giannakis et al. (2001); Sadough (2008). Usually, these semi-blind approaches are implemented in an iterative scheme when channel coding is performed. That is, channel estimation is performed iteratively together with signal detection and channel decoding Sadough, Ichir, Duhamel & Jaffrot (2009). We, hence, continue Section 2 by considering semi-blind estimation for the case of time-multiplexed pilots.

The drawback of the PSAM scheme is the encountered loss in the spectral efficiency by the periodic insertion of pilot symbols. As an alternative to this method, overlay pilots (OP) can be employed, where pilot symbols are sent in parallel with data symbols Hoeher & Tufvesson (1999). We introduce the OP approach in Section 3 and explain, in particular, pilot-only-based

and semi-blind estimation approaches. The pros and cons of OP with respect to PSAM are discussed too.

Whatever the channel estimation technique, in practice, the receiver can only obtain an *imperfect* estimate of the channel. Classically, for signal detection, the estimated channel is considered as the perfect estimate. This sub-optimal approach is usually called the *mismatched* receiver. Its sub-optimality is due to the fact that the receiver does not take into account the presence of channel estimation errors Sadough & Duhamel (2008); Sadough (2008). A more appropriate approach is to take into account channel estimation inaccuracies in the formulation of the detector. We firstly consider in Section 4 the effect of estimation errors on the receiver performance and the impact of the employed space-time coding scheme.

Next, in Section 5, we consider maximum-likelihood (ML) signal detection and show how to integrate the imperfect channel knowledge into the design of the detector. More precisely, we consider two iterative detectors based on maximum a posteriori (MAP) and soft parallel interference cancellation (soft-PIC), and propose for each case modifications to the MIMO detectors for taking into account the channel estimation errors. The implementation complexity issues are also discussed. We present some numerical results to demonstrate the performance improvement obtained via the use of the improved detectors. Finally, Section 8 concludes the chapter.

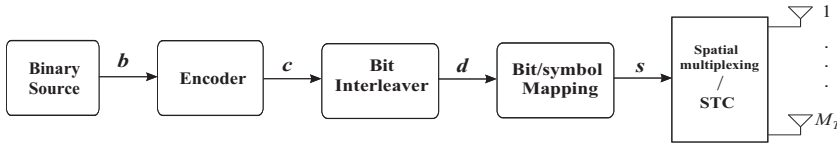


Fig. 1. Transmitter architecture of MIMO-BICM scheme.

1.1 Assumptions and notations

We consider a single-user MIMO system with M_T transmit and M_R receive antennas, transmitting over a frequency non-selective channel and refer to it as an $(M_R \times M_T)$ MIMO channel. Unless otherwise mentioned, single carrier modulation and block fading channel model is considered where the channel is assumed to remain almost constant over the duration of a block of symbols.

Figure 1 shows the block diagram of the transmitter that employs the bit-interleaved coded modulation (BICM) scheme which is known to be a simple and efficient method for exploiting channel time-selectivity. The binary data sequence \mathbf{b} is encoded by a forward error correction (FEC) code before being interleaved by a quasi-random interleaver. The output bits \mathbf{d} are mapped to constellation symbols \mathbf{s} and then either multiplexed spatially or encoded according to a space-time scheme before being sent through the wireless channel. Let us denote by \mathbf{x} and \mathbf{y} respectively the $(M_T \times 1)$ and $(M_R \times 1)$ vectors of transmit and received symbols at a given time reference. For simplicity, we assume for now the simple spatial multiplexing scheme. We have:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z} \quad (1)$$

where \mathbf{H} denotes the $(M_R \times M_T)$ channel matrix and \mathbf{z} is the vector of additive complex white Gaussian noise of zero mean and covariance matrix $\Sigma_z = \sigma_z^2 \mathbb{I}_{M_R}$, where \mathbb{I}_n denotes an $(n \times n)$ Identity matrix. We assume here that σ_z^2 is perfectly known at the receiver and focus on the estimation of \mathbf{H} .

2. Time-multiplexed Pilots and Data

Most current systems use a training-based channel estimation scheme in the form of time-multiplexed pilot symbols. In what follows, we present the PSAM approach for MIMO systems and explain two cases of pilot-only-based and semi-blind estimation; the latter is implemented in an iterative receiver.

2.1 Pilot-only-based PSAM channel estimation

When using the PSAM method, we have a trade-off between the channel estimation quality and the data throughput. With an increased number of pilots, a better channel estimate can be obtained, but at the same time, the spectral efficiency is sacrificed more. There is a minimum number of channel-uses that should be devoted to the transmission of pilots in order that the MIMO channel be identifiable at the receiver. As a general case, if L denotes the maximum length of the underlying subchannels' impulse response, the number of pilot channel-uses N_p should satisfy Balakrishnan et al. (2000):

$$(N_p - L + 1) \geq M_T L \quad (2)$$

Under flat fading conditions where $L = 1$, this implies: $N_p \geq M_T$. Several works have been done on the optimal placement of pilots in a frame of symbols, as well as on the optimal power allocation between pilot and data symbols Hassibi & Hochwald (2003); Dong & Tong (2002); Adireddy et al. (2002); Ma et al. (2003). They consider the criteria of mean-square channel estimation error, channel capacity, or the Cramér-Rao bounds. In particular, it is shown in Hassibi & Hochwald (2003) that, if power optimization over pilot and data symbols is allowed, the optimum N_p is $N_{p\text{opt}} = M_T$. In such a case, we should place pilot symbols with lower power at the beginning and the end of the frame, and those with higher power in the middle of the frame Dong & Tong (2002). However, if equal power has to be allocated to pilot and data symbols, then $N_{p\text{opt}}$ can be larger than M_T Hassibi & Hochwald (2003).

Considering the simple case of flat block-fading MIMO channel, to estimate the channel, corresponding to each fading block, we send N_p pilot symbol vectors with the same power as the data symbols. Let $\mathbf{x}_p[k]$ denote an $(M_T \times 1)$ pilot symbol vector at the time sample k . We denote the received vector corresponding to $\mathbf{x}_p[k]$ by $\mathbf{y}_p[k]$. We can constitute the $(M_T \times N_p)$ matrix \mathbf{X}_p by stacking in its columns the pilot vectors $\mathbf{x}_p[k]$, $k = 0, 1, N_p - 1$, i.e., $\mathbf{X}_p = [\mathbf{x}_p[0], \dots, \mathbf{x}_p[N_p - 1]]$.

According to (1), during a given channel training interval, we have:

$$\mathbf{Y}_p = \mathbf{H} \mathbf{X}_p + \mathbf{Z}_p. \quad (3)$$

The definitions of \mathbf{Y}_p and \mathbf{Z}_p are similar to that of \mathbf{X}_p . We denote by \mathbf{E}_p the average power of the training symbols on any subcarrier as

$$\mathbf{E}_p \triangleq \frac{1}{N_p M_T} \text{tr}(\mathbf{X}_p \mathbf{X}_p^\dagger). \quad (4)$$

The maximum likelihood (ML) channel estimate $\hat{\mathbf{H}}$, which is equivalent to the least-squares solution, is Balakrishnan et al. (2000):

$$\hat{\mathbf{H}} = \left(\sum_{k=0}^{N_p-1} \mathbf{y}_p[k] \mathbf{x}_p^\dagger[k] \right) \left(\sum_{k=0}^{N_p-1} \mathbf{x}_p[k] \mathbf{x}_p^\dagger[k] \right)^{-1}, \quad (5)$$

which can be written in a more compact form as:

$$\hat{\mathbf{H}} = \mathbf{Y}_p \mathbf{X}_p^\dagger (\mathbf{X}_p \mathbf{X}_p^\dagger)^{-1}, \quad (6)$$

where \cdot^\dagger denotes transpose-conjugate.

Let us denote by \mathcal{E} the matrix of estimation errors, that is, $\mathcal{E} = \hat{\mathbf{H}} - \mathbf{H}$. From (3) and (6), it is easy to show that

$$\mathcal{E} = \mathbf{Z}_p \mathbf{X}_p^\dagger (\mathbf{X}_p \mathbf{X}_p^\dagger)^{-1}. \quad (7)$$

It is known that the best channel estimate is obtained with mutually orthogonal training sequences, which result in uncorrelated estimation errors. In other words, we should choose \mathbf{X}_p with orthogonal rows such that

$$\mathbf{X}_p \mathbf{X}_p^\dagger = N_p E_p \mathbb{I}_{M_T}. \quad (8)$$

Then, the j -th column \mathcal{E}_j of \mathcal{E} has the covariance matrix Σ given by

$$\Sigma_e = \mathbb{E}[\mathcal{E}_j \mathcal{E}_j^\dagger] = \sigma_e^2 \mathbb{I}_{M_R}, \quad \text{where } \sigma_e^2 = \frac{\sigma_z^2}{N_p E_p}. \quad (9)$$

2.1.1 Statistics of the Channel Estimation Errors

We saw that the estimated channel matrix $\hat{\mathbf{H}}$ can be viewed as a noisy version of the perfect channel matrix \mathbf{H} . In Section 5 we show that for channel estimators having this feature, the detection performance can be improved if the statistics of the channel estimation errors are known.

Let us reconsider the pilot-based ML channel estimator of equation (7). The good feature of this estimator is that the statistics of the channel estimation error matrix \mathcal{E} are known (see equation (9)). By using these statistics and equation (7), the conditional pdf of $\hat{\mathbf{H}}$ given \mathbf{H} can be easily expressed as:

$$p(\hat{\mathbf{H}}|\mathbf{H}) = \mathcal{CN}(\mathbf{H}, \mathbb{I}_{M_T} \otimes \Sigma_e), \quad (10)$$

where \otimes denotes the Kronecker product and \mathcal{CN} denotes the complex Gaussian distribution. Furthermore, we assume that the channel matrix \mathbf{H} has a normal *prior* distribution as:

$$\mathbf{H} \sim \mathcal{CN}(\mathbf{0}, \mathbb{I}_{M_T} \otimes \Sigma_H) = \frac{1}{\pi^{M_R M_T} \det\{\Sigma_H\}^{M_T}} \exp \left\{ -\text{tr}(\mathbf{H} \Sigma_H^{-1} \mathbf{H}^\dagger) \right\} \quad (11)$$

where Σ_H is the $(M_R \times M_R)$ covariance matrix of the columns of \mathbf{H} , and $\det\{\cdot\}$ denotes matrix determinant. We assume that the entries of \mathbf{H} , i.e., the fading coefficients of different subchannels, are i.i.d. Then, Σ_H is a diagonal matrix with equal diagonal entries σ_h^2 .

By using the prior pdf of \mathbf{H} from (11) and the pdf of $(\hat{\mathbf{H}}|\mathbf{H})$ from (10), we can derive the *posterior* distribution of the perfect channel matrix, conditioned on its ML estimate, as follows (see Sadough & Duhamel (2008) for the details of the derivation):

$$p(\mathbf{H}|\hat{\mathbf{H}}) = \mathcal{CN}(\Sigma_\Delta \hat{\mathbf{H}}, \mathbb{I}_{M_T} \otimes \Sigma_\Delta \Sigma_e), \quad (12)$$

where

$$\Sigma_\Delta = \Sigma_H (\Sigma_e + \Sigma_H)^{-1}. \quad (13)$$

Under the above-mentioned assumptions, we have

$$\Sigma_\Delta = \delta \mathbb{I}_{M_R} \quad (14)$$

where

$$\delta = \frac{\sigma_h^2}{\sigma_h^2 + \sigma_e^2} . \quad (15)$$

In particular, when the number of pilot symbols tends to infinity, it is not difficult to see that $\delta \rightarrow 1$ and $\delta \sigma_e^2 \rightarrow 0$ and consequently $p(\mathbf{H}|\hat{\mathbf{H}})$ tends to a Dirac delta function. The availability of the estimation error distribution is an interesting feature of pilot-only-based PSAM channel estimation that we used to derive the posterior distribution (12). This distribution constitutes a Bayesian framework which is exploited in Section 5 for the design of detectors by taking into account channel estimation inaccuracies.

2.2 Semi-blind PSAM channel estimation

In order to preserve the spectral efficiency for the transmission of data symbols, we are interested in minimizing the number of pilot symbols in a frame. However, by reducing the number of pilot symbols, the channel may be learned improperly and channel estimation errors may become important. This can result in a considerable performance degradation and in the need to data retransmission. This performance degradation can be compensated by smart signal processing at the receiver. In fact, instead of estimation methods based on pilot symbols only, we can use *semi-blind* approaches that in addition to pilot symbols, make use of data symbols in channel estimation. In this way, a considerable performance improvement can be achieved at the price of increased receiver complexity de Carvalho & Slock (2001); Sadough, Ichir, Duhamel & Jaffrot (2009). We present here two semi-blind channel estimation schemes that we implement in an iterative receiver. The first semi-blind method that we consider is the *thresholded hard-decisions* (Th-HD) method and the second one is based on the expectation maximization (EM) algorithm Dempster et al. (1977); Moon (1996). For both methods, at the first iteration, we calculate a primary channel estimate based on the pilot sequences only, which allows the semi-blind estimator, used in the succeeding iterations, to bootstrap. Before describing these methods, we present in the following details on the iterative receiver.

2.2.1 Iterative signal detection

We usually consider in this paper iterative signal detection in the case of using non-orthogonal space-time codes at the transmitter. As shown in Fig. 2, the receiver mainly consists of a combination of two sub-blocks that exchange soft information with each other. The first sub-block, referred to as soft detector or demapper, produces extrinsic soft information from the input symbols and send it to the second sub-block, the soft-input soft-output (SISO) channel decoder. Here, we consider SISO channel decoding based on the well known forward-backward algorithm Bahl et al. (1974). Soft MIMO signal detection and soft-input SISO channel decoding are performed iteratively and the estimates of the channel coefficients are updated at each iteration of the turbo-detector Sadough, Ichir, Duhamel & Jaffrot (2009); Berthet et al. (2001). The blocks Π and Π^{-1} denote bit-level interleaver and de-interleaver, respectively, corresponding to the BICM scheme used at the transmitter.

2.2.2 Th-HD semi-blind estimation

In the Th-HD method, in addition to pilot symbols, we use in channel estimation the symbols detected with high reliability at each iteration Khalighi & Boutros (2006); Sellathurai & Haykin (2002). For instance, consider the *a posteriori* probability $P_i^{(m)}$ at the decoder output at iteration m , corresponding to the coded bit c_i . We compare it with a threshold $0.5 < P_{TH} < 1$. If $P_i^{(m)} > P_{TH}$, we make the hard decision $\hat{c}_i^{(m+1)} = 1$; otherwise,

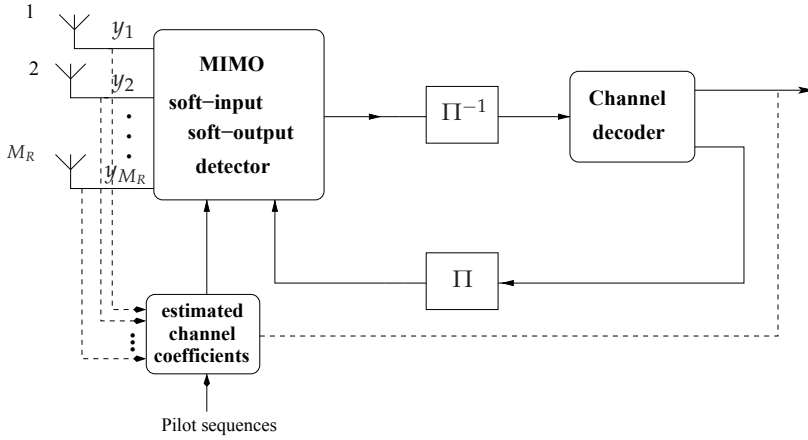


Fig. 2. Iterative channel estimation and data detection, y_i denotes the received signal on the i th antenna.

if $P_i^{(m)} < (1 - P_{TH})$, we make the hard decision $\hat{c}_i^{(m+1)} = 0$; and if none of these conditions are verified, we give up the channel-use corresponding to c_i and do not consider it in channel estimation. If a hard decision is made on all BM_T constituting bits of a channel-use, we use the resulting hard-detected symbol vector in channel estimation, in the same way as pilot symbols. The resulting channel estimate is then used in the next iteration of the detector.

The performance of Th-HD depends highly on the choice of the threshold P_{TH} that determines whether or not the SISO decoder soft-outputs are reliable enough. The practical limitation is that the optimum threshold value depends on the MIMO structure, i.e., the number of transmit and receive antennas, as well as on the actual SNR Khalighi & Boutros (2006). Note that, if we take P_{TH} very close to 0.5, we effectively make hard decisions on all detected symbols and use them in channel estimation. This coincides with the so called decision-feedback channel estimation Visoz & Berthet (2003).

2.2.3 EM-based semi-blind estimation

The interest of the EM algorithm is that it is guaranteed to be stable and to converge to an ML estimate Moon & Stirling (2000). We do not present here the details on the formulation of the EM-based estimator and refer the reader to Khalighi & Boutros (2006), for instance. A simple and classical formulation is when data and pilot symbols are used in the same way in channel estimation. By this approach, the estimated channel matrix is given below:

$$\hat{\mathbf{H}} = \bar{\mathbf{R}}_{yx} \bar{\mathbf{R}}_x^{-1}, \quad (16)$$

where

$$\bar{\mathbf{R}}_{yx} = \sum_{k=1}^{N_s} \mathbf{y}[k] \tilde{\mathbf{x}}^\dagger[k] \quad (17)$$

and

$$\bar{\mathbf{R}}_{x,i,j} = \begin{cases} N_s & ; \quad i = j \\ \sum_{k=1}^{N_s} \tilde{\mathbf{x}}_i[k] \tilde{\mathbf{x}}_j^*[k] & ; \quad i \neq j \end{cases} \quad (18)$$

Here, N_s denotes the number of channel-uses per frame, $\mathbf{y}[k]$ is the received symbol vector at the time reference k , and $\hat{\mathbf{x}}[k]$ is the corresponding soft-estimates of the transmitted symbol vector, calculated using the SISO channel decoder outputs at the preceding iteration. For $k = 1, \dots, N_p$, we have $\hat{\mathbf{x}}[k] = \mathbf{x}_p[k]$. Also, $\bar{\mathbf{R}}_{x,i,j}$ denotes the (i,j) th entry of matrix $\bar{\mathbf{R}}_x$. The formulation that we provided for EM can be modified further to improve the receiver performance. Interested reader may refer to Khalighi & Boutros (2006); Khalighi et al. (2006) for details.

2.2.4 Case study

BER curves versus E_b/N_0 are shown in Fig. 3 for the case of (8×8) and (8×6) MIMO structures using the PSAM technique. Rayleigh independent quasi-static fading model is considered with blocks of length $N_s = 64$ channel-uses. The number of channel-uses devoted to pilot transmission is $N_p = 10$. Three cases of pilot-only-based estimation, and Th-HD and EM-based semi-blind estimations are considered. The case of perfect-CSIR is also provided as reference. The simple spatial multiplexing scheme is used at the transmitter and MIMO signal detection is based on soft parallel interference cancellation (Soft-PIC) Sellathurai & Haykin (2002); Lee et al. (2006). Results shown in Fig. 3 correspond to the eighth iteration of the receiver. We notice that the performance of the semi-blind Th-HD and EM-based estimator are very close to each other and they outperform the pilot-only-based method. Yet, their performance is about 2 dB away from the perfect-CSIR case that is due to the relatively high co-antenna interference, as we have $M_T = 8$. We can approach further the perfect-CSIR case by increasing N_p .

3. Superimposed Pilots and Data

The main drawback of the PSAM approach is that, for finite-length blocks, if channel estimation is to be done on each block of symbols, the periodic insertion of pilot symbols can result in a considerable reduction of the achievable data rate. This loss in the data rate becomes important, specially for large number of transmit antennas, at low SNR, and when the channel undergoes relatively fast variations Hassibi & Hochwald (2003). As an alternative, we can use overlay pilots (OP), also called superimposed or embedded pilots, for channel estimation Hoeher & Tufvesson (1999); Zhu et al. (2003). In this approach, a pilot sequence is superimposed on the data sequence before transmission, as shown in Fig. 4; thus, no separate time slot is dedicated to pilot transmission.

3.1 Channel estimation using OP

By using OP, we prevent the loss in the data throughput but we experience degradation in the quality of the channel estimate due to the unknown data symbols Hoeher & Tufvesson (1999). As a matter of fact, here also there is a trade-off between high quality channel estimation and the information throughput: To obtain a better channel estimate, we should increase the percentage of the power dedicated to pilot symbols; this, however, reduces the SNR for the detection of data symbols Tapio & Bohlén (2004). In general, OP may be preferred to PSAM for high SNR, not too short channel coherence times, and larger number of receive than transmit antennas Khalighi et al. (2005).

Consider again the block fading channel model. Assuming uncorrelated data and pilot sequences, we can estimate channel coefficients by calculating the cross-correlation between the received sequences on each antenna and the transmitted pilot sequences, known to the

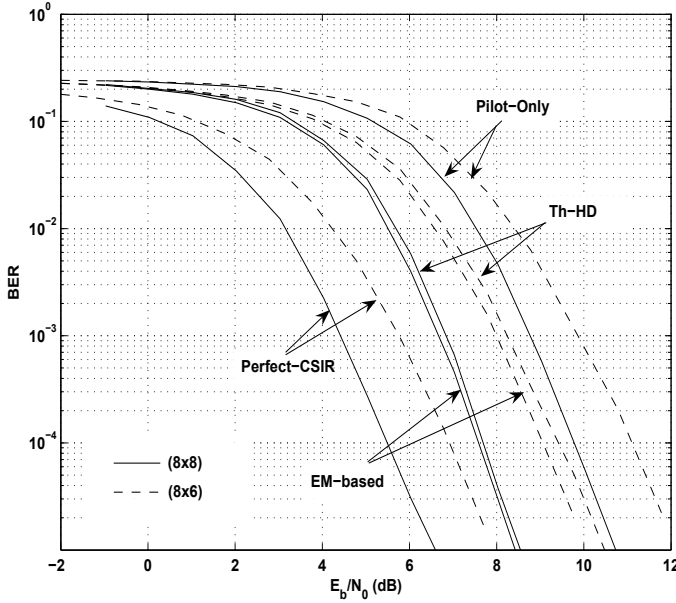


Fig. 3. Comparison of different estimation methods based on PSAM, iterative Soft-PIC detection, 8th iteration of the receiver (almost full convergence); (8×8) and (8×6) systems, i.i.d. Rayleigh quasi-static fading, QPSK modulation, $(5,7)_8$ NRNSC rate 1/2 channel code, orthogonal pilot sequences with $N_p = 10$, $N_s = 64$ channel-uses. E_b/N_0 takes into account the receiver antenna gain M_R .

receiver. As the pilot sequences transmitted from different antennas are orthogonal, the estimation errors arise from noise and the data sequences. Indeed, the main problem in the estimation of channel coefficients concerns the latter interference component, i.e., the unknown data symbols. In fact, although data and pilot sequences are *statistically* uncorrelated, the cross-correlation is calculated over a block of symbols of limited length, over which the channel coefficients are supposed to remain unchanged. The smaller is the block length (i.e., the faster the channel fading), the more important this cross-correlation is. This can result in an error floor in the receiver BER performance Jungnickel et al. (2001), especially at high SNR, and make the OP scheme lose its interest.

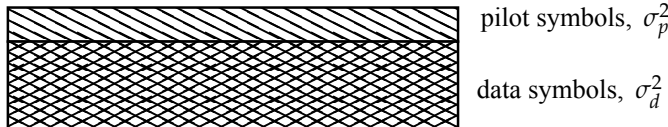


Fig. 4. Overlay pilot scheme

3.2 Iterative channel estimation for OP

Iterative data detection and channel estimation can be a solution to the problem of error floor Zhu et al. (2003); Khalighi et al. (2005); Cui & Tellambura (2005). We consider in the following, two such iterative schemes; a pilot-only-based decision-directed estimator Khalighi et al. (2005) and an EM-based semi-blind estimator Khalighi & Bourennane (2007).

Let us denote by \mathbf{x}_d and \mathbf{x}_p the vectors of data and pilot symbols, respectively, corresponding to the transmitted $(M_T \times 1)$ vector \mathbf{x} : $\mathbf{x} = \mathbf{x}_d + \mathbf{x}_p$. We denote by σ_d^2 and σ_p^2 the power allocated to the entries of \mathbf{x}_d and \mathbf{x}_p , respectively.

3.2.1 Pilot-only-based decision-directed estimator

Consider the estimation of the entry H_{ij} of the channel matrix \mathbf{H} . As explained above, this estimate can be obtained by calculating the cross-correlation Γ_{ij} between the sequence received on the antenna $\#i$, \mathbf{y}_i , and the pilot sequence transmitted on the antenna $\#j$, $\mathbf{x}_{p,j}$. By the decision-directed method that we denote by DD, we use in each iteration, the soft-estimates \tilde{x}_d of transmitted data symbols using a posteriori LLRs at the SISO decoder output, and cancel their effect in Γ_{ij} .

3.2.2 EM-based semi-blind estimator

To obtain a better performance, similar to the case of PSAM, we can employ semi-blind estimation methods Khalighi & Bourennane (2008); Bohlin & Tapio (2004); Meng & Tugnait (2004) at the price of increased Rx complexity. For instance, a semi-blind estimator based on the EM algorithm may be used. Its formulation is analogous to the case of PSAM and can be found in Khalighi & Bourennane (2007).

3.2.3 Data-dependent overlay pilots

A solution for getting rid of the interference from data symbols in channel estimation is to use data-dependent overlay pilot sequences such that the corresponding pilot and data sequences are orthogonal Ghogho et al. (2005). The drawback of this method is that it results in nulls in the equivalent channel impulse response (seen by data symbols), and hence, in a performance degradation. This degradation could be reduced by performing iterative *channel equalization* Lam et al. (2008). On the other hand, such a scheme leads to increased envelope fluctuations of the transmitted signal. We do not consider this scheme here.

3.2.4 Case study

Let us denote by α the ratio of the power of pilot symbols to the total transmit power at a symbol time, i.e., $\alpha = \sigma_p^2 / (\sigma_p^2 + \sigma_d^2)$. For a (2×4) MIMO system, we have presented in Fig. 5, BER curves versus SNR for DD, EM-based, and perfect channel estimation, and different values of α . Pilot sequences for M_T antennas can be QPSK modulated and chosen according to the Walsh-Hadamard series to ensure their orthogonality. Results correspond to the fifth iteration of the Rx where almost full convergence is attained. SNR in Fig. 5 stands for the *actual average* received SNR, i.e., $M_R(\sigma_d^2 + \sigma_p^2) / \sigma_n^2$, in contrast to E_b / N_0 that takes into account only σ_d^2 . In this way, we can directly see the compromise between the channel estimation quality and the data detection performance, e.g. by increasing α . Again, Soft-PIC MIMO detection is performed. For both estimation methods, we notice an error floor at high SNR for $\alpha < 15\%$ which is especially visible for $\alpha = 2\%$ and $\alpha = 5\%$. On the other hand, by increasing α , better channel estimates are obtained, but at the same time, less power is dedicated to data symbols. So, increasing α too much, will result in an overall performance degradation. Comparing

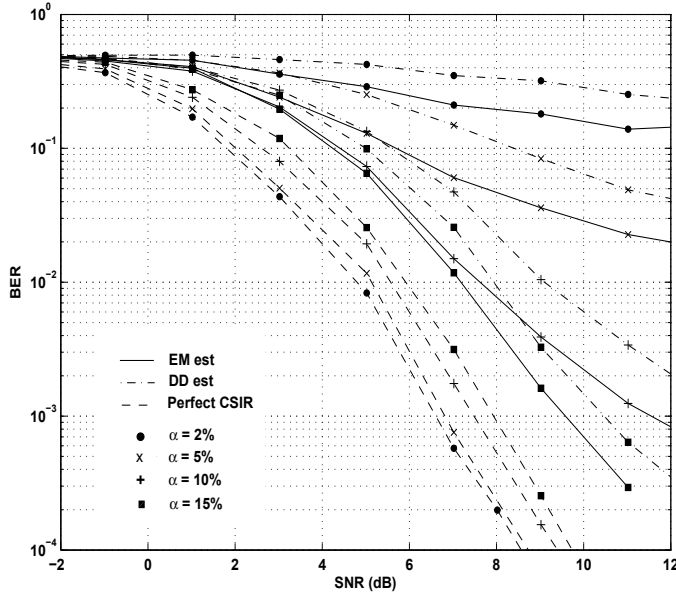


Fig. 5. Overlay Pilot scheme: BER after five receiver iterations for EM-based semi-blind, DD, and perfect channel estimation ; (2×4) MIMO system, i.i.d. Rayleigh quasi-static fading, QPSK-modulated pilot and data, $(133,171)_8$ NRNSC rate 1/2 channel code, $N_s = 100$. SNR takes into account M_R .

the performance curves of the EM-based and DD estimators, we see that the performance improvement by semi-blind estimation is quite considerable. This improvement is more important for smaller α values. For increased α , more power is dedicated to pilot symbols, and hence, the performance of the pilot-only-based estimator approaches that of the semi-blind estimator.

4. Impact of Space-Time Scheme

An important aspect is the the impact of channel estimation errors on the receiver performance for different ST schemes. We would like to compare the two general classes of ST schemes, i.e., orthogonal and non-orthogonal schemes. In the case of perfect channel knowledge, a comparison is made between these two categories in Khalighi et al. (2009), where iterative signal detection based on Soft-PIC is proposed for the case of non-orthogonal schemes. Note that optimal signal detection is too computationally complex for these schemes. Conditioned to the presence of sufficient (time or frequency) diversity, it is shown that a substantial gain is obtained by using the appropriate non-orthogonal schemes for moderate-to-high spectral efficiency MIMO systems, as compared to orthogonal schemes, which justifies the increased complexity of the receiver Khalighi et al. (2009).

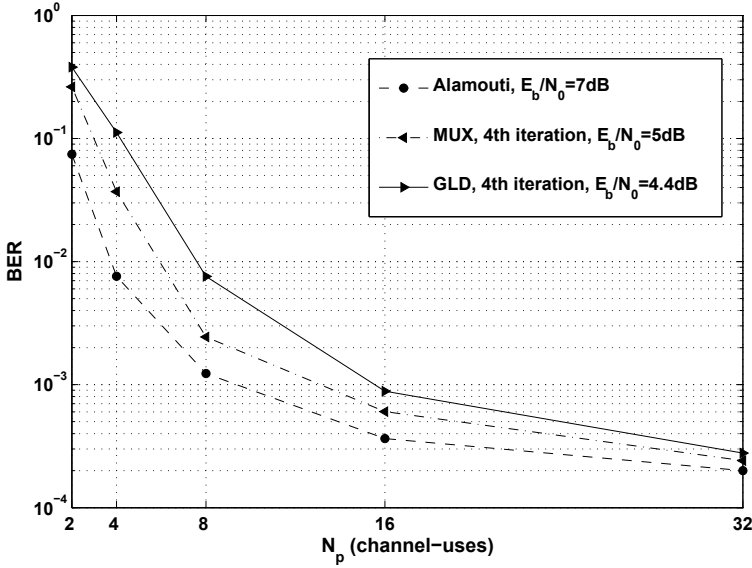


Fig. 6. Sensitivity to channel estimation errors; (2×2) MIMO channel, $\eta = 2$ bps/Hz, $(133, 171)_8$ NRNSC rate 1/2 channel code, i.i.d. Rayleigh flat block fading channel with 32 independent fades per frame of 768 channel-uses. Iterative Soft-PIC used for MUX and GLD.

Concerning the practical case of imperfect channel estimate, in fact, lower-rate orthogonal schemes could be more sensitive to channel estimation errors as, in general, they have to use a larger signal constellation to attain a desired spectral efficiency. Concerning non-orthogonal schemes, we need, in general, smaller constellation sizes as compared to orthogonal ones. However, the iterative detector for the non-orthogonal schemes could be more sensitive to channel estimation errors because its convergence, and hence, its performance is affected by these errors.

To study the effect of channel estimation errors, let us consider pilot-only-based channel estimation using time-multiplexed pilots. For each fading block, we devote N_p channel-uses to the transmission of power-normalized mutually orthogonal QPSK pilot sequences from M_T transmit antennas Khalighi & Boutros (2006). For a (2×2) MIMO system, we have shown in Fig. 6 the average BER after four detector iterations versus N_p for a spectral efficiency of $\eta = 2$ bps/Hz. We have considered the Alamouti code Alamouti (1998) as the orthogonal scheme, and two cases of spatial multiplexing (denoted here by MUX) and Golden coding Belfiore et al. (2005) (denoted by GLD) as non-orthogonal schemes. The E_b/N_0 for each ST scheme is set to what results in $\text{BER} \approx 10^{-4}$ in the case of perfect channel knowledge. From Fig. 6 we notice an almost equivalent sensitivity to the channel estimation errors for MUX and GLD schemes. This comparison makes sense as the SNRs for these schemes are close to each other. On the other hand, we see that the Alamouti scheme has the lowest sensitivity. This is due to the orthogonal structure of the code, and the fact that the SNR is higher, compared to those for MUX and GLD schemes, and as a result, the quality of channel estimate is much better.

5. Improved Signal Detection in the Presence of Channel Estimation Errors

For the case of time-multiplexed pilot and data, as we explained previously, in order to obtain a good channel estimate, we should increase N_p , which in turn, results in a larger loss in the spectral efficiency, specially for relatively fast time-varying communication channels Hassibi & Hochwald (2003). One solution is to use semi-blind channel estimation in order to reduce the number of channel-uses devoted to pilot transmission, as seen in Section 2.2. The disadvantage of semi-blind approaches is the increased receiver complexity. For a reduced-complexity semi-blind joint channel estimator and data detector the reader is referred to Sadough, Ichir, Duhamel & Jaffrot (2009).

An alternative to this solution is to modify the detector so as to take into account channel estimation errors. As a matter of fact, the classical approach is to use the channel estimate in the detection part in the same way as if it was a perfect estimate, what is known as *mismatched* signal detection. Obviously, this approach is suboptimal and can degrade considerably the receiver performance in the presence of channel estimation errors.

In this section we provide the general formulation of a detection rule that takes into account the available imperfect CSIR and refer to it as the *improved* detector. To this end, we consider the model (1) and denote by $J(\mathbf{y}, \mathbf{x}, \mathbf{H})$ the quantity (cost function) that would let us to decide in favor of a particular \mathbf{x} at the receiver if the channel was *perfectly* known. Note that depending on the detection criteria, the quantity $J(\mathbf{y}, \mathbf{x}, \mathbf{H})$ can be the posterior pdf $p(\mathbf{x}|\mathbf{y}, \mathbf{H})$, the logarithm of the likelihood function $p(\mathbf{y}|\mathbf{H}, \mathbf{x})$, the mean square error (as in Sadough, Khalighi & Duhamel (2009); Sadough & Khalighi (2007)), etc. Assume a channel estimator in which the statistics of the estimation errors are known. Such a scenario occurs for instance in pilot-only-based PSAM channel estimation studied in Subsection 2.1 where we saw that the estimation process can be characterized by the posterior pdf of the channel (12). In this case, we propose a detector based on the minimization of a new cost function defined as

$$\bar{J}(\mathbf{y}, \mathbf{x}, \hat{\mathbf{H}}) = \int_{\mathbf{H}} J(\mathbf{y}, \mathbf{x}, \mathbf{H}) p(\mathbf{H}|\hat{\mathbf{H}}) d\mathbf{H} = E_{\mathbf{H}|\hat{\mathbf{H}}} [J(\mathbf{y}, \mathbf{x}, \mathbf{H}) | \hat{\mathbf{H}}] \quad (19)$$

where by using the posterior distribution (12), we have averaged the cost function J over all realizations of the unknown channel \mathbf{H} conditioned on its available estimate $\hat{\mathbf{H}}$. Note that the *mismatched* detector is based on the minimization of the cost function $J(\mathbf{y}, \mathbf{x}, \hat{\mathbf{H}})$. This latter cost function is obtained by using the estimated channel $\hat{\mathbf{H}}$ in the same metric that would be used if the channel was perfectly known, i.e., $J(\mathbf{y}, \mathbf{x}, \mathbf{H})$. Using the metric of (19) differs from the mismatched detection on the conditional expectation $E_{\mathbf{H}|\hat{\mathbf{H}}}[\cdot]$ which provides a robust design by averaging the cost function $J(\mathbf{y}, \mathbf{x}, \mathbf{H})$ over all (true) channel realizations which could correspond to the available estimate.

Consider the problem of detecting symbol vector \mathbf{x} from the observation model (1) in the ML sense, i.e., so as to maximize the likelihood function $p(\mathbf{y}|\mathbf{H}, \mathbf{x})$.

It is well known that under *perfect* channel knowledge and i.i.d. Gaussian noise, detecting \mathbf{x} by maximizing the likelihood $p(\mathbf{y}|\mathbf{x}, \mathbf{H})$ is equivalent to minimizing the Euclidean distance \mathcal{D}_{ML} as

$$\hat{\mathbf{x}}_{\text{ML}}(\mathbf{H}) = \arg \min_{s_0, \dots, s_{M-1} \in \mathcal{C}} \{ \mathcal{D}_{\text{ML}}(\mathbf{x}, \mathbf{y}, \mathbf{H}) \}, \quad (20)$$

with $\mathcal{D}_{\text{ML}}(\mathbf{x}, \mathbf{y}, \mathbf{H}) \triangleq -\log p(\mathbf{y}|\mathbf{H}, \mathbf{x}) \propto \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2$, where \propto means “is proportional to” and \mathcal{C} denotes the set of constellation symbols of size M . Assuming B bits per symbol, we have $M = 2^B$.

The detection rule (20) requires the knowledge of the *perfect* channel matrix \mathbf{H} . The sub-optimal mismatched ML detector consists in replacing the exact channel by its estimate $\hat{\mathbf{H}}$ in the receiver metric as

$$\hat{\mathbf{x}}_{\text{MM}}(\hat{\mathbf{H}}) = \underset{s_0, \dots, s_{M-1} \in \mathcal{C}}{\operatorname{argmin}} \{ \mathcal{D}_{\text{MM}}(\mathbf{x}, \mathbf{y}, \hat{\mathbf{H}}) \} = \underset{s_0, \dots, s_{M-1} \in \mathcal{C}}{\operatorname{argmin}} \{ \|\mathbf{y} - \hat{\mathbf{H}}\mathbf{x}\|^2 \},$$

where

$$\mathcal{D}_{\text{MM}}(\mathbf{x}, \mathbf{y}, \mathbf{H}) \triangleq \mathcal{D}_{\text{ML}}(\mathbf{x}, \mathbf{y}, \mathbf{H}) \Big|_{\mathbf{H}=\hat{\mathbf{H}}}, \quad (21)$$

and the subscript \cdot_{MM} denotes mismatched. Obviously, the sub-optimality of this detection technique is due to the mismatch introduced by the channel estimation errors; while the decision metric is derived from the likelihood function $p(\mathbf{y}|\mathbf{H}, \mathbf{x})$ conditioned on the perfect channel \mathbf{H} , the receiver uses an estimate $\hat{\mathbf{H}}$ different from \mathbf{H} in the detection process.

As an alternative to this mismatched detection, an improved ML detection metric is proposed in Tarokh et al. (1999); Taricco & Biglieri (2005). This metric is based on modified likelihood $p(\mathbf{y}|\hat{\mathbf{H}}, \mathbf{x})$ which is conditioned on the imperfect channel $\hat{\mathbf{H}}$. The pdf $p(\mathbf{y}|\hat{\mathbf{H}}, \mathbf{x})$ can be derived as follows:

$$p(\mathbf{y}|\hat{\mathbf{H}}, \mathbf{x}) = \int_{\mathbf{H} \in \mathcal{C}} p(\mathbf{y}, \mathbf{H}|\hat{\mathbf{H}}, \mathbf{x}) d\mathbf{H} = \int_{\mathbf{H} \in \mathcal{C}} p(\mathbf{y}|\mathbf{H}, \mathbf{x}) p(\mathbf{H}|\hat{\mathbf{H}}) d\mathbf{H} = \mathbb{E}_{\mathbf{H}|\hat{\mathbf{H}}} \left[p(\mathbf{y}|\mathbf{H}, \mathbf{x}) \right], \quad (22)$$

where $p(\mathbf{H}|\hat{\mathbf{H}})$ is the channel posterior distribution of equation (12) and \mathcal{C} denotes the set of complex matrices of size $(M_R \times M_T)$. In fact, equation (22) shows that $p(\mathbf{y}|\hat{\mathbf{H}}, \mathbf{x})$ can be simply derived from the general formulation in (19). It is shown in Sadough & Duhamel (2008) that the averaged likelihood in (22) is shown to be a complex Gaussian distributed vector given by

$$p(\mathbf{y}|\hat{\mathbf{H}}, \mathbf{x}) = \mathcal{CN}(\mathbf{m}_{\mathcal{M}}, \mathbf{\Sigma}_{\mathcal{M}}), \quad (23)$$

where $\mathbf{m}_{\mathcal{M}} = \delta \hat{\mathbf{H}}\mathbf{x}$, and $\mathbf{\Sigma}_{\mathcal{M}} = \mathbf{\Sigma}_z + \delta \sigma_e^2 \|\mathbf{x}\|^2$. Finally, the estimate of the symbol \mathbf{x} is

$$\hat{\mathbf{x}}_{\mathcal{M}}(\hat{\mathbf{H}}) = \underset{s_0, \dots, s_{M-1} \in \mathcal{C}}{\operatorname{argmin}} \{ \mathcal{D}_{\mathcal{M}}(\mathbf{x}, \mathbf{y}, \hat{\mathbf{H}}) \}, \quad (24)$$

where

$$\mathcal{D}_{\mathcal{M}}(\mathbf{x}, \mathbf{y}, \hat{\mathbf{H}}) \triangleq -\log p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{H}}) = M_R \log \pi (\sigma_z^2 + \delta \sigma_e^2 \|\mathbf{x}\|^2) + \frac{\|\mathbf{y} - \delta \hat{\mathbf{H}}\mathbf{x}\|^2}{\sigma_z^2 + \delta \sigma_e^2 \|\mathbf{x}\|^2} \quad (25)$$

is referred to as the *improved* ML decision metric under imperfect CSIR.

Note that when CSIR tends to the exact value, which is obtained when the number of pilot symbols tends to infinity, we have $\delta \rightarrow 1$, $\sigma_e^2 \rightarrow 0$, and the improved metric (25) tends to the mismatched metric:

$$\lim_{N \rightarrow \infty} \frac{\mathcal{D}_{\mathcal{M}}(\mathbf{x}, \mathbf{y}, \hat{\mathbf{H}})}{\mathcal{D}_{\text{MM}}(\mathbf{x}, \mathbf{y}, \hat{\mathbf{H}})} = 1. \quad (26)$$

In the following two sections, we apply the proposed receiver design method of equation (19) for improving the performance of two usually-used MIMO receivers working under imperfect channel estimation: one based on the maximum *a posteriori* (MAP) criterion, and the other on Soft-PIC. For both cases, we consider the simple spatial multiplexing as the space-time scheme at the transmitter, and iterative MIMO detection and channel decoding at the receiver.

6. Reception Scheme I: Iterative MAP Detection

Here, we consider MIMO signal detection based on the MAP algorithm. In the following, we make use of the improved ML metric derived in the previous section to modify the MAP detector part for the case of imperfect CSIR Sadough et al. (2007). Let us denote by $\mathbf{x}[k]$ and $\mathbf{y}[k]$, the transmitted and received symbol vectors corresponding to the time slot k , simply by \mathbf{x}_k and \mathbf{y}_k , respectively. Also, let $d_{k,j}$ denote the j -th ($j = 1, \dots, BM_T$) coded and interleaved bit corresponding to \mathbf{x}_k . We denote by $L(d_{k,j})$ the coded log-likelihood ratio (LLR) of the bit $d_{k,j}$ at the output of the detector. Conditioned on the imperfect CSIR $\hat{\mathbf{H}}_k$, $L(d_{k,j})$ is given by:

$$L(d_{k,j}) = \log \frac{P_{\text{dem}}(d_{k,j} = 1 | \mathbf{y}_k, \hat{\mathbf{H}}_k)}{P_{\text{dem}}(d_{k,j} = 0 | \mathbf{y}_k, \hat{\mathbf{H}}_k)}, \quad (27)$$

where $P_{\text{dem}}(d_{k,j} | \mathbf{y}_k, \hat{\mathbf{H}}_k)$ is the probability of transmission of $d_{k,j}$ at the detector output. We partition the set \mathcal{C} that contains all possibly-transmitted symbol vectors \mathbf{x}_k into two sets \mathcal{C}_0^m and \mathcal{C}_1^m , for which the j -th bit of \mathbf{x}_k equals "0" or "1", respectively. We have:

$$L(d_{k,j}) = \log \frac{\sum_{\mathbf{x}_k \in \mathcal{C}_1^m} e^{-\mathcal{D}_{\mathcal{M}}(\mathbf{x}_k, \mathbf{y}_k, \hat{\mathbf{H}}_k)} \prod_{\substack{i=1 \\ i \neq j}}^{BM_T} P_{\text{dec}}^1(d_{k,i})}{\sum_{\mathbf{x}_k \in \mathcal{C}_0^m} e^{-\mathcal{D}_{\mathcal{M}}(\mathbf{x}_k, \mathbf{y}_k, \hat{\mathbf{H}}_k)} \prod_{\substack{i=1 \\ i \neq j}}^{BM_T} P_{\text{dec}}^0(d_{k,i})}, \quad (28)$$

where $P_{\text{dec}}^1(d_{k,j})$ and $P_{\text{dec}}^0(d_{k,j})$ are *prior* probabilities on the bit $d_{k,j}$ coming from the SISO decoder.

Note that using the metric $\mathcal{D}_{\mathcal{M}}(\mathbf{x}_k, \mathbf{y}_k, \hat{\mathbf{H}}_k)$ for the evaluation of the LLRs in (28) is an alternative to using the mismatched ML metric $\mathcal{D}_{\text{MM}}(\mathbf{x}_k, \mathbf{y}_k, \hat{\mathbf{H}}_k)$ which replaces at each iteration, the exact channel \mathbf{H}_k by its estimate $\hat{\mathbf{H}}_k$ in $\mathcal{D}_{\text{ML}}(\mathbf{x}_k, \mathbf{y}_k, \mathbf{H}_k)$. By doing so, the LLRs are adapted to the imperfect channel knowledge available at the receiver and consequently the impact of channel uncertainty on the SISO decoder performance is reduced. We refer to the latter approach as *improved* MAP detector Sadough et al. (2007).

The summations in (28) are taken over the product of the likelihood $p(\mathbf{y}_k | \mathbf{x}_k, \hat{\mathbf{H}}_k) = e^{-\mathcal{D}_{\mathcal{M}}(\mathbf{x}_k, \mathbf{y}_k, \hat{\mathbf{H}}_k)}$ given a symbol \mathbf{x}_k and the estimated channel coefficient $\hat{\mathbf{H}}_k$, and of the *a priori* probability on \mathbf{x}_k (the term $\prod P_{\text{dec}}$), fed back from the SISO decoder at the previous iteration. In this latter term, the *a priori* probability of the bit $d_{k,j}$ itself has been excluded, so as to let the exchange of *extrinsic* informations between the channel decoder and the soft detector. Also, note that this term assumes independent coded bits $d_{k,j}$, which is a reasonable approximation for random interleaving of large size. At the first iteration, no *a priori* information is available on bits $d_{k,j}$, therefore the probabilities $P_{\text{dec}}^0(d_{k,j})$ and $P_{\text{dec}}^1(d_{k,j})$ are set to 1/2. The decoder accepts the LLRs of all coded bits and computes the LLRs of information bits, which are used for decision, at the last iteration.

6.1 Case study

We now present some numerical results. First, the BER performance of the improved and mismatched detectors are compared. Let us first address the case of BICM iterative decoding with 16-QAM and Gray labeling for a 2×2 MIMO channel. It can be seen from Fig. 7 that for $N_p = 2$ (the shortest possible training sequence), the improvement in terms of required E_b/N_0

in order to attain a BER of 10^{-5} is about 1 dB, compared to the mismatched solution, while staying still 3 dB away from the perfect channel knowledge case. We also notice that, logically, these quantities are reduced when increasing the length of the training sequence, that is, the performances of mismatched and improved detectors get closer to the case of perfect channel knowledge.

Similar plots are shown in Fig. 8 for the case of 16-QAM and set-partition (SP) labeling on the (2×2) MIMO channel. These show the behavior of the detectors with respect to the type of bit-symbol labeling. At a BER of 2×10^{-4} with $N_p = 2$, we obtain an SNR gain of about 1.4 dB by using the improved detector. In other words, iterative decoding with SP labeling benefits more from the improved metric than the one with Gray labeling. Otherwise, similar conclusions hold between the SP-labeling curves.

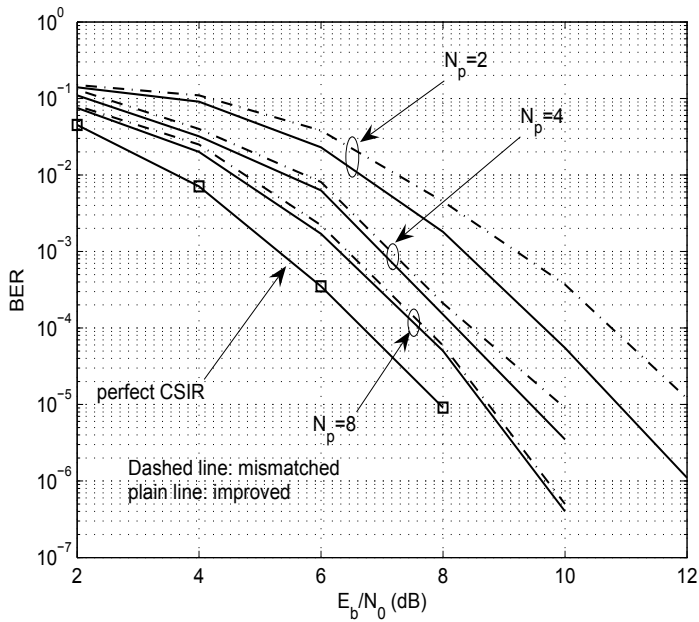


Fig. 7. BER performance improvement over (2×2) MIMO channel with i.i.d. Rayleigh fading for various training sequence lengths. 16-QAM modulation with Gray labeling, iterative MAP detection after four receiver iterations.

7. Reception Scheme II: Iterative Soft-PIC Detection

MAP detection is the optimal solution under perfect CSIR in the sense of bit error rate but its complexity grows exponentially with the number of transmit antennas and the signal constellation size. For this reason, suboptimal detection techniques are usually preferred. One interesting solution is that based on Soft-PIC and linear minimum mean-square error (MMSE) filtering Wang & Poor (1999); Sellathurai & Haykin (2002); Lee et al. (2006), what we considered in Sections 2 to 4. In fact, in these parts, we considered in the iterative detector a sim-

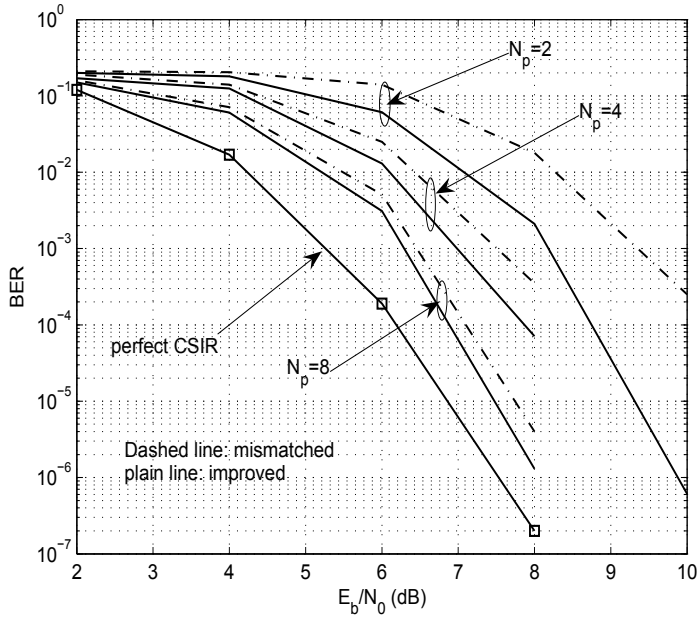


Fig. 8. BER performance improvement over (2×2) MIMO channel with i.i.d. Rayleigh fading for various training sequence lengths. 16-QAM modulation with set-partition labeling, iterative MAP detection after four receiver iterations.

plified formulation of Soft-PIC, which assumes perfect interference cancellation after the first iteration. In this section, however, we consider the exact formulation of Soft-PIC. In order to better understand the formulation of the improved detector, we present in the following the formulation of (exact) Soft-PIC under perfect channel knowledge at the receiver. Then, we present the improved Soft-PIC detector in the presence of channel estimation errors in Subsection 7.2.

7.1 Soft-PIC detection under perfect channel knowledge

Consider the general block diagram of Fig. 2. Here, to detect a symbol transmitted from a given antenna, we first make use of the soft information available from the SISO channel decoder to reduce and hopefully to cancel the interfering signals arising from other transmit antennas. At the first iteration where this information is not available, we perform a classical MMSE filtering.

Let us consider the transmitted vector $\mathbf{x}_k = [x_k^1, \dots, x_k^{M_T}]^T$ at time k and assume that we are interested in the detection of its i -th symbol x_k^i . We start by evaluating the parameters \hat{x}_k^i and

$\sigma_{x_k^j}^2$ for the interfering symbols $x_k^j, j \neq i$, from the SISO decoder as follows:

$$\hat{x}_k^j = \mathbb{E}[x_k^j] = \sum_{j=1}^{2^B} x_k^j P[x_k^j] \quad (29)$$

$$\sigma_{x_k^j}^2 = \mathbb{E}[|x_k^j|^2] = \sum_{j=1}^{2^B} |x_k^j|^2 P[x_k^j] \quad (30)$$

where $P[x_k^j]$ is the probability of the transmission of x_k^j and is evaluated using the probabilities $P_{\text{dec}}(d_k^{j,n})$ at the decoder output:

$$P[x_k^j] = K \prod_{n=1}^B P_{\text{dec}}(d_k^{j,n}),$$

where K is a normalization factor. We further introduce the following definitions.

$\underline{\mathbf{H}}_i$ is the $(M_R \times (M_T - 1))$ matrix constructed from \mathbf{H} by discarding its i -th column, namely \mathbf{h}_i . We also define the $((M_T - 1) \times 1)$ vectors

$$\underline{\mathbf{x}}_k^i \triangleq [x_k^1, x_k^2, \dots, x_k^{i-1}, x_k^{i+1}, \dots, x_k^{M_T}]^T$$

and

$$\underline{\hat{\mathbf{x}}}_k^i \triangleq [\hat{x}_k^1, \hat{x}_k^2, \dots, \hat{x}_k^{i-1}, \hat{x}_k^{i+1}, \dots, \hat{x}_k^{M_T}]^T,$$

where \hat{x}_k^j are estimated in (29).

Now, given the received signal vector \mathbf{y}_k , a soft interference cancellation is performed on \mathbf{y}_k for detecting the symbol x_k^i by subtracting to \mathbf{y}_k the estimated signals of the other transmit antennas as Sadough, Khalighi & Duhamel (2009):

$$\underline{\mathbf{y}}_k^i = \mathbf{y}_k - \underline{\mathbf{H}}_i \underline{\hat{\mathbf{x}}}_k^i = \mathbf{h}_i x_k^i + \underline{\mathbf{H}}_i \underline{\mathbf{x}}_k^i - \underline{\mathbf{H}}_i \underline{\hat{\mathbf{x}}}_k^i + \mathbf{z}_k, \quad \text{for } i = 1, \dots, M_T. \quad (31)$$

Except under perfect prior information on the symbols which leads to $\hat{x}_k^j = x_k^j$, there remains a residual interference in $\underline{\mathbf{y}}_k^i$. In order to reduce further this interference, an instantaneous linear MMSE filter \mathbf{w}_k^i is applied to $\underline{\mathbf{y}}_k^i$ to minimize the mean square value of the error e_k^i defined as

$$e_k^i = x_k^i - r_k^i \quad (32)$$

where the filter output r_k^i is equal to

$$r_k^i = \mathbf{w}_k^i \underline{\mathbf{y}}_k^i. \quad (33)$$

Here, \mathbf{w}_k^i is obtained as

$$\mathbf{w}_k^i = \arg \min_{\mathbf{w}_k^i \in \mathbb{C}^{M_R}} \mathbb{E}_{\mathbf{x}_k, \mathbf{z}_k} [|x_k^i - \mathbf{w}_k^i \underline{\mathbf{y}}_k^i|^2]. \quad (34)$$

By invoking the orthogonality principle Scharf (1991), the coefficients of the MMSE filter \mathbf{w}_k^i are given by

$$\mathbf{w}_k^i = \mathbf{h}_i^\dagger \left[\mathbf{h}_i \mathbf{h}_i^\dagger + \frac{\underline{\mathbf{H}}_i (\Lambda_{k,i} - \tilde{\Lambda}_{k,i}) \underline{\mathbf{H}}_i^\dagger}{\sigma_{x_k^i}^2} + \frac{\sigma_z^2}{\sigma_{x_k^i}^2} \mathbb{I}_{M_R} \right]^{-1} \quad (35)$$

where

$$\Lambda_{k,i} = \mathbb{E}[\underline{\mathbf{x}}_k^i \underline{\mathbf{x}}_k^{i\dagger}] \approx \text{diag}\left(\mathbb{E}[|x_k^1|^2], \dots, \mathbb{E}[|x_k^{i-1}|^2], \mathbb{E}[|x_k^{i+1}|^2], \dots, \mathbb{E}[|x_k^{M_T}|^2]\right), \quad \text{and}$$

$$\tilde{\Lambda}_{k,i} = \hat{\underline{\mathbf{x}}}_k^i \hat{\underline{\mathbf{x}}}_k^{i\dagger} \approx \text{diag}\left(|\hat{x}_k^1|^2, \dots, |\hat{x}_k^{i-1}|^2, |\hat{x}_k^{i+1}|^2, \dots, |\hat{x}_k^{M_T}|^2\right).$$

Note that the off-diagonal entries in $\Lambda_{k,i}$ and $\tilde{\Lambda}_{k,i}$ have been neglected to reduce the complexity without causing significant performance loss Lee et al. (2006).

At the first decoding iteration, we have no prior information available on the transmitted data, i.e., $\Lambda_{k,i} = \sigma_{x_k}^2 \mathbb{I}_{M_T-1}$ and $\tilde{\Lambda}_{k,i} = \mathbf{0}_{M_T-1}$. Consequently, (35) reduces to

$$\mathbf{w}_k^i = \mathbf{h}_i^\dagger \left[\mathbf{H} \mathbf{H}^\dagger + \frac{\sigma_z^2}{\sigma_{x_k}^2} \mathbb{I}_{M_R} \right]^{-1} \quad (36)$$

which is no more than the linear MMSE detector for x_k^i .

Before passing the detected symbols r_k to the SISO decoder, we convert them to LLR. This is done assuming a Gaussian distribution for the residual interference after Soft-PIC detection (see Wang & Poor (1999) for details on the LLR conversion).

7.2 Improved Soft-PIC Detection Under Imperfect Channel Estimation

As we see from (31) and (35), we need the channel \mathbf{H} for both interference canceling and MMSE filtering. As the receiver has only an imperfect channel estimate $\hat{\mathbf{H}}$, the suboptimal *mismatched* solution consists in replacing $\underline{\mathbf{H}}_i$ and \mathbf{h}_i in (31) and (35) by their estimates $\hat{\underline{\mathbf{H}}}_i$ and $\hat{\mathbf{h}}_i$, respectively. As a first step toward a realistic design, we make use of the available channel estimate $\hat{\mathbf{H}}$ for interference cancellation. That is, equation (31) is rewritten as

$$\underline{\mathbf{y}}_k^i = \mathbf{y}_k - \hat{\underline{\mathbf{H}}}_i \hat{\underline{\mathbf{x}}}_k^i = \mathbf{h}_i x_k^i + \underline{\mathbf{H}}_i \underline{\mathbf{x}}_k^i - \hat{\underline{\mathbf{H}}}_i \hat{\underline{\mathbf{x}}}_k^i + \mathbf{z}_k, \quad \text{for } i = 1, \dots, M_T \quad (37)$$

where $\hat{\underline{\mathbf{H}}}_i$ is the $(M_R \times (M_T - 1))$ matrix constructed from $\hat{\mathbf{H}}$ by discarding its i -th column, namely $\hat{\mathbf{h}}_i$. We note that (37) naturally depends on the unknown channel matrix \mathbf{H} of which the receiver has only an imperfect estimate available. Instead of replacing the unknown channel by its estimate (i.e., the mismatched approach), we use the posterior distribution (12) and make two modifications to the detector described in Subsection 7.1, as follows (see Sadough, Khalighi & Duhamel (2009) and Sadough & Khalighi (2007) for more details).

The first modification concerns the design of the filter \mathbf{w}_k^i in (34). The modified filter $\tilde{\mathbf{w}}_k^i$ should minimize the average of the mean square error over all realizations of channel estimation errors. In other words,

$$\tilde{\mathbf{w}}_k^i = \arg \min_{\tilde{\mathbf{w}}_k^i \in \mathbb{C}^{M_R}} \mathbb{E}_{\mathbf{H}, \mathbf{x}_k, \mathbf{z}_k} \left[|x_k^i - \tilde{\mathbf{w}}_k^i \underline{\mathbf{y}}_k^i|^2 \middle| \hat{\mathbf{H}} \right] = \arg \min_{\tilde{\mathbf{w}} \in \mathbb{C}^{M_R}} \mathbb{E}_{\mathbf{H} | \hat{\mathbf{H}}} \left[\mathbb{E}_{\mathbf{x}_k, \mathbf{z}_k} \left[|x_k^i - \tilde{\mathbf{w}}_k^i \underline{\mathbf{y}}_k^i|^2 \right] \right] \quad (38)$$

where we have assumed the independence between \mathbf{H} , \mathbf{x}_k , and \mathbf{z}_k . After some simple algebraic manipulations Sadough, Khalighi & Duhamel (2009); Scharf (1991), we obtain:

$$\tilde{\mathbf{w}}_k^i = \bar{\mathbf{R}}_{x_k^i \underline{\mathbf{y}}_k^i} \bar{\mathbf{R}}_{\underline{\mathbf{y}}_k^i}^{-1} \quad (39)$$

where

$$\bar{\mathbf{R}}_{x_k^i \mathbf{y}_k^i} = \delta \sigma_{x_k^i}^2 \hat{\mathbf{h}}_i^\dagger + (\delta - 1) \mathbf{m}_{k,i} \hat{\mathbf{H}}_i^\dagger \quad (40)$$

with $\mathbf{m}_{k,i} = \hat{x}_k^i \hat{\mathbf{z}}_k^{i\dagger}$ and δ is given by (15), and

$$\begin{aligned} \bar{\mathbf{R}}_{\mathbf{y}_k^i} &= \delta^2 \sigma_{x_k^i}^2 \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^\dagger + \delta^2 \hat{\mathbf{H}}_i \Lambda_{k,i} \hat{\mathbf{H}}_i^\dagger + (\delta^2 - \delta) \hat{\mathbf{h}}_i \mathbf{m}_{k,i} \hat{\mathbf{H}}_i^\dagger + (\delta^2 - \delta) \hat{\mathbf{H}}_i \mathbf{m}_{k,i}^\dagger \hat{\mathbf{h}}_i^\dagger \\ &\quad + (1 - 2\delta) \hat{\mathbf{H}}_i \tilde{\Lambda}_{k,i} \hat{\mathbf{H}}_i^\dagger + (\sigma_z^2 + (1 - \delta) \sigma_{x_k^i}^2 + (1 - \delta) \text{tr}(\Lambda_{k,i})) \mathbb{I}_{M_R}. \end{aligned} \quad (41)$$

To get more insight on the proposed detector, let us consider the ideal case where perfect channel knowledge is available at the receiver, i.e., $\hat{\mathbf{H}} = \mathbf{H}$ and $\sigma_e^2 = 0$. We note that in this case, $\delta = 1$ and the posterior pdf (12) reduces to a Dirac delta function; consequently the two filters $\tilde{\mathbf{w}}_k^i$ and \mathbf{w}_k^i coincide. Similarly, under near-perfect CSIR, obtained either when $\sigma_e^2 \rightarrow 0$ or when $N_p \rightarrow \infty$, we have $\delta \rightarrow 1$, and the filter $\tilde{\mathbf{w}}_k^i$ gives a similar expression as \mathbf{w}_k^i in (35). However, in the presence of estimation errors, the proposed improved and mismatched detectors become different due to the inherent averaging in (38), which provides a robust design that adapts itself to the channel estimate available at the receiver.

The second modification concerns the application of the derived filter $\tilde{\mathbf{w}}_k^i$ to the received signal \mathbf{y}_k^i . As this latter depends on \mathbf{H} (see (37)), we average the filter output r_k^i as follows:

$$\tilde{r}_k^i = \mathbb{E}_{\mathbf{H}|\hat{\mathbf{H}}}[r_k^i] = \underbrace{\delta \tilde{\mathbf{w}}_k^i \hat{\mathbf{h}}_i x_k^i}_{\mu_{k,i}} + \underbrace{\delta \tilde{\mathbf{w}}_k^i \hat{\mathbf{H}}_i \hat{\mathbf{z}}_k^i - \tilde{\mathbf{w}}_k^i \hat{\mathbf{H}}_i \hat{\mathbf{z}}_k^i + \tilde{\mathbf{w}}_k^i \mathbf{z}_k}_{\eta_{k,i}} = \mu_{k,i} x_k^i + \eta_{k,i}, \quad (42)$$

where $\eta_{k,i}$ contains interference and noise. From (42) it is clear that the output of the improved MMSE filter can be viewed as an equivalent AWGN channel having x_k^i at its input. The parameters $\mu_{k,i}$ and $\sigma_{\eta_{k,i}}^2$ are calculated at each time-slot by using the symbols statistics. In order to transform the detected symbols at the output of the MMSE filter to LLRs on the corresponding bits, we approximate $\eta_{k,i}$ by a zero-mean Gaussian random variable with variance $\sigma_{\eta_{k,i}}^2$ (see Sadough, Khalighi & Duhamel (2009) for details on the calculation of this variance). Let $d_k^{i,m}$ denote the m -th ($m = 1, \dots, B$) bit corresponding to x_k^i . The LLR on $d_k^{i,m}$ is given by:

$$L(d_k^{i,m}) = \log \frac{P_{\text{dem}}(d_k^{i,m} = 1 | \tilde{r}_k^i, \mu_{k,i})}{P_{\text{dem}}(d_k^{i,m} = 0 | \tilde{r}_k^i, \mu_{k,i})} = \log \frac{\sum_{x_k^i \in \mathcal{S}_1^m} \exp \left\{ -\frac{|\tilde{r}_k^i - \mu_{k,i} x_k^i|^2}{\sigma_{\eta_{k,i}}^2} \right\} \prod_{n=1, n \neq m}^B P_{\text{dec}}^1(d_k^{i,n})}{\sum_{x_k^i \in \mathcal{S}_0^m} \exp \left\{ -\frac{|\tilde{r}_k^i - \mu_{k,i} x_k^i|^2}{\sigma_{\eta_{k,i}}^2} \right\} \prod_{n=1, n \neq m}^B P_{\text{dec}}^0(d_k^{i,n})}. \quad (43)$$

Note that here the cardinality of the sets \mathcal{S}_1^m and \mathcal{S}_0^m equals 2^{B-1} .

7.2.1 Case study

Figure 9 shows BER curves of the mismatched and improved receivers for the case of QPSK modulation and a (2×2) MIMO system. The number of channel uses for pilot transmission is $N_p \in \{2, 4, 8\}$. As a reference, we have also presented the BER curve for the case of perfect CSIR. We observe that the gain in SNR of the improved detector to attain the BER of 10^{-5} is about 1.4 dB, 0.5 dB, and 0.2 dB, respectively for $N_p = 2, 4$, and 8.

The interesting point is that the two detectors have almost the same convergence trend and the major improvement is obtained after the second iteration for both detectors Sadough, Khalighi & Duhamel (2009). So, if for the reasons of complexity reduction, we only process two receiver iterations, we still have a considerable performance gain by using the improved detector.

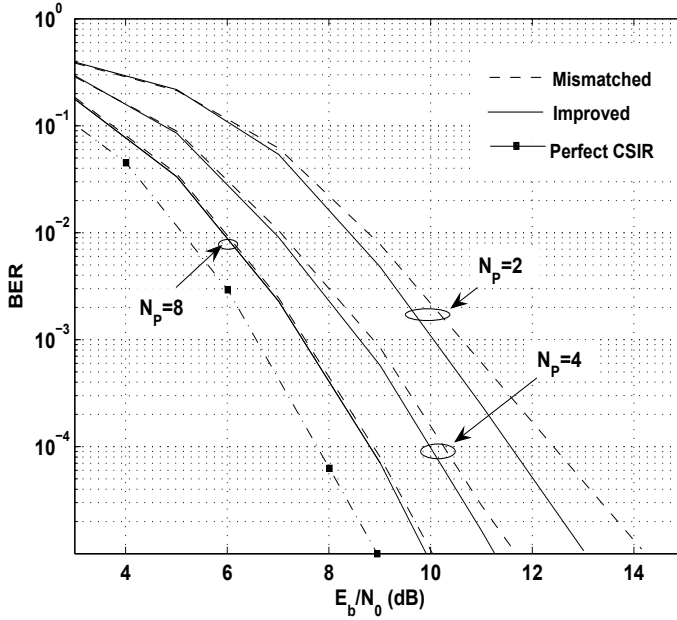


Fig. 9. BER performance of improved and mismatched iterative Soft-PIC; (2×2) MIMO with MUX ST scheme, i.i.d. Rayleigh block-fading channel with 4 fades per frame, QPSK modulation, training sequence length $N_p \in \{2, 4, 8\}$.

8. Conclusions

We studied in this chapter the interaction between iterative data detection and channel estimation in realistic wireless communication systems where the receiver disposes only of an *imperfect* estimate of the unknown channel parameters. To obtain the CSIR, we considered different recent and classically-used techniques. First, we presented pilot-only based channel estimation and showed that an accurate estimate of the channel through this method would require a large number of pilots per frame, which can result in a considerable loss in the system data throughput. Overlay pilots may be preferred to time-multiplexed solution from this point of view, however, the quality of channel estimate is, in general, worse, as compared to PSAM. We also presented semi-blind channel estimation methods that, in addition to pilot symbols, make use of the data symbols in the estimation process. Although iterative semi-blind channel estimation outperforms pilot-only assisted channel estimation, it has a higher complexity, which may be of critical concern for practical implementations.

Regardless of the channel estimation technique, an important point is the impact of estimation errors on the receiver performance. The usually-used approach is to consider the (imperfect) channel estimate as perfect and to use it in data detection. We called this the mismatched approach. In such case, we saw that, the impact of estimation errors is somehow similar for orthogonal and non-orthogonal space-time schemes. We then considered the improved approach by which we take into account the channel estimation inaccuracies in data detection. More precisely, by using the statistics of the channel estimation errors, we use a new detection rule instead of the sub-optimal mismatched detector. Applying this detection design rule to MAP and Soft-PIC detectors, we showed that a significant improvement can be obtained as compared to the mismatched detector. Finally, it is worth mentioning that adopting the improved reception scheme does not increase considerably the complexity. In fact, the improved detectors require just a few more matrix additions and multiplications, which does not have an important impact on the receiver complexity.

9. References

- Adireddy, S., Tong, L. & Viswanathan, H. (2002). Optimal placement of training for frequency-selective block-fading channels, *IEEE Trans. Inf. Theory* **48**(8): 2338–2353.
- Alamouti, S. (1998). A simple transmit diversity technique for wireless communications, *IEEE J. Sel. Areas Commun.* **16**(8): 1451–1458.
- Bahl, L., Cocke, J., Jelinek, F. & Raviv, J. (1974). Optimal decoding of linear codes for minimizing symbol error rate, *IEEE Trans. Inform. Theory* pp. 284–287.
- Balakrishnan, J., Rupp, M. & Viswanathan, H. (2000). Optimal channel training for multiple antenna systems, in *Proc. MMT*. Miami, FL.
- Belfiore, J.-C., Rekaya, G. & Viterbo, E. (2005). The golden code: a 2×2 full-rate space-time code with nonvanishing determinants, *IEEE Trans. Inform. Theory* **51**(4): 1432–1436.
- Berthet, A., Unal, B. & Visoz, R. (2001). Iterative decoding of convolutionally encoded signals over multipath Rayleigh fading channels, *IEEE J. Sel. Areas Commun.* **19**(9): 1729–1743.
- Bohlin, P. & Tapio, M. (2004). Optimized data aided training in MIMO systems, in *Proc. VTC* pp. 679–683. Milan, Italy.
- Cavers, J. K. (1991). An analysis of pilot symbol assisted modulation for Rayleigh fading channels, *IEEE Trans. Veh. Technol.* **40**(4): 686–693.
- Cui, T. & Tellambura, C. (2005). Superimposed pilot symbols for channel estimation in OFDM systems, in *Proc. Globecom* pp. 2229–2233. St. Louis, MO.
- de Carvalho, C. & Slock, D. T. M. (2001). *Semi-blind Methods for FIR Multi-channel Estimation*, in *Signal Processing Advances in Wireless and Mobile Communications*; Vol. I: Trends in Channel Estimation and Equalization, Editors: G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, Upper Saddle River, NJ: Prentice Hall.
- Dempster, A., Laird, N. & Rubin, D. (1977). Maximum-likelihood from incomplete data via the EM algorithm, *J. Royal Statistical Society* **39**(1): 1–38.
- Dong, M. & Tong, L. (2002). Optimal design and placement of pilot symbols for channel estimation, *IEEE Trans. Signal Processing* **50**(12): 3055–3069.
- Ghogho, M., McLernon, D., Alameda-Hernandez, E. & Swami, A. (2005). Channel estimation and symbol detection for block transmission using data-dependent superimposed training, *IEEE Signal Processing Letters* **12**(3): 226–229.
- Giannakis, G. B., Hua, Y., Stoica, P. & Tong, L. (2001). *Signal Processing Advances in Wireless and Mobile Communications*, Vol. 1, Trends in Channel Estimation and Equalization, Prentice Hall PTR, Upper Saddle River, NJ.

- Hassibi, B. & Hochwald, B. M. (2003). How much training is needed in multiple-antenna wireless links?, *IEEE Trans. Inform. Theory* **49**(4): 951–963.
- Hoeher, P. & Tufvesson, F. (1999). Channel estimation with superimposed pilot sequence, in *Proc. Globecom* **4**: 2162–2166. Rio de Janeiro, Brazil.
- Jungnickel, V., Haustein, T., Jorswieck, E., Pohl, V. & von Helmolt, C. (2001). Performance of a MIMO system with overlay pilots, in *Proc. Globecom* **1**: 594–598. San Antonio, TX.
- Khalighi, M. A., Berriche, L. & H  lard, J.-F. (2005). Overlaid or time-multiplexed pilots for channel estimation in iterative MIMO receivers, in *Proc. ISSPA* pp. 483–486. Sydney, Australia.
- Khalighi, M. A. & Bourennane, S. (2007). Semi-blind channel estimation based on superimposed pilots for single-carrier MIMO systems, in *Proc. VTC* pp. 1480–1484. Dublin, Ireland.
- Khalighi, M. A. & Bourennane, S. (2008). Semi-blind single-carrier MIMO channel estimation using overlay pilots, *IEEE Trans. Veh. Technol.* **57**(5): 1951–1956.
- Khalighi, M. A., Boutros, J. & H  lard, J.-F. (2006). Data-aided channel estimation for Turbo-PIC MIMO detectors, *IEEE Commun. Lett.* **10**(5): 350–352.
- Khalighi, M. A. & Boutros, J. J. (2006). Semi-blind channel estimation using EM algorithm in iterative MIMO APP detectors, *IEEE Trans. Wireless Commun.* **5**(11): 3165–3173.
- Khalighi, M. A., H  lard, J.-F., Sadough, S. M. S. & Bourennane, S. (2009). Suitable combination of channel coding and space-time schemes for moderate-to-high spectral efficiency mimo systems, *AE   International Journal of Electronics and Communications* . to appear.
- Lam, C., Falconer, D. & Danilo-Lemoine, F. (2008). Iterative frequency domain channel estimation for DFT-precoded OFDM systems using in-band pilots, *IEEE Journal on Selected Areas in Communication* **26**(2): 348–358.
- Lee, H., Lee, B. & Lee, I. (2006). Iterative detection and decoding with improved V-BLAST for MIMO-OFDM systems, *IEEE J. Selected Areas Commun.* **24**(3): 504–513.
- Ma, X., Giannakis, G. & Ohno, S. (2003). Optimal training for block transmission over doubly selective wireless fading channels, *IEEE Trans. Signal Processing* **51**(5): 1351–1366.
- Meng, X. & Tugnait, J. (2004). MIMO channel estimation using superimposed training, in *Proc. ICC* pp. 2663–2667. Paris, France.
- Moon, T. (1996). The expectation-maximization algorithm, *IEEE Signal Processing Mag.* **13**(6): 47–60.
- Moon, T. K. & Stirling, W. C. (2000). *Mathematical Methods and Algorithms for Signal Processing*, Upper Saddle River, NJ: Prentice Hall.
- Raoof, K., Khalighi, M. A. & Prayongpun, N. (2008). “MIMO Systems: Principles, Iterative Techniques and Advance Polarization” in *ADAPTIVE SIGNAL PROCESSING FOR WIRELESS COMMUNICATIONS*, CRC Press.
- De Carvalho, E. & Slock, D. T. M. (1997). Cramer-Rao bounds for semi-blind, blind, and training antenna arrays, a tutorial study, *International workshop on Signal Processing Advances in Wireless Communications (SPAWC)* pp. 129–132. Paris, France.
- Sadough, S. M. S. (2008). *Ultra Wideband OFDM Systems: Channel Estimation and Improved Detection Accounting for Estimation Inaccuracies*, PhD thesis, Universit   Paris-Sud 11.
- Sadough, S. M. S. & Duhamel, P. (2008). Improved iterative detection and achieved throughputs of OFDM systems under imperfect channel estimation, *IEEE Trans. Wireless Commun.* **7**(12): 5039 – 5050.

- Sadough, S. M. S., Ichir, M. M., Duhamel, P. & Jaffrot, E. (2009). Wavelet based semi-blind channel estimation for ultra wideband OFDM systems, *IEEE Trans. Veh. Technol.* **58**(3): 1302–1314.
- Sadough, S. M. S. & Khalighi, M. A. (2007). Optimal turbo-blast detection of MIMO-OFDM systems with imperfect channel estimation, *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*.
- Sadough, S. M. S., Khalighi, M. A. & Duhamel, P. (2009). Improved iterative MIMO signal detection accounting for channel estimation errors, *IEEE Trans. Veh. Technol.* . to appear.
- Sadough, S., Piantanida, P. & Duhamel, P. (2007). MIMO-OFDM optimal decoding and achievable information rates under imperfect channel estimation, in *Proc. Signal Process. Advances Wireless Commun. (SPAWC)*.
- Scharf, L. (1991). *Statistical Signal Processing*, Addison-Wesley.
- Sellathurai, M. & Haykin, S. (2002). Turbo-BLAST for wireless communications: theory and experiments, *IEEE Trans. Signal Process.* **50**(10): 2538– 2546.
- Tapio, M. & Bohlin, P. (2004). A capacity comparison between time-multiplexed and superimposed pilots, in *Proc. Asilomar Conf. on Signals, Systems and Computers* pp. 1049–1053. Pacific Grove, CA.
- Taricco, G. & Biglieri, E. (2005). Space-time decoding with imperfect channel estimation, *IEEE Trans. Wireless Commun.* **4**(4): 1874–1888.
- Tarokh, V., Naguib, A., Seshadri, N. & Calderbank, A. R. (1999). Space-time codes for high data rate wireless communication: Performance criteria in the presence of channel estimation errors, mobility, and multiple paths, *IEEE Trans. Commun.* **47**: 199–207.
- Telatar, E. (1999). Capacity of multi-antenna Gaussian channels, *European Trans. Telecommun.* **10**(6): 585–595.
- Visoz, R. & Berthet, A. (2003). Iterative decoding and channel estimation for space-time BICM over MIMO block fading multipath AWGN channel, *IEEE Trans. Commun.* **51**(8): 1358– 1367.
- Wang, X. & Poor, H. V. (1999). Iterative (turbo) soft interference cancellation and decoding for coded CDMA, *IEEE Trans. Commun.* **47**(7): 1046–1061.
- Zhu, H., Farhang-Boroujeny, B. & Schlegel, C. (2003). Pilot embedding for joint channel estimation and data detection in MIMO communication systems, *IEEE Commun. Lett.* **7**(1): 30–32.

Cooperative MIMO Systems in Wireless Sensor Networks

M. Riduan Ahmad¹, Eryk Dutkiewicz², Xiaojing Huang³ and M. Kadim Suaidi⁴

^{1,4}*Universiti Teknikal Malaysia Melaka*, ²*Macquarie University*, ³*CSIRO ICT Centre*

^{1,4}*Malaysia*, ^{2,3}*Australia*

1. Introduction

The Multiple-Input Multiple-Output (MIMO) term originally describes the use of the multiple antennas concept or exploitation of spatial diversity techniques. In early research work, the MIMO concept was proposed to fulfil the demand for providing reliable high-speed wireless communication links in harsh environments. Subsequently, MIMO technology has been proposed to be used in wireless local area networks and cellular networks, particularly at the base station and access point sides to tackle the challenges of low transmission rates and low reliability with no constraints on energy efficiency. In contrast, Wireless Sensor Networks (WSNs) have to deal with energy constraints due to the fact that each sensor node depends on its battery for its operation. In harsh environments, sensor nodes must be provided with reliable communication links. However, current WSN design requirements do not require high transmission rates.

The concept of cooperative MIMO was introduced in WSNs by utilizing the collaborative nature of dense sensor nodes with the broadcast wireless medium to provide reliable communication links in order to reduce the total energy consumption for each sensor node. Therefore, instead of using multiple antennas attached to one node or device such in the traditional MIMO concept, cooperative MIMO presents the concept of multiple sensor nodes cooperating to transmit and/or receive signals. Multiple sensor nodes are physically grouped together to cooperatively transmit and/or receive. Within a group, sensor nodes can communicate with relatively low power as compared to inter-group communication. Furthermore, by using this cooperative MIMO concept, we can provide the advantages of traditional MIMO systems to WSNs, particularly in terms of energy efficient operation.

This chapter introduces the concepts of diversity techniques and their relationship with cooperative MIMO systems and to discuss their practicality for implementation in WSNs. The approaches that we are going to use in this chapter are the comparative study and performance evaluation of major diversity techniques and their implementations in cooperative MIMO systems. The outcomes can be used as the basis for further study in order to find the most suitable cooperative MIMO scheme to be implemented in the WSN environment.

The rest of the chapter is organized as follows. Section 2 introduces the concept of cooperative MIMO and Section 3 explains various types of diversity techniques proposed in the current literature. A comparative study between the major diversity techniques is presented in Section 4 and this is followed by performance evaluation in Section 5. Finally, in Section 6 we conclude the chapter.

2. Cooperative MIMO Concepts

The MIMO term originally describes the use of the multiple antennas concept or exploitation of spatial diversity techniques. In early research work, the MIMO concept was proposed to fulfil the demand for providing reliable high-speed wireless communication links in harsh environments. Subsequently, MIMO technology has been proposed to be used in wireless local area networks and cellular networks (Proakis, 2001), particularly at the base station and access point sides to tackle the challenges of low transmission rates and low reliability with no constraints on energy efficiency. In contrast, WSNs have to deal with energy constraints due to the fact that each sensor node depends on its battery for its operation. In harsh environments, sensor nodes must be provided with reliable communication links. However, current WSN design requirements do not require high transmission rates.

The concept of cooperative MIMO was introduced in WSNs by utilising the collaborative nature of dense sensor nodes with the broadcast wireless medium to provide reliable communication links in order to reduce the total energy consumption for each sensor node. Therefore, instead of using multiple antennas attached to one node or device such in the traditional MIMO concept, cooperative MIMO presents the concept of multiple sensor nodes cooperating to transmit and/or receive signals. Multiple sensor nodes are physically grouped together to cooperatively transmit and/or receive. Within a group, sensor nodes can communicate with relatively low power as compared to inter-group communication (Singh and Prasanna, 2003; Gupta and Younis, 2003; Yuksel and Erkip, 2004). Furthermore, by using this cooperative MIMO concept, we can provide the advantages of traditional MIMO systems to WSNs, particularly in terms of energy efficient operation.

3. Techniques of Diversity

In this section we discuss various diversity techniques to reduce the deep fading problem in WSNs which requires higher retransmission rates. By tackling this problem, we are clearly satisfying two design requirements: energy efficient operation and higher communication link reliability. It is important to observe that deep fading contributes to packet errors (if a portion of the packet is affected) or packet loss (if the whole packet is totally lost which can be common as data packets in WSNs are normally small (Kohvakka et. al., 2006; Karl and Willig, 2007)). The basic concept of diversity is to provide the receiver with copies of independently faded transmitted packets with the hope that at least one of these copies will be received correctly. Diversity can be implemented in various ways such as frequency diversity, spatial diversity, time diversity, modulation diversity and polarisation diversity to suit different design requirements.

Frequency diversity is achieved when the same signal is transmitted over different frequency bands. The separation of the frequency bands has to be more than the coherence

bandwidth of the channel (Duman and Ghrayeb, 2007). Time diversity is achieved when the same signal is transmitted with redundancy using different time intervals. The separation of the time intervals has to be more than the coherence time of the channel. Also, time diversity can be achieved by means of channel coding. The idea is to transmit the different parts of the codeword corresponding to a particular symbol using different time intervals. The most practical channel codes discussed in the literature include block codes, convolutional codes and trellis-coded modulation (Duman and Ghrayeb, 2007).

Spatial diversity is achieved by the use of multiple antennas or nodes at either end or at both ends of the MIMO communication link. The separation between the antennas or nodes has to be more than half a wavelength in a uniform scattering environment. Systems with multiple antennas are also referred to as MIMO systems (Duman and Ghrayeb, 2007). Therefore we can refer to systems with multiple nodes as cooperative MIMO systems. Spatial diversity gains increase channel capacity which leads to higher data throughputs and significant improvement in data transmission reliability. These advantages are achieved without any expansion of bandwidth or higher transmit power which makes this technique very suitable to be implemented in energy constrained WSNs.

Diversity techniques can be combined to achieve greater improvements in reliability and achievable transmission rates. Perhaps the most popular combination technique is between space diversity and time diversity by using channel coding. The combination yields the space-time coding (STC) scheme. The variants of the STC scheme depend on the channel coding being used. For example, space-time block coding (STBC) schemes are based on block coding and space-time trellis coding scheme (STTC) schemes are based on trellis-coded modulation.

Multiple antennas or nodes can be exploited in different ways at both ends of the MIMO communication link. Early work in this area concerned designs of multiple antennas at the receiver side to achieve receive spatial diversity in order to boost link reliability as the number of receiving antennas grows. Among the earliest users of receive spatial diversity schemes are mobile communications systems to improve uplink performance by implementing multiple receive antennas at the base station (Proakis, 2001). If only a single transmit antenna and multiple receive antennas are used, the resulting system is referred to as a Single-Input Multiple-Output (SIMO) system.

In later work transmit spatial diversity was achieved by exploiting multiple transmit antennas with proper coding or weighting of the transmitted data signals. It is important to note that both spatial diversity schemes achieve improved transmission reliability at the cost of transmission rates comparable to the Single-Input Single-Output (SISO) systems. Clearly the achievement of higher link reliability is a trade-off with transmission rates. When multiple transmit antennas and single receive antenna are used, the resulting system is referred to as a Multiple-Input Single-Output (MISO) system.

Further research to achieve higher transmission rates and higher capacity has been done using multiple antennas at both ends of the communications link. These multiple transmit antennas and multiple receive antenna systems are referred to as MIMO systems. One of the common techniques to boost the transmission rates is to provide the receivers with independent streams of the same data signal from different transmit antennas. In this way, the transmit antennas are exploited to boost the transmission rates at the cost of lower link reliability. However, when operating under certain constraints, the same scheme can achieve full diversity gain leading to higher link reliability.

A comparison between the different spatial diversity schemes discussed above is shown in Table 1 where M and N denote the number of transmit antennas and receive antennas, respectively.

Scheme	M	N	Example	Benefits
SISO	1	1	No transmit or receive diversity	No diversity gain.
SIMO	1	> 1	Receive diversity	Diversity proportional to N .
MISO	> 1	1	Transmit diversity	Diversity proportional to M .
MIMO	> 1	> 1	Use of multiple antennas at both the transmitter and receiver	Diversity proportional to the product of M and N . Array gain (coherent combining assuming prior channel estimation).

Table 1. Comparison of main spatial diversity schemes

4. Multiple Nodes in Wireless Sensor Networks

A major design requirement of WSNs is to reduce the total energy consumption of the sensor nodes. The transmission power can be reduced by providing the highest diversity gain possible which leads to higher link reliability and thus lower the retransmission rates. Therefore the exploitation of multiple nodes in WSNs (referred to as cooperative MIMO) is inevitable in order to provide higher reliability communication link and reduce transmission power.

Most of the previous work in the area of cooperative MIMO has assumed that the cooperating sensor nodes are perfectly synchronised during transmission and reception (Jagannathan et. al., 2004). Recently, the impact of imperfect synchronisation effects on the performance of cooperative MIMO operation in WSNs has gained more attention (Li, 2004; Li et. al., 2004; Li and Hwu, 2006). Imperfect synchronisation could occur due to the lack of carrier synchronisation or because of imperfect timing in frame and bit level synchronisation. In this chapter, we consider the impact of imperfect synchronisation caused by clock jitter alone. Each cooperating transmit node from a set of M cooperative nodes experiences clock jitter with the jitter around a reference clock, T_o denoted as T_j^m where $1 \leq m \leq M$. The detailed system model of clock jitter will be explained in Section 5.

The following discussion explains in detail the three major MIMO schemes in both synchronous and asynchronous scenarios and their practicality in WSNs. Synchronous operation assumes perfect synchronisation between cooperating transmitting nodes and asynchronous operation refers to scenarios where imperfect synchronisation occurs. The three MIMO schemes are:

- a) SIMO System
- b) MISO System
- c) Spatial Multiplexing MIMO System

4.1 SIMO System

Perhaps the first technique in diversity particularly related to spatial utilisation is the receive diversity (SIMO) technique. The transmitter can choose to perform frequency, time, or polarisation due to the fact that the source of diversity does not affect the method of combination at the receiver side (with the exception of transmit spatial diversity (Jafarkhani, 2005)). At the receiver side, more than one antenna or node must be used, $N \geq 2$, to gain spatial diversity which leads to higher reliability by increasing the average signal-to-noise ratio (SNR) and lowering the bit error rate (BER).

There are four popular combining methods that are utilised at the receiver: Maximum Ratio Combining (MRC), Equal Gain Combining (EGC), Selection Combining (SC) and Switched Combining (Simon and Alouini, 2000; Rappaport, 2002; Jafarkhani, 2005; Duman and Ghayeb, 2007). MRC achieves diversity gain equal to the number of the receive antennas, N , with N Radio Frequency (RF) chains as shown in Figure 1. EGC is a special case of MRC with equal weights' amplitudes where all the received signals are co-phased and then combined together with equal gain. The EGC receiver's circuit is less complex but at the cost of lower diversity gain than for MRC.

Assume that the receiver receives N replicas of the transmitted signal, s through N independent paths. The k^{th} received signal is defined by:

$$r_k = h_k s + \eta_k \quad (1)$$

where $k = 1, 2, \dots, N$, η_k is the complex white Gaussian noise sample vector added to the k^{th} copy of the signal with zero mean and σ^2 variance, $\eta_k \sim N_c(0, \sigma^2)$ and h_k is the complex channel fading gain vector with zero mean and ρ^2 variance, $h_k \sim N_c(0, \rho^2)$. We assume that the receiver is coherent where the channel information is known and perfectly estimated at the receiver. If the average power of the transmitted symbol is $E[|s|^2]$, the instantaneous SNR of the k^{th} receiver is given by (Jafarkhani, 2005):

$$\gamma_k = |h_k|^2 \cdot \frac{E[|s|^2]}{\sigma^2}. \quad (2)$$

The attempt to recover s can be given by the following MRC linear combination:

$$\tilde{s} = \sum_{k=1}^N h_k^* r_k = \sum_{k=1}^N |h_k|^2 s + \sum_{k=1}^N \eta_k h_k^* \quad (3)$$

where \tilde{S} is the resulting decision variable with S mean and $\frac{\sigma^2}{\sum_{k=1}^N |h_k|^2}$ variance which can

be represented as $\tilde{S} \sim \left(S, \frac{\sigma^2}{\sum_{k=1}^N |h_k|^2} \right)$. The resulting effective SNR at the output of the MRC

block is proportional to $\sum_{k=1}^N |h_k|^2$ and given as:

$$\gamma = \sum_{k=1}^N |h_k|^2 \cdot \frac{E[|S|^2]}{\sigma^2} = \sum_{k=1}^N \gamma_k. \quad (4)$$

From Equation (4), the effective SNR of the system with a receive diversity scheme is equivalent to the sum of the instantaneous receive SNRs for N different paths. If we assume that all the different paths have the same average SNR, then the average of the effective SNR at the output of the MRC block is:

$$\bar{\gamma} = \sum_{k=1}^N E[|h_k|^2] \cdot \frac{E[|S|^2]}{\sigma^2} = \sum_{k=1}^N E[\gamma_k] = N \cdot \bar{\gamma}_k. \quad (5)$$

By increasing the number of receiving antennas N , the receive average SNR can be increased by N -fold which leads to the lowest possible BER for the system such that at the high SNRs regime, the error probability decays as SNR_a^{-N} (Proakis, 2001). On the other hand, when N increases, the receiver becomes more complex and larger in size. It seems that we have to trade off the cost, size and complexity of the devices or nodes for higher link reliability.

Selection combining was introduced to reduce the N^{th} RF chains complexity with only one RF chain used where the receiver performs signal selection with the highest SNR for decoding as shown in Figure 2. Also, channel state information is not needed which means that selection combining can be used for both coherent and non-coherent receivers (Simon and Alouni, 2000). The average SNR at the output of the selection combiner is given as:

$$\bar{\gamma} = \bar{\gamma}_k \sum_{k=1}^N \frac{1}{k}. \quad (6)$$

As can be seen from Equation (6), selection combining does not achieve the full N diversity gain which clearly trades off the diversity gain for lower complexity.

Later, a hybrid selection/MRC technique was proposed to balance the requirements between higher diversity gain and lower complexity (Jafarkhani, 2005).

Switched combining employs scanning and selection operation where the receiver scans all the diversity branches and selects a particular branch with the SNR above a certain predetermined threshold (Jafarkhani, 2005). The signals from the selected branch are selected as the output until its SNR drops below the threshold. Then the receiver starts to scan again and switches to another branch. This scheme is simpler since it does not require any channel knowledge but at the cost of lower achievable diversity gain.

In the context of practicality in WSNs, a SIMO with MRC scheme is more practical and promising for implementation as shown in Figure 3. This is due to the fact that each node in the network represents a single path processing including the RF chain processing. It seems that the complexity issue in the traditional SIMO approach can be reduced with the cooperative SIMO implementation while providing the highest SNR possible. Moreover, the transmission by the transmit node can be done without the need for time synchronisation, thus the cooperative SIMO system is not affected by clock jitter.

On the other hand, there are other issues that we have to consider such as the fact that data signals received by all the receiving nodes must be forwarded to a common destination node in order to combine and decode them successfully. Moreover, the diversity gain does not contribute to the reduction of the total transmission power and the use of N receiving nodes can contribute to the higher circuit power in the network.

4.2 MISO System

The main motivation for using of multiple antennas at the transmitter is to reduce the required processing power and complexity at the receiver which leads to lower power consumption, lower size and lower cost. However, the MISO concept is not easy to exploit and to implement (Naquib and Calderbank, 2000). Additional signal processing is required at both the transmitters and receiver in order to correctly decode the received signals. Also, another challenge is that the transmitter does not know the channel conditions unless channel information is fed back by the receiver to the transmitter (Liu et. al., 2001).

A number of MISO schemes have been proposed in the literature and can be categorised into two major classes:

- a) Closed-loop MISO schemes with feedback
- b) Open-loop MISO schemes without feedback

The difference between the two types of schemes is that the former relies on channel state information which has to be fed back to the transmitter and the latter eliminate the need for channel state information at the transmitter.

4.2.1 Closed-loop MISO System

The modulated signals are weighted with different weighting factors and transmitted with multiple antennas M at the transmitter. The weighting factors are chosen with the assistance of the channel state information so that the received SNR can be maximised at the receiver. The weighting factors must be optimised in order to achieve full diversity gain. One of the drawbacks of this system is that when the weighting factors are not optimised due to imperfect channel estimation, the received SNR is decreased.

Two of the most popular closed-loop MISO schemes are switched diversity (Winters, 1983) and digital beamforming (Litva and Lo, 1996). Among the two schemes, the best solution, if the transmitter has perfect knowledge of the channel, is the MISO beamforming scheme. The MISO beamforming scheme is less complex and easier to deploy which makes it more practical to implement in WSNs.

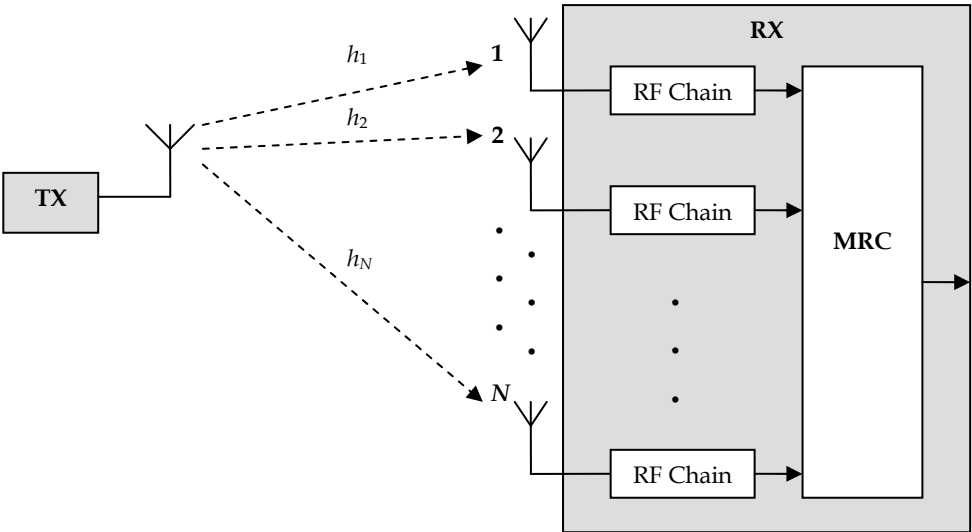


Fig. 1. A system with one transmit antenna and N receive antennas with MRC.

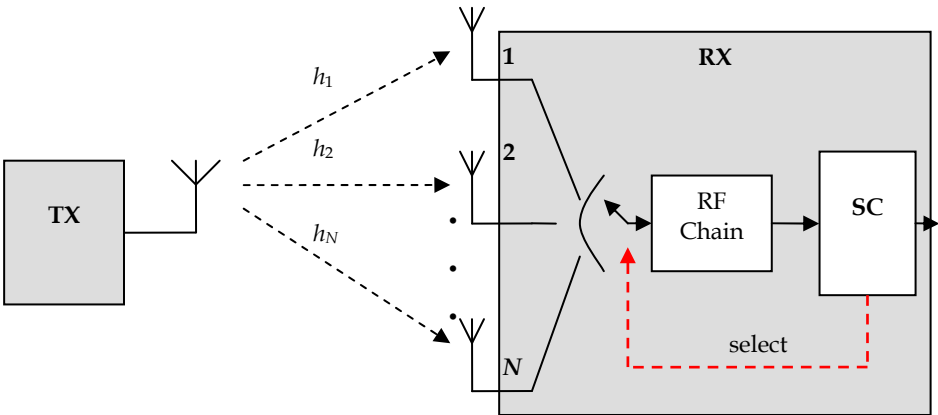


Fig. 2. A system with one transmit antenna and N receive antennas with SC.

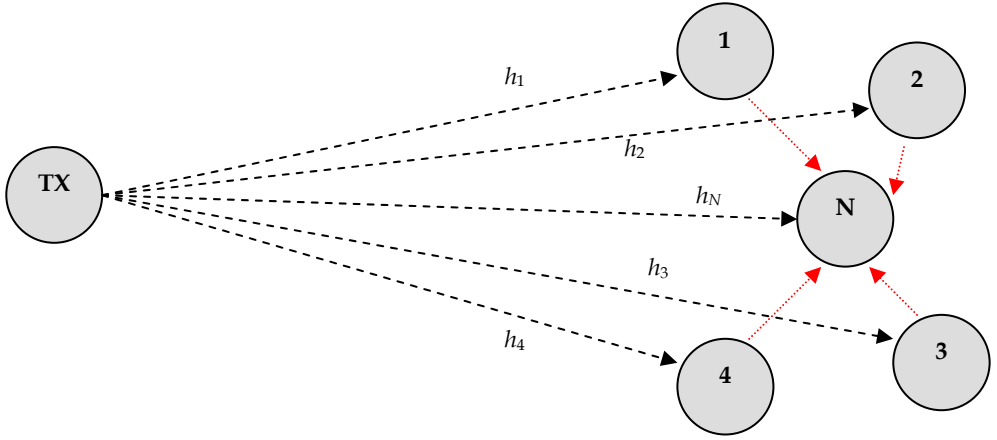


Fig. 3. A cooperative receive diversity system with one transmit node and N receive nodes.

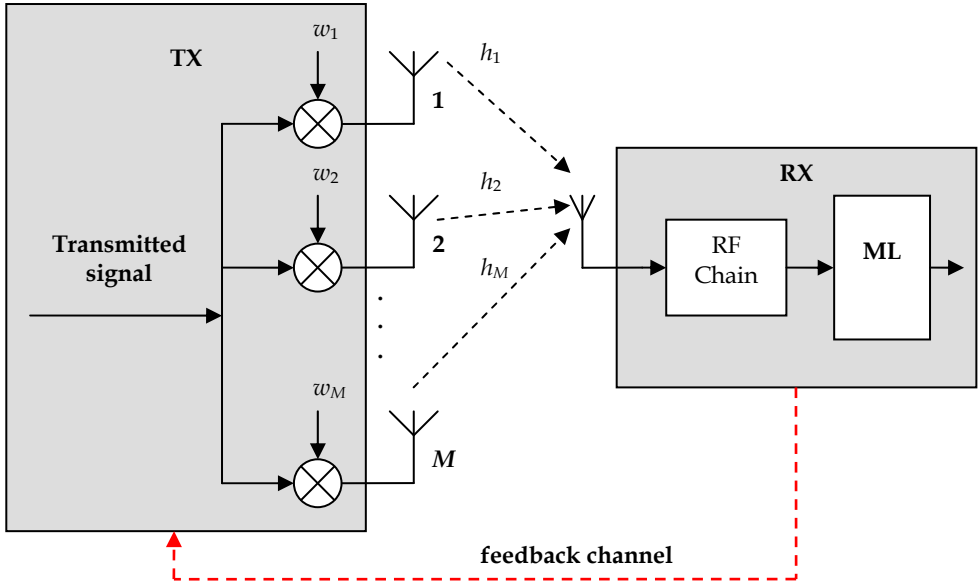


Fig. 4. A beamforming transmit diversity system with M transmit antennas and 1 receive antenna.

Let us consider a digital MISO beamforming system with M antennas at the transmitter and one antenna at the receiver as shown in Figure 4. The transmitter transmits M weighted transmitted signals, $w_k s$ through M independent paths. The received signal is then given as:

$$r = \sum_{k=1}^M h_k w_k s + \eta \quad (7)$$

where $k = 1, 2, \dots, M$, η is the complex white Gaussian noise sample added to the received signal with zero mean and σ^2 variance, $\eta \sim N_c(0, \sigma^2)$ and h_k is the complex channel fading gain vector with zero mean and ρ^2 variance, $h_k \sim N_c(0, \rho^2)$. We assume that the receiver is coherent where the channel information is known and perfectly estimated at the receiver.

The decision variable \tilde{S} for the Maximum Likelihood (ML) detector has S mean and $\frac{\sigma^2}{\sum_{k=1}^M h_k w_k}$ variance. If the average power of the transmitted symbol is $E[|s|^2]$, the

instantaneous effective SNR is then given as:

$$\gamma = \sum_{k=1}^M h_k w_k \cdot \frac{E[|s|^2]}{\sigma^2}. \quad (8)$$

Given the transmitter knows the channel perfectly through feedback from the receiver, the weights are scaled and optimised proportionally to h_k^* so as to maximise the SNR in \hat{S} . The resulting instantaneous SNR is proportional to $\sum_{k=1}^M |h_k|^2$ and given as:

$$\gamma = \sum_{k=1}^M |h_k|^2 \cdot \frac{E[|s|^2]}{\sigma^2}. \quad (9)$$

Let $\gamma_k = |h_k|^2 \cdot \frac{E[|s|^2]}{\sigma^2}$, then the effective receive SNR can be written as:

$$\gamma = \sum_{k=1}^M |h_k|^2 \cdot \frac{E[|s|^2]}{\sigma^2} = \sum_{k=1}^M \gamma_k. \quad (10)$$

From Equation (10), the effective SNR of the system with the transmit MISO beamforming diversity scheme is equivalent to the sum of the instantaneous receive SNRs for M different paths. If we assume that all the different paths have the same average SNR, then the average of the effective SNR at the output of the ML block is:

$$\bar{\gamma} = \sum_{k=1}^M E[|h_k|^2] \cdot \frac{E[|s|^2]}{\sigma^2} = \sum_{k=1}^M E[\gamma_k] = M \cdot \bar{\gamma}_k. \quad (11)$$

As M grows, the receive average SNR is increased by M -fold which leads to the lowest possible BER for the system. Furthermore, the total radiated power for all M antennas is the same as the total transmission power of one single transmit antenna, as in the receive diversity cases (Larsson and Stoica, 2003). It is clear that the transmission power P_t has been reduced down to P_t/M as the diversity gain increases up to M .

In the context of practicality in WSNs, the main obstacle for MISO beamforming implementation is the issue of how to provide each transmitting node with the knowledge of the channel. A multi-channel approach can be used where one channel is dedicated for the feedback and the other channel for data transmission. The channel is estimated periodically through training sequences on the feedback channel. However, a multi-channel approach is not practical for WSNs because such an approach increases the hardware and processing complexity at both the transmitter and the receiver. Also, such an approach requires tight frequency synchronisation to maintain the dual channel utilisation which obviously increases the total energy consumption of the network.

A better and practical alternative approach is to exploit the existing control protocols in WSNs such as those utilising RTS-CTS packets to provide the channel state information to the transmitter as shown in Figure 5. Both the control and data communications can be maintained over a single-channel with less complexity and loose synchronisation. Moreover, the transmission power of each transmitting node is reduced down to P_t/M which leads to the reduction of the total power consumption in the network. In addition, the RTS-CTS implementation can also reduce the hidden node problem in such densely distributed sensor networks.

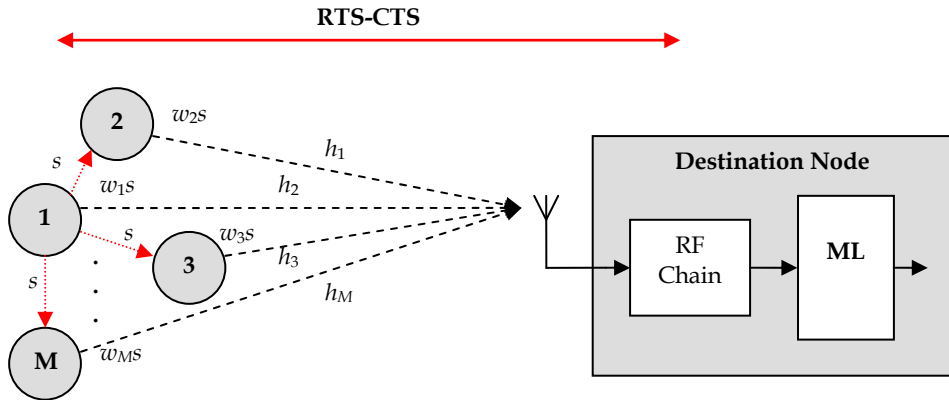


Fig. 5. A cooperative beamforming transmit diversity system with M transmit nodes and 1 destination.

4.2.2 Open-loop MISO System

The modulated signals must be processed at the transmitter first before being transmitted from multiple antennas M . The main motivation is to reduce the complexity of the feedback schemes in the closed-loop MISO systems. The transmitter design is enhanced with more advanced signal processing and/or a combination of various diversity techniques in order to provide the receiver with the capability to exploit full diversity gain from the received signals.

One of the early proposed open-loop MISO schemes is the antenna hopping scheme (Seshadri and Winters, 1993; Wittneben, 1991). The modulated signals are transmitted from M antennas with different time intervals. At the receiver, the delayed signals introduce a multipath-like distortion for the intended signal.

The multipath-like distortion can be resolved at the receiver by using a ML detector or a Minimum Mean Square Error (MMSE) detector to obtain M diversity gain. The antenna hopping scheme has been shown to achieve fully diversity gain up to M without any bandwidth expansion but at the cost of a lower spatial rate.

In order to gain the full spatial rate, the diversity gain achieved from a multiple antennas implementation is combined with the coding gain achieved from the error control and channel coding schemes. The combination schemes of error control coding and multiple antennas have gained the full spatial rate in addition to the diversity benefit but at the cost of bandwidth losses due to code redundancy (Vucetic and Yuan, 2003). A better and practical approach is a joint design of multiple antennas with channel coding schemes. This approach can be achieved when the multiple antennas and channel coding schemes are designed as a single signal processing module. Coding techniques for multiple antenna communications are called space-time coding (STC). STC schemes provide redundant transmission in both spatial and temporal domains. In addition to the diversity gain, full spatial rate and no bandwidth expansion advantages, STC schemes can be combined with multiple receive antennas to achieve capacity gain.

The most popular STC scheme is due to Alamouti (Alamouti, 1998) who studied the case of two transmit antennas. The Alamouti space-time encoder picks up two symbols s_1 and s_2 from an arbitrary constellation and the two symbols are transmitted in two consecutive time slots as shown in Figure 6. In the first time slot, s_1 is transmitted from the first antenna while s_2 is transmitted from the second antenna. Consecutively in the second time slot, $-s_2^*$ is transmitted from the first antenna while s_1^* is transmitted from the second antenna. Since both the symbols are transmitted in two time slots, the overall rate is given as one symbol per channel use. The key concept of the Alamouti STC scheme is the orthogonal design of the transmit sequences. The inner product of the sequences x_1 and x_2 is given as:

$$x_1 \cdot x_2 = s_1 s_2^* - s_1 s_2^* = 0. \quad (12)$$

The transmitted code matrix has the following property:

$$X \cdot X^H = \begin{bmatrix} s_1 & -s_2^* \\ s_2 & s_1^* \end{bmatrix} \cdot \begin{bmatrix} s_1^* & s_2^* \\ -s_2 & s_1 \end{bmatrix} = \begin{bmatrix} |s_1|^2 + |s_2|^2 & 0 \\ 0 & |s_1|^2 + |s_2|^2 \end{bmatrix}. \quad (13)$$

Assume that both the paths experience quasi-static fading where the fading coefficients are constant across the two consecutive symbol transmission intervals which can be expressed as:

$$\begin{aligned} h_1(t) &= h_1(t+T) = h_1 = |h_1|e^{j\theta_1} \\ h_2(t) &= h_2(t+T) = h_2 = |h_2|e^{j\theta_2} \end{aligned} \quad (14)$$

where $|h_k|$ and θ_k , $k = 1, 2$, are the amplitude gain and phase shift for the path from transmit antenna k to the receiver antenna and T is the symbol duration. The received signal in the first time slot is given as:

$$r(t) = r_1 = h_1 s_1 + h_2 s_2 + \eta_1 \quad (15)$$

and in the second time slot, the received signal is given as:

$$r(t+T) = r_2 = -h_1 s_2^* + h_2 s_1^* + \eta_2 \quad (16)$$

where η_1 and η_2 are the complex white noise with zero mean and variance σ^2 for the first time slot and second time slot, respectively. The received signal vector is defined at the receiver as:

$$r = \begin{bmatrix} r_1 \\ r_2^* \end{bmatrix} \quad (17)$$

which can be written as:

$$r = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2^* \end{bmatrix}. \quad (18)$$

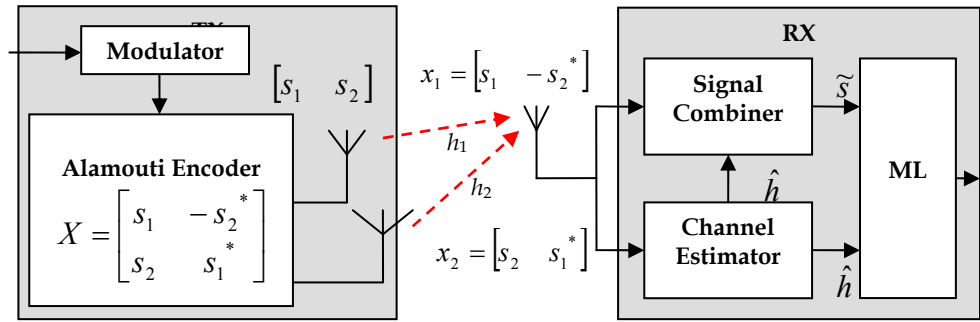


Fig. 6. An alamouti STC transmit diversity system with 2 transmit antennas and 1 receive antenna.

Assume that the receiver is coherent and optimal. Then the attempt to recover s_1 and s_2 can be given by the following linear combination:

$$\tilde{s} = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix}^H \begin{bmatrix} r_1 \\ r_2^* \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^M |h_k|^2 s_1 + h_1^* \eta_1 + h_2 \eta_2^* \\ \sum_{k=1}^M |h_k|^2 s_2 + h_2^* \eta_1 - h_1 \eta_2^* \end{bmatrix}. \quad (19)$$

The resulting decision variables in Equation (19) are equivalent to the one obtained with receive diversity using the MRC scheme. The only difference is the phase rotations on the noise components which do not degrade the effective SNR (Alamouti, 1998).

The decision variable vector \tilde{S} with S mean and $\frac{\sigma^2}{\sum_{k=1}^M h_k}$ variance is sent to the ML

detector. If the average power of the transmitted symbols is $E[|s_n|^2]$, the receive SNR in each sub-channel is given as:

$$\gamma = \sum_{k=1}^M h_k \cdot \frac{E[|s_n|^2]}{2\sigma^2}. \quad (20)$$

We can observe from Equation (20) that the linear processing in Equation (19) transforms the space-time channel into two parallel and independent scalar channels. If we assume that the symbols are Phase Shift Keying (PSK) modulated signals with equal energy constellations, the total transmission power is effectively doubled as shown in Equation (20) compared to

the SIMO MRC and MISO beamforming schemes. Let $\gamma_k = |h_k|^2 \cdot \frac{E[|s_n|^2]}{2\sigma^2}$, then the effective receive SNR can be written as:

$$\gamma = \sum_{k=1}^M |h_k|^2 \cdot \frac{E[|s_n|^2]}{2\sigma^2} = \sum_{k=1}^M \gamma_k. \quad (21)$$

If we assume that all the different paths have the same average SNR, then the average of the effective SNR at the output of the ML block is:

$$\bar{\gamma} = \sum_{k=1}^M E[|h_k|^2] \cdot \frac{E[|s_n|^2]}{2\sigma^2} = \sum_{k=1}^M E[\gamma_k] = M \cdot \bar{\gamma}_k. \quad (22)$$

As we can observe, the MISO Alamouti STC scheme provides the same diversity gain as the SIMO MRC and MISO beamforming schemes with M equal to two but with 3 dB loss in error performance (Larsson and Stoica, 2003). In addition, the MISO Alamouti STC scheme can be applied for a system with 2 transmitting antennas and N receiving antennas to gain higher capacity. Although such systems are very important for high-speed networks, careful consideration of circuit and processing energies and decoder complexity at the receiver in WSNs keeps our discussion to systems with only one receive antenna which corresponds to one receive node in cooperative MISO transmission.

The Alamouti STC scheme can be generalised from two transmit antennas up to M transmit antennas by using the same theory of orthogonal design (Tarokh et. al., 1999). The generalised scheme is referred to as Orthogonal Space-Time Block Codes (OSTBC). In general, OSTBC can be categorised into two types: real and complex, based on the signal constellation.

The basic operation of OSTBC is shown in Figure 7 where the scheme can achieve full transmit diversity up to M order with M transmit antennas while allowing the use of a very simple ML decoding algorithm and linear combining at the receiver. However, OSTBC trades off full diversity gain for lower spatial rate when $M > 2$. In order to provide a compromise between full diversity and full rate, a Quasi-Orthogonal STBC (Quasi OSTBC) scheme was proposed in (Jafarkhani, 2005).

Another class of STCs is the Space-Time Trellis Codes (STTC) (Tarokh et. al., 1998). STTC achieves higher coding gain and is comparable to STBC in terms of achieving full transmit diversity gain. However, the encoder design based on trellis-coded modulation leads to a more complex receiver with a Viterbi algorithm decoding implementation. The ML decoder complexity grows exponentially with the number of bits per symbol, thus limiting the achievable data rates.

In the context of practicality in WSNs, the main obstacle of MISO STBCs and STTCs schemes implementation is the issues of how to provide each transmitting node with the transmit sequences knowledge and how different transmit sequences are assigned to each node in order to provide an orthogonal or quasi-orthogonal design.

A better and practical approach as suggested in (Yang et. al., 2007) is when the source node broadcasts the transmit sequences to its particular neighbours in order to provide the transmit sequences knowledge together with the original data signal. Such an approach introduces an increasing packet overhead as M increases, prior data packet transmission. The overhead is a compromise with full diversity gain which achieves higher reliability and lower transmission energy.

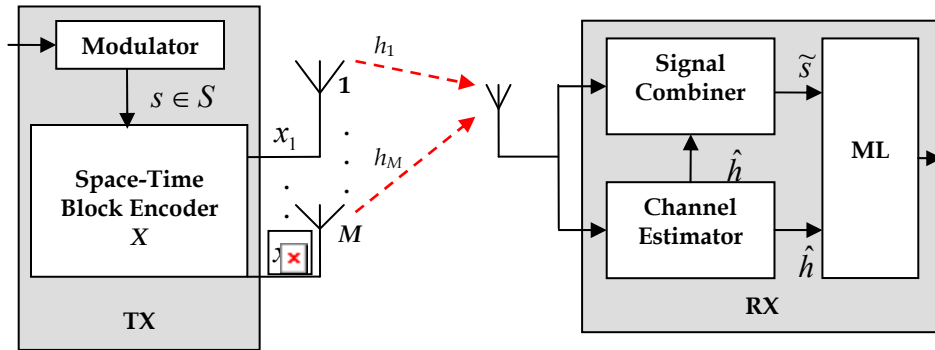


Fig. 7. A STBCs transmit diversity system with M transmit antennas and 1 receive antenna.

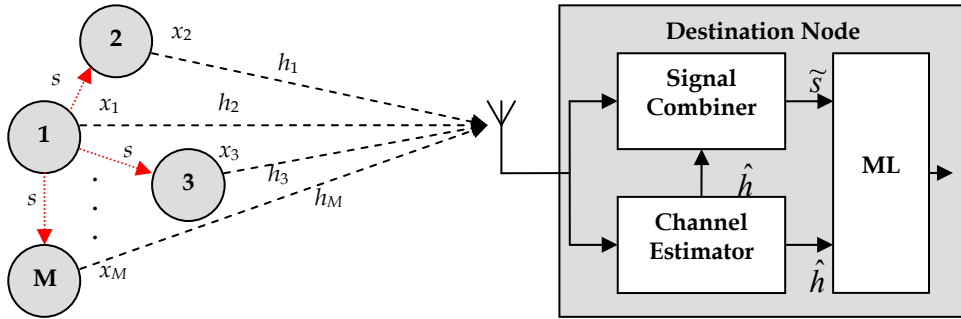


Fig. 8. A cooperative STBC transmit diversity system with M transmit nodes and 1 destination.

As a comparison, MISO STBC is more practical and promising to be implemented in WSNs due to a simpler decoding algorithm which leads to lower processing energy at the receiver. On the other hand, the simpler encoding and decoding algorithms of MISO STBC come at the cost of higher transmission power compared to the MISO beamforming scheme. The pictorial concept of cooperative MISO STBC is shown in Figure 8.

4.2.3 Spatial Multiplexing MIMO System

The main motivation of spatial multiplexing (SM) scheme is to achieve a higher data rate while maintaining the same full diversity gain. Thus the main purpose of SM schemes is basically to complement the lack of spatial rate in MISO STBC and STTC schemes. Therefore SM schemes are designed purposely for high data rate applications such as mobile communications systems and wireless local area networks. Though the current WSNs target only low to medium data rate applications, future generations of WSNs may require to operate with such high data rate applications which makes the investigation of cooperative SM in WSNs relevant and useful.

The main concept of SM (also referred as Layered Space-Time Codes - LSTC (Foschini, 1996)) is to provide simultaneous transmissions of M information streams in the same frequency band from M transmit antennas. However, by using such a transmission method, a constraint is introduced where the number of receive antennas must be equal or greater than the number of transmit antennas ($N \geq M$) in order to separate and detect the M transmitted signals. The separation process involves a combination of interference suppression and cancellation. The achievable spatial rate is given as $R_c b M$ where R_c denotes the rate of the channel code whenever channel coding is employed and 2^b is the signal constellation size. When full channel code is achieved with $R_c = 1$ and Binary PSK is used with $b = 1$, we can show that the spatial rate is increased linearly with M .

Among the simplest SM schemes is Bell Laboratories Layered Space-Time (BLAST) (Golden et. al., 1999). There are various versions of the BLAST schemes in the literature such as Vertical BLAST (VBLAST), Horizontal BLAST (HBLAST) and Diagonal BLAST (DBLAST). The simplest version is VBLAST due to the simplest encoder architecture compared to HBLAST and DBLAST (Vucetic and Yuan, 2003). VBLAST is also referred to as an uncoded LST scheme while HBLAST and DBLAST are classified as coded LST schemes. The simple encoder architecture makes VBLAST the most practical version of SM schemes for

implementation in WSNs in order to keep the complexity and power consumption as low as possible. There are other SM schemes such as threaded LSTCs and multilayered LSTCs, with higher spatial rate but they come at the cost of more complex encoding and decoding mechanisms. Obviously these schemes are not practical to be implemented in WSNs and thus are not considered in our work.

A VBLAST encoder is shown in Figure 9. As shown in the figure, the bit stream is demultiplexed into M sub-streams. Each M sub-stream is then modulated and transmitted from M transmit antennas. The transmitted signal matrix is given as:

$$X = \begin{bmatrix} s_k^t \end{bmatrix} \quad (23)$$

where $k = 1, 2, \dots, M$ and $t = 1, 2, \dots, L$ with L is the transmission block length. At a given time t , the transmitter transmits the t^{th} column from the transmission matrix, one symbol from each k^{th} antenna. The given transmission mechanism represents vertical structuring referring to transmission sequences of the matrix columns in the space-time domains (Vucetic and Yuan, 2003). Given the system constraint of $N \geq M$, the achievable spatial rate is bM and the achievable spatial diversity depends on the detection scheme employed at the receiver. When a Zero Forcing (ZF) or a Minimum Mean Square Error (MMSE) decoder is used at the receiver for the separation and detection, the achievable spatial diversity varies between 1 to N (Vucetic and Yuan, 2003). In order to gain full spatial diversity equal to N , a ML decoder must be employed at the receiver at the cost of a more complex decoder compared to ZF and MMSE. The complexity of the ML decoder increases linearly with bN . Thus the use of a modulation scheme with the smallest constellation size (e.g. $b = 1$) is very helpful to reduce the decoder complexity while achieving higher spatial diversity gain.

In the context of practicality in WSNs, there are three major issues that must be tackled: how to provide the data packet stream to $M-1$ transmitting nodes, how to transmit each M data packet stream simultaneously from M nodes and how to forward the receive data packets by $N-1$ receiving nodes to the destination. An example of architecture for cooperative SM which is based on the VBLAST scheme was proposed in (Yang et. al., 2007) and is shown in Figure 10. The cooperative SM scheme has the following operations:

- a) Source node broadcasts the original data packet stream to its $M-1$ neighbour nodes with very low power and all the M transmitting nodes send the same data packet streams simultaneously after the sending timer expires.
- b) N receiving nodes receive the data packet streams from M transmitting nodes and each receiving node employs a ML decoder to decode the data packet and forward the data packet to the destination node.

In order to gain both spatial diversity and spatial rate, the constraint of the traditional SM scheme such as $N \geq M$ also works for the cooperative SM scheme. Consider the transmission route in Figure 10. The error rate in each route is given as:

$$P_e = P_{eM-1} + P_{epp(dst)} - P_{epp(dst)} P_{eM-1(recv)} \quad (24)$$

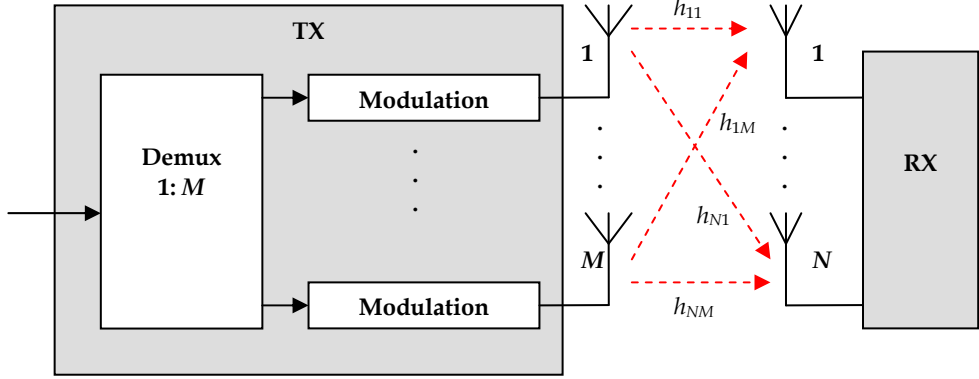


Fig. 9. A VBLAST spatial multiplexing system with M transmit antennas and N receive antennas.

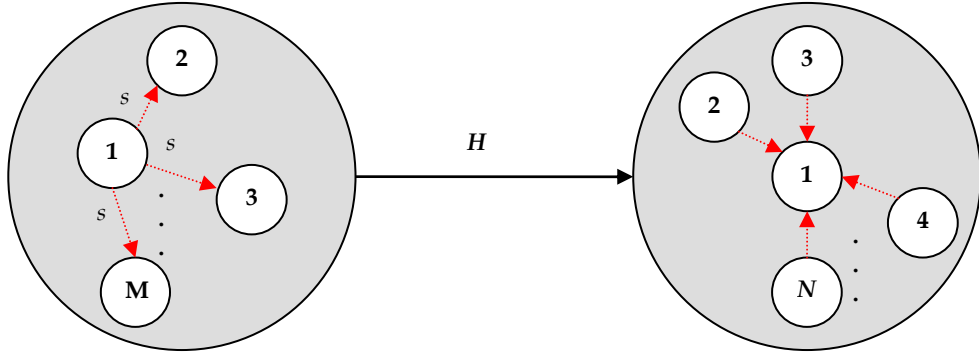


Fig. 10. A cooperative spatial multiplexing system with M transmit nodes and N receive nodes.

where P_{eM_1} is the error rate for M nodes cooperatively sending to one receiving node which relates to the power summation from multiple paths M , and different fading characteristics that may occur in different signal transmission paths. $P_{ep(dst)}$ is the error rate from one receiving node to the destination. A simple majority decision rule is employed at the destination node when multiple packets are received from $N-1$ nodes (Yang et. al., 2007). The data packet stream with the lowest BER, which means that the SNR is maximised, is selected at the destination node. If each receiving node in the receiving group has the same BER, the BER in the destination node after the reception from the N nodes forming the reception group is given as:

$$P_{eM_N} = \sum_{k=N/2}^N \binom{N}{k} P_e^k (1-P_e)^{N-k}. \quad (25)$$

5. Performance Analysis

In this section, we study the performance of cooperative MIMO schemes discussed earlier on, namely cooperative MISO Beamforming (BF), MISO STBC and MIMO SM schemes. The clock jitter impact is modelled as a timing error function in Section 5.1. The error performance for each cooperative scheme is modelled in Section 5.2 while the results are discussed in Section 5.3.

5.1 Timing Error Modeling

We consider the impact of imperfect synchronisation which is caused by clock jitter alone. Each cooperative sending nodes experiences clock jitter with the jitter around a reference clock, T_o denoted as T_j^m where $1 \leq m \leq M$. The worst case scenario is considered here with only 2 cooperative transmitting nodes where the clock jitters are fixed at the extreme ends, $T_j^1 = -\frac{\Delta T_b}{2}, T_j^2 = +\frac{\Delta T_b}{2}$ where $0 \leq \Delta T_b \leq T_b$ and T_b is the bit duration. Thus the clock jitters difference is $\Delta T_j = T_j^1 - T_j^2 = \Delta T_b$. The effect of imperfect synchronisation can be modelled as a degrading function of the bit period which consequently degrades the received bit energy. Therefore the timing error as a function of the bit period and clock jitters difference is given as:

$$T_e = T_b - \Delta T_j. \quad (26)$$

5.2 Error Performance Modeling

We derive the two most important performance parameters to measure the channel condition and to evaluate the link reliability: BER and PER. Without Forward Error Correction (FEC), the relationship between Packet Error Rate (PER), P_p and BER, P_b is given by:

$$P_p = 1 - (1 - P_b)^{N_{data}} \quad (27)$$

where N_{data} is the packet length in bits. Consider the case of BPSK modulation under quasi-static Rayleigh fading with fading gain h , experiencing a square law path loss without channel codes. In the SISO system, the conditional SNR is given by (Proakis, 2001):

$$\gamma_{bSISO} = \frac{P_t |h|^2 G_t G_r}{N_o M_l \left(\frac{4\pi d}{\lambda} \right)^2} \quad (28)$$

where P_t is the transmission power, d is the distance between the sending and destination node, G_t and G_r are the transmission and reception antenna gain, λ is the carrier wave length, M_l is the link margin and N_o is single-sided thermal noise power spectral density (PSD) given as -171 dBm/Hertz.

The probability density function (PDF) of γ_{bSISO} is given by:

$$p(\gamma_{bSISO}) = \frac{1}{\bar{\gamma}_{bSISO}} \exp\left(-\frac{\gamma_{bSISO}}{\bar{\gamma}_{bSISO}}\right) \quad (29)$$

where $\bar{\gamma}_{bSISO}$ is the average SNR. Assume that $E[|h|^2] = 1$ (Larsson and Stoica, 2003), then the value of $\bar{\gamma}_{bSISO}$ is given by:

$$\bar{\gamma}_{bSISO} = \frac{P_t E[|h|^2] G_t G_r}{N_o M_l \left(\frac{4\pi d}{\lambda}\right)^2} = \frac{P_t G_t G_r}{N_o M_l \left(\frac{4\pi d}{\lambda}\right)^2}. \quad (30)$$

The average BER can be expressed as:

$$E_h[P_{bSISO}] = E_h[Q(\sqrt{2\gamma_{bSISO}})]. \quad (31)$$

The upper bound of the average BER can be derived as (Proakis, 2001):

$$Q(\sqrt{2\gamma_{bSISO}}) = p(x \geq \sqrt{2\gamma_{bSISO}}) \leq \exp\left(-\frac{(\sqrt{2\gamma_{bSISO}})^2}{2}\right) \quad (32)$$

$$E[Q(\sqrt{2\gamma_{bSISO}})] \leq E[\exp(-\gamma_{bSISO})]. \quad (33)$$

The moment generating function of γ_{bSISO} is given by (Proakis, 2001):

$$\Phi(S) = E[\exp(\gamma_{bSISO} S)] = \frac{1}{S \bar{\gamma}_{bSISO}} \quad (34)$$

$$E_h[P_{bSISO}] \leq E[\exp(-\gamma_{bSISO})] = \Phi(-1) = (1 + \bar{\gamma}_{bSISO})^{-1}. \quad (35)$$

If there are M nodes in the sending group and we are comparing between optimal cooperative BF and STBC schemes, the SNRs for both schemes for perfect synchronisation scenario at the destination node can be given by:

$$\gamma_{bBF} = \sum_{k=1}^M |h_k|^2 \frac{P_t G_t G_r}{N_o M_l \left(\frac{4\pi d}{\lambda}\right)^2} = \sum_{k=1}^M \gamma_{BFk} \quad (36)$$

$$\gamma_{bSTBC} = \sum_{k=1}^M |h_k|^2 \frac{P_t G_t G_r}{M \cdot N_o M_l \left(\frac{4\pi d}{\lambda} \right)^2} = \sum_{k=1}^M \gamma_{STBCk} \quad (37)$$

and the SNRs for imperfect synchronisation scenario at the destination node can be given by:

$$\gamma_{bBF} = \sum_{k=1}^M |h_k|^2 \frac{P_t G_t G_r}{N_o M_l \left(\frac{4\pi d}{\lambda} \right)^2} \cdot \frac{T_e}{T_b} = \sum_{k=1}^M \gamma_{BFk} \quad (38)$$

$$\gamma_{bSTBC} = \sum_{k=1}^M |h_k|^2 \frac{P_t G_t G_r}{M \cdot N_o M_l \left(\frac{4\pi d}{\lambda} \right)^2} \cdot \frac{T_e}{T_b} = \sum_{k=1}^M \gamma_{STBCk} \quad (39)$$

where γ_{BFk} and γ_{STBCk} are the instantaneous SNR on the k^{th} channel.

The PDF of γ_{BFk} and γ_{STBCk} are:

$$p(\gamma_{BFk}) = \frac{1}{\bar{\gamma}_{BFk}} \exp^{-\frac{\gamma_{BFk}}{\bar{\gamma}_{BFk}}} \quad (40)$$

$$p(\gamma_{STBCk}) = \frac{1}{\bar{\gamma}_{STBCk}} \exp^{-\frac{\gamma_{STBCk}}{\bar{\gamma}_{STBCk}}} \quad (41)$$

Assume that $E[\|h_k\|^2] = 1$ (Larsson and Stoica, 2003), then the values of $\bar{\gamma}_{BFk}$ and $\bar{\gamma}_{STBCk}$ for perfect synchronisation scenario become:

$$\bar{\gamma}_{BFk} = \frac{P_t E[\|h_k\|^2] G_t G_r}{N_o M_l \left(\frac{4\pi d}{\lambda} \right)^2} = \frac{P_t G_t G_r}{N_o M_l \left(\frac{4\pi d}{\lambda} \right)^2} \quad (42)$$

$$\bar{\gamma}_{STBCk} = \frac{P_t E[\|h_k\|^2] G_t G_r}{M \cdot N_o M_l \left(\frac{4\pi d}{\lambda} \right)^2} = \frac{P_t G_t G_r}{M \cdot N_o M_l \left(\frac{4\pi d}{\lambda} \right)^2} \quad (43)$$

and the average SNRs for imperfect synchronisation scenario can be given by:

$$\bar{\gamma}_{BFk} = \frac{P_t E[\|h_k\|^2] G_t G_r}{N_o M_l \left(\frac{4\pi d}{\lambda} \right)^2} \cdot \frac{T_e}{T_b} = \frac{P_t G_t G_r}{N_o M_l \left(\frac{4\pi d}{\lambda} \right)^2} \cdot \frac{T_e}{T_b} \quad (44)$$

$$\bar{\gamma}_{STBCk} = \frac{P_t E[\|h_k\|^2] G_t G_r}{M \cdot N_o M_l \left(\frac{4\pi d}{\lambda} \right)^2} \cdot \frac{T_e}{T_b} = \frac{P_t G_t G_r}{M \cdot N_o M_l \left(\frac{4\pi d}{\lambda} \right)^2} \cdot \frac{T_e}{T_b}. \quad (45)$$

The moment generating functions of γ_{bBF} and γ_{bSTBC} are (Proakis, 2001):

$$\Phi(S) = E[\exp(\gamma_{bBF} S)] = \prod_{k=1}^M \frac{1}{S \bar{\gamma}_{BFk}} \quad (46)$$

$$E_h[P_{bBF}] \leq E[\exp(-\gamma_{bBF})] = \Phi(-1) = (1 + \bar{\gamma}_{BFk})^{-M} \quad (47)$$

$$\Phi(S) = E[\exp(\gamma_{bSTBC} S)] = \prod_{k=1}^M \frac{1}{S \bar{\gamma}_{STBCk}} \quad (48)$$

$$E_h[P_{bSTBC}] \leq E[\exp(-\gamma_{bSTBC})] = \Phi(-1) = (1 + \bar{\gamma}_{STBCk})^{-M}. \quad (49)$$

The average BER for the cooperative SM scheme in (Yang et. al., 2007) is given as:

$$P_{bSM} = \sum_{k=N/2}^N \binom{N}{k} P_e^k (1 - P_e)^{N-k} \quad (50)$$

$$P_e = E_h[P_{bMISO}] + E_h[P_{bSISO}] - E_h[P_{bSISO}] E_h[P_{bMISO}] \quad (51)$$

where P_e is the error rate in each route and N is the number of nodes forming the reception group. The average SNR of the MISO scheme in Equation (51) is the same as the average SNR of the cooperative MISO BF scheme (Yang et. al., 2007). Thus we assume that the average BER is the same for both schemes. Table 2 lists the system parameters used for evaluating BER performance of the three cooperative MIMO schemes.

5.3 Performance Results and Discussions

Figures 11, 12 and 13 show the corresponding results for perfect synchronisation scenarios. For comparison, those figures also show the BER performance of the corresponding SISO scheme. As we can see, in general, cooperative BF outperforms the other schemes except for the special case below the 10mW transmit power where cooperative SM performs better.

However, this special case may not have a significant impact due to the fact that the operating transmission power for WSNs is in the range between 25mW to 50mW (Kohvakka et. al., 2006). Also, we can observe that the diversity gain of cooperative SM depends on N and not M as shown in Figures 12 and 13. In addition, the cooperative SM achieves spatial rate equal to M .

Figures 14 and 15 show the corresponding results for imperfect synchronisation scenarios. As we can see, in general SISO outperforms other schemes above $0.8T_b$ and cooperative SM outperforms the other schemes when the diversity gain is getting higher. However, when the diversity gain of all the cooperative schemes is the same, cooperative BF outperforms the other schemes.

Symbol	Quantity
f_c	2.4 GHz
$G_t G_r$	5 dBi [5]
M_t	40 dB [5]
d	100 meters
d_m	10 meters
R_b	250 Kbps

Table 2. System parameter for ber and per modeling

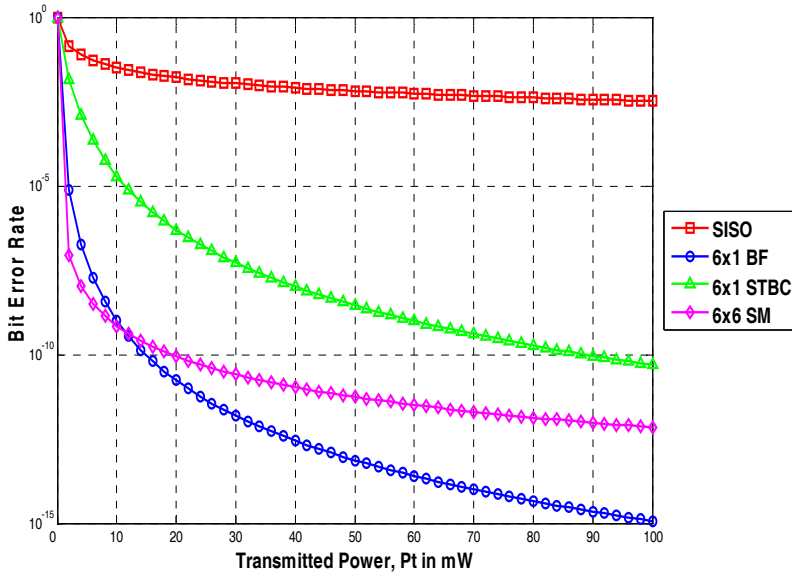


Fig. 11. BER vs. transmission power for various cooperative schemes with $M = 6$ and $N = 1$ (Cooperative BF and Cooperative STBC) and $M = N = 6$ (Cooperative SM).

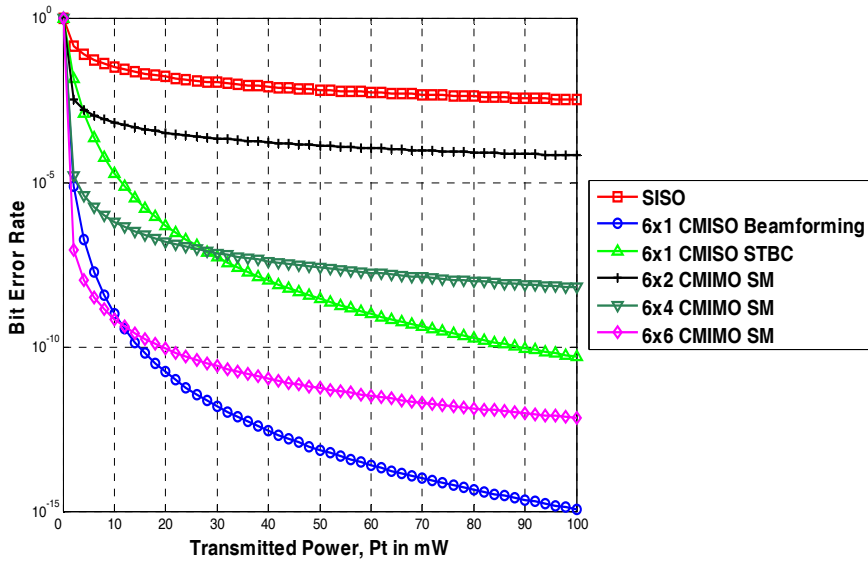


Fig. 12. BER vs. transmission power for various cooperative schemes with $M = 6$ and $N = 1$ (Cooperative BF and Cooperative STBC) and various $N = 2, 4$ and 6 for Cooperative SM with $M = 6$.

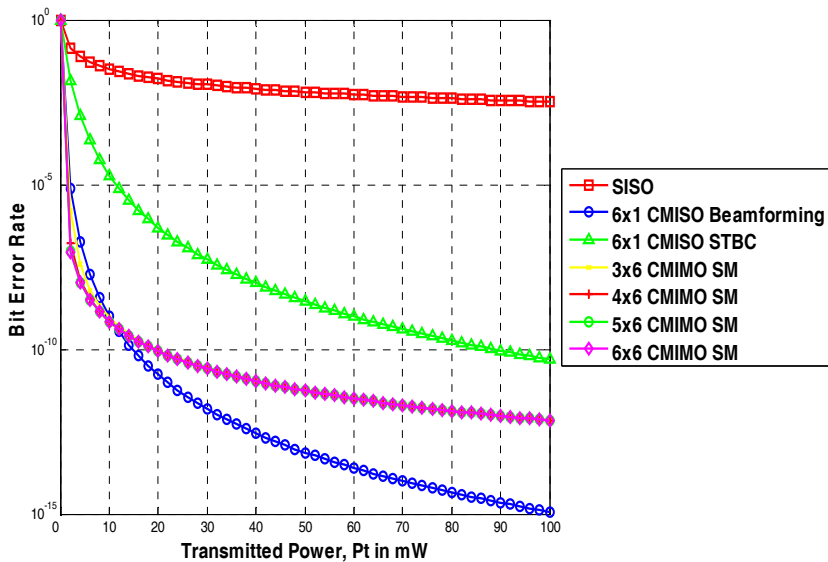


Fig. 13. BER vs. transmission power for various cooperative schemes with $M = 6$ and $N = 1$ (Cooperative BF and Cooperative STBC) and various $M = 3, 4, 5$ and 6 for Cooperative SM with $N = 6$.

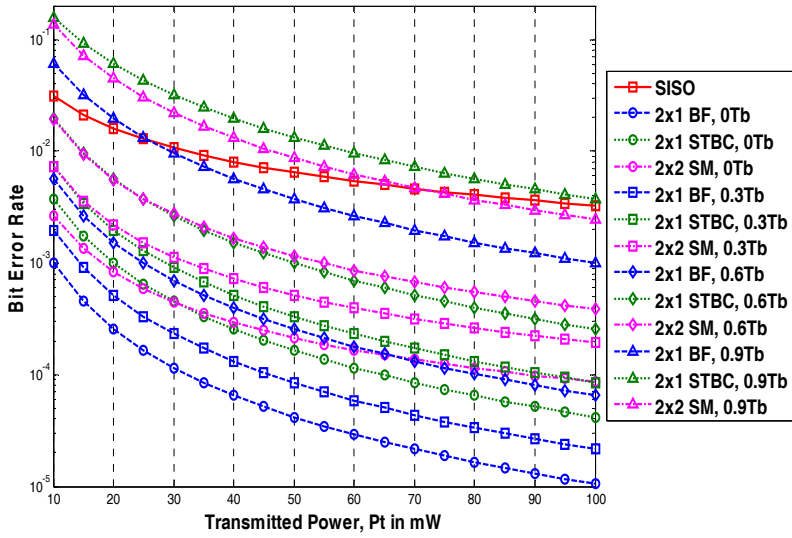


Fig. 14. BER vs. transmission power for various imperfect synchronisation cooperative schemes with $M = 2$ and $N = 1$ (Cooperative BF and Cooperative STBC) and $M = N = 2$ (Cooperative SM).

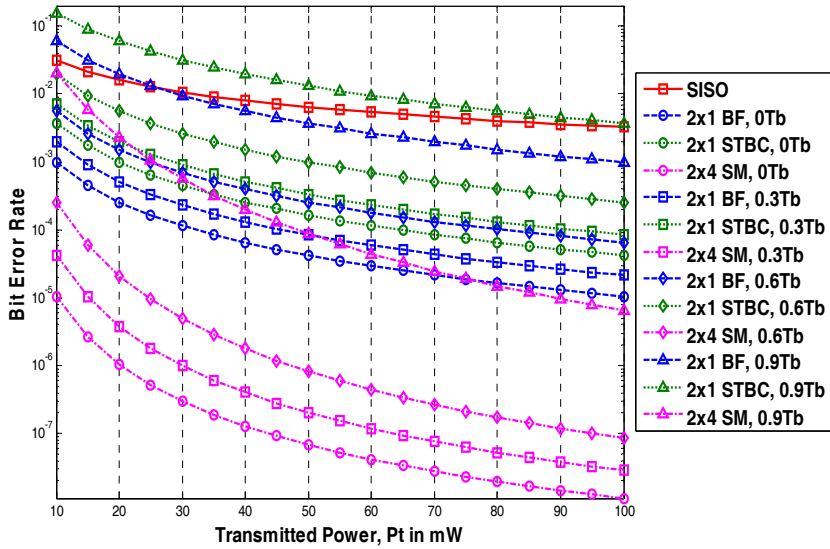


Fig. 15. BER vs. transmission power for various imperfect synchronisation cooperative schemes with $M = 2$ and $N = 1$ (Cooperative BF and Cooperative STBC) and $N = 4$ for Cooperative SM with $M = 2$.

6. Conclusion

This chapter has examined the major diversity techniques and various cooperative configurations, including BF, STBC and SM schemes in conjunction with performance evaluation and comparative literature. Both cooperative BF and STBC schemes utilise the MISO concept while the SM scheme utilises the MIMO concept. We have shown that the cooperative MISO BF is the most promising scheme to be implemented in WSNs due to the lowest error performance among others with the same diversity gain. Also, cooperative MISO BF outperforms other cooperative schemes in imperfect synchronisation scenarios. On the other hand, cooperative MIMO SM is more practical in terms of lower error performance and tolerance to clock jitter error when its diversity gain is higher than the others. In addition, cooperative MIMO SM provides a higher spatial rate as M grows.

The comparative study relates the diversity gain with the reduction in the transmission power by increasing the communication link reliability. However, in order to find the best or optimal scheme to be used in WSNs, we have to compare all the three schemes in terms of total energy consumption which must include both the transmission power and circuit power for each sensor node in the network. The discussion in this chapter can provide a basis for further study to find the optimal cooperative MIMO scheme when both transmission power and circuit power are considered for all required energy components of cooperative communications in WSNs.

7. References

- Alamouti, S.M. (1998). A Simple Transmit Diversity Technique for Wireless Communications, IEEE Journal on Selected Areas in Communications, vol. 16, pp. 1451-1458.
- Duman, T.M. & Ghayeb, A. (2007). *Coding for MIMO Communication Systems*, First ed. West Sussex, England: John Wiley & Sons Ltd.
- Foschini, G.J. (1996). Layered Space-time Architecture for Wireless Communication in A Fading Environment when using Multi-element antennas, Bell Labs Technical Journal, pp. 41-59.
- Golden, G.D.; Foschini, G.J.; & Valenzuela, R.A. (1999). Detection Algorithm and Initial Laboratory Results using the V-BLAST Space-time Communication Architecture, Electronic Letters, vol. 35, pp. 14-15.
- Gupta, G. & Younis, M. (2003). Fault-Tolerant Clustering of Wireless Sensor Networks, presented at IEEE Wireless Communications and Networking Conference (WCNC).
- Jafarkhani, H. (2005). *Space-Time Coding: Theory and Practice*, First ed: Cambridge University Press.
- Jagannathan, S.; Aghajan, H.; & Goldsmith, A. (2004). The effect of time synchronization errors on the performance of cooperative MISO systems, presented at IEEE Global Communications Conference (Globecom), Dallas, Texas, USA.
- Karl, H. & Willig, A. (2007). MAC Protocols, In: *Protocols and Architectures for Wireless Sensor Networks*, pp. 111-148, John Wiley & Sons, 978-0-470-09510-2, West Sussex, England.
- Kohvakka, M.; Kuorilehto, M.; Hannikainen, M. & Hamalainen, T.D. (2006). Performance Analysis of IEEE 802.15.4 and Zigbee for Large-scale Wireless Sensor Network

- Applications, *Proceedings of ACM International Workshop on Performance Evaluation of Wireless Ad hoc, Sensor, and Ubiquitous Networks*, pp. 1-6, Malaga, Spain.
- Kuorilehto, M.; Kohvakka, M.; Suhonen, J.; Hamalainen, P.; Hannikainen, M. & Hamalainen, T.D. (2007). MAC Protocols, In: *Ultra-Low Energy Wireless Sensor Networks in Practice*, pp. 73-88, John Wiley & Sons, 978-0-470-05786-5, West Sussex, England.
- Larsson, E.G. & Stoica, P. (2003). *Space-Time Block Coding for Wireless Communications*, First ed. Cambridge, UK: Cambridge University Press.
- Li, X. & Hwu, J. (2006). Performance of Cooperative Transmissions in Flat Fading Environment with Asynchronous Transmitters, presented at Military Communications Conference (MILCOM 2006), Washington DC, USA.
- Li, X. (2004). Space-Time Coded Multi-Transmission Among Distributed Transmitters Without Perfect Synchronization, *IEEE Signal Processing Letters*, vol. 11, pp. 948-951.
- Li, X.; Chen, M.; & Liu, W. (2004). Cooperative Transmissions in Wireless Sensor Networks with Imperfect Synchronization, presented at Conference on Signals, Systems and Computers, Pacific Grove, CA.
- Litva, J. & Lo, T.K.Y. (1996). *Digital Beamforming in Wireless Communications*: Artech House Publisher.
- Liu, Z.; Giannakis, G.B.; Zhuo, S.; & Muquet, B. (2001). Space-time coding for broadband wireless communications, *Wireless Communications and Mobile Computing*, vol. 1, pp. 35-53.
- Naquib, A.F. & Calderbank, R. (2000). Space-time codes for high data rate wireless communications, in *IEEE Signal Processing*, vol. 47, pp. 76-92.
- Proakis, J.G. (2001). Probability and Stochastic Processes, In: *Digital Communications*, pp. 17-36, McGraw-Hill, 0-07-232111-3, Singapore, Singapore.
- Rappaport, T.S. (2002). *Wireless Communications: Principles and Practice*, second ed. Upper Saddle River, NJ, USA: Pearson Education International.
- Seshadri, N. & Winters, J.H. (1993). Two Signaling Schemes for Improving the Error Performance of Frequency-division-duplex (FDD) Transmission Systems using Transmitter Antenna Diversity, presented at IEEE Vehicular Technology Conference.
- Simon, M.K. & Alouini, M.S. (2000). *Digital Communication over Fading Channels: A Unified Approach to Performance Analysis*: John Wiley & Sons.
- Singh, M. & Prasanna, V.K. (2003). A Hierarchical Model for Distributed Collaborative Computation in Wireless Sensor Networks, presented at IEEE International Parallel and Distributed Processing Symposium.
- Tarokh, V.; Jafarkhani, H.; & Calderbank, A.R. (1999). Space-time Block Codes from Orthogonal Designs, *IEEE Transactions on Information Theory*, vol. 45, pp. 1456-1467.
- Tarokh, V.; Seshadri, N.; & Calderbank, A.R. (1998). Space-time Codes for High Data Rate Wireless Communication: Performance Criterion and Code Construction, *IEEE Transactions on Information Theory*, vol. 44, pp. 744-765.
- Vucetic, B. & Yuan, J. (2003). *Space-Time Coding*. West Sussex, England: John Wiley & Sons Ltd.
- Winters, J.H. (1983). Switched diversity with feedback for DPSK mobile radio systems, *IEEE Transactions on Vehicular Technology*, vol. 32, pp. 134-150.

- Wittneben, A. (1991). Base Station Modulation Diversity for Digital SIMULCAST, presented at IEEE Vehicular Technology Conference.
- Yang, H.; Shen, H.-Y. & Sikdar, B. (2007). A MAC Protocol for Cooperative MIMO Transmissions in Sensor Networks, *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM)*, pp. 636-640, Washington, USA.
- Yuksel, M. & Erkip, E. (2004). Diversity Gains and Clustering in Wireless Relaying, presented at IEEE International Symposium of Information Theory.

Optimal Cooperative MIMO Scheme in Wireless Sensor Networks

M. Riduan Ahmad¹, Eryk Dutkiewicz², Xiaojing Huang³ and M. Kadim Suaidi⁴

^{1,4}*Universiti Teknikal Malaysia Melaka*, ²*Macquarie University*, ³*CSIRO ICT Centre*

^{1,4}*Malaysia*, ^{2,3}*Australia*

1. Introduction

Cooperative Multiple-Input Multiple-Output (MIMO) has been proposed as a transmission strategy to combat the fading problem in Wireless Sensor Networks (WSNs) to reduce the retransmission probability and lower the transmission energy. Among the earliest work on cooperative MIMO in WSNs is the analysis of the Space-Time Block Coding (STBC) scheme to achieve lower Bit Error Rate (BER) and significant energy savings. The work is continued with the implementation of the Low-Energy Adaptive Clustering Hierarchy (LEACH) Medium Access Control (MAC) protocol for clustered-based architectures. The combination of STBC and the LEACH scheme resulted in a significant improvement in transmission energy efficiency compared to the Single-Input Single Output (SISO) scheme.

Further study is conducted to compare the performance of STBC and various Spatial Multiplexing (SM) schemes such as Vertical Bell Labs Layered Space-Time (V-BLAST) and Diagonal BLAST. In this study, LEACH MAC was also utilized and lower transmission energy and latency were achieved against the SISO scheme. However, the centralized architecture leads to energy wastage and higher latency compared to a distributed architecture. On the other hand, the implementation of a distributed architecture needs to consider synchronisation issues. Thus a practical cooperative MIMO scheme for distributed asynchronous WSNs is needed.

Moreover, a practical MAC that can suit cooperative transmission is required. A combination of a practical MAC protocol and an efficient MIMO scheme for asynchronous cooperative transmission leads to a more energy efficient and lower latency cooperative MIMO system. A combination of a MAC protocol and a cooperative SM scheme for cooperative MIMO transmission has been proposed in previous study where the combined scheme achieves significant energy efficiency and lower latency.

Furthermore, a transmit Maximum Ratio Combiner (MRC) scheme is suggested to be more tolerant to the jitter difference than the Alamouti STC scheme in network with imperfect transmitting nodes synchronisation. In this chapter, we expand these studies to two other cooperative MIMO schemes, namely Beamforming (BF) and STBC for both network scenarios: perfect and imperfect transmitting nodes synchronisation. The optimal cooperative MIMO scheme combined with an appropriate MAC protocol should lead to the lowest energy consumption and lowest packet latency.

The rest of this chapter is organised as follows. Section 2 describes the system model considered in this chapter. Section 3 and Section 4 model the system performance and are followed by Section 5 presenting the analytical results for the three cooperative MIMO schemes (BF, SM and STBC) in terms of total energy consumption and packet latency. Finally the chapter is concluded in Section 6.

2. System Model

The baseline system for cooperative MIMO communication is equipped with a CMA_{CON} protocol as proposed and evaluated in (Yang et. al., 2007). Sleep cycles are not implemented in order to ensure that the cooperative nodes are always available to perform cooperative transmission and reception. In order to avoid collision, we assume that during the cooperative transmission and reception, other nodes in the vicinity that are not involved in the transmission are put in the silent mode for the whole transmission duration. The duration to remain silent is obtained from the Network Allocation Vector (NAV).

Also in this chapter we consider the impact of imperfect synchronisation caused by clock jitter alone. Each cooperative transmitting node experiences clock jitter with the jitter around a reference clock, T_o denoted as T_j^m where $1 \leq m \leq M$. The worst case scenario is considered here with only 2 cooperative transmitting nodes where the clock jitters are fixed at the extreme ends, $T_j^1 = -\frac{\Delta T_b}{2}, T_j^2 = +\frac{\Delta T_b}{2}$ where $0 \leq \Delta T_b \leq T_b$ and T_b is the bit

duration. Thus the clock jitters difference is $\Delta T_j = T_j^1 - T_j^2 = \Delta T_b$. The effect of imperfect synchronisation can be modelled as a degrading function of the bit period which consequently degrades the received bit energy.

The baseline network configurations for MISO BF and STBC are shown in Figures 1 and 2 while for MIMO SM it is shown in Figure 3. The network is assumed to be distributed without any infrastructure and the nodes are fixed once they are deployed. A new node that wants to join the network should broadcast a packet after powering up to acknowledge its presence in the neighbourhood. A node checks its remaining energy regularly and when its total remaining energy is below the threshold, which indicates that its death is near, it informs the other nodes in the vicinity of the expected death time. Therefore the neighbouring nodes will exclude this node from any future cooperative MIMO transmission. The distance between the cooperating nodes either at the transmitting or receiving side is assumed to be very small compared to the distance between the source node and the destination node, d . We assume that there are M cooperative transmitting nodes and one receiving node for the perfect synchronisation scenario and $M = 2$ cooperative transmitting nodes and one receiving node for the imperfect synchronisation scenario. A special case for the spatial multiplexing scheme is used where the number of the cooperative receivers is assumed to be N .

In this section, we introduce two kinds of network configurations. The first network configuration involves data transmission from M cooperating transmitting nodes to one destination node by utilizing either of the two MIMO schemes: BF or STBC. An RTS-CTS handshaking method is performed as described in (Yang et. al., 2007) and the source node broadcasts the original data packet to its $M-1$ neighbours.

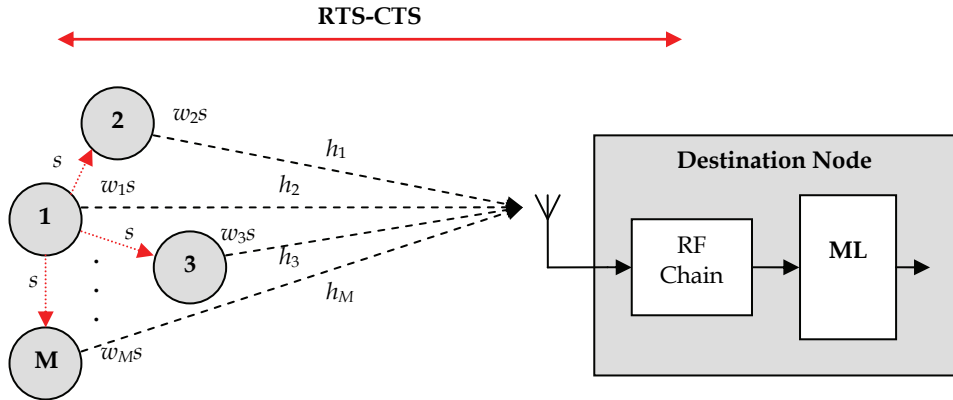


Fig. 1. A cooperative beamforming transmit diversity system with M transmit nodes and 1 destination.

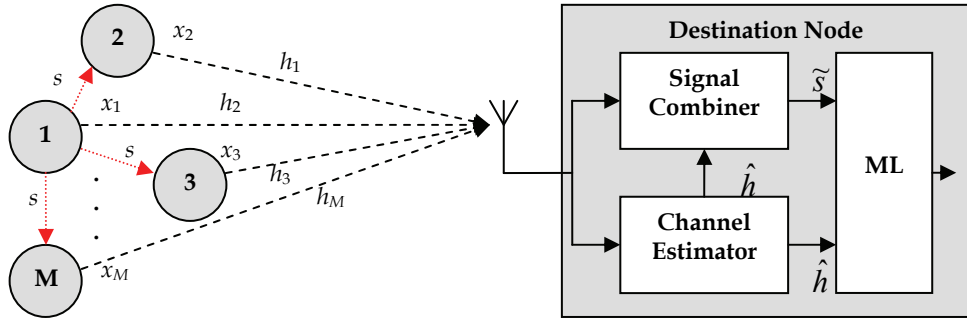


Fig. 2. A cooperative STBC transmit diversity system with M transmit nodes and 1 destination.

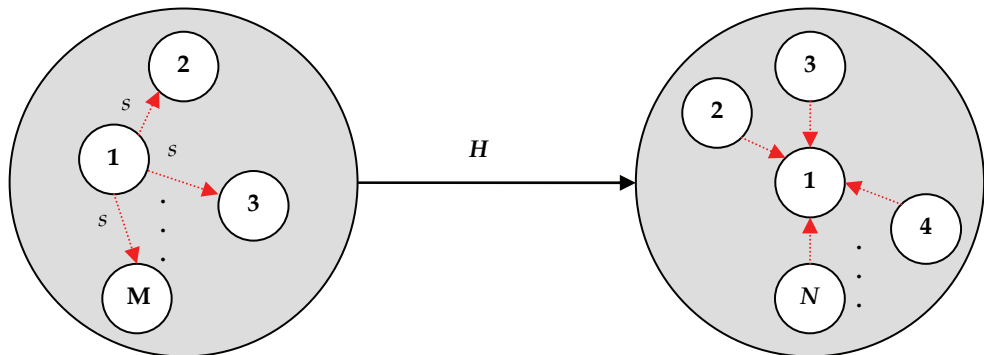


Fig. 3. A cooperative spatial multiplexing system with M transmit nodes and N receive nodes. In the case of the BF scheme, the channel information is estimated and optimized from the CTS packet by all the M nodes in order to weight the data packet. In the case of the STBC

scheme, all the M nodes encode the original data packet with the information supplied by the source node in the broadcast packet. Both schemes utilize a Maximum Likelihood (ML) detector and a coherent receiver is used. The second network configuration is the data transmission from M cooperating transmitting nodes to N cooperating receiving nodes by utilizing the concept of SM. The recovered data from $N-1$ nodes is forwarded to the destination node.

3. Energy Consumption Performance Analysis

The energy consumed by a sensor node can be categorized into two major parts (Cui et. al, 2004; Nguyen et. al., 2007): energy expended during running the transceiver circuits, P_c and energy expended during packet transmission, P_t . Therefore, both energy components must be considered when comparing the total energy consumption of cooperative MIMO and SISO transmission schemes. All the nodes in vicinity that are not involved in the transmission and reception are assumed to be in the sleep mode. Also for simplicity, the energy consumed during the transient mode from the sleep mode to the active mode and by the digital signal processing blocks is neglected.

3.1 SISO System

To model transmission energy for the non-cooperative or SISO system, we start with the power consumed by the power amplifier, P_{pa} . As given in (Cui et. al., 2004; Nguyen et. al., 2007), P_{pa} is dependent on the transmit power P_t and can be approximated as:

$$P_{pa} = (1 + \alpha)P_t \quad (1)$$

where $\alpha = \frac{\xi}{\eta} - 1$ with ξ denoting the drain efficiency of the Radio Frequency (RF) power amplifier and η denoting the Peak-to-Average Ratio (PAR) which depends on the modulation scheme and the associated constellation size. The total circuit power is given by:

$$P_c \approx (M \times P_{ct}) + (N \times P_{cr}) \quad (2)$$

where $P_{cr} = P_{LNA} + P_{mix} + P_{IFA} + P_{filr} + P_{ADC} + P_{syn}$ and $P_{ct} = P_{DAC} + P_{mix} + P_{filt} + P_{syn}$ are values for the power consumption of the Digital-to-Analogue Converter (DAC), mixer, Low Noise Amplifier (LNA), Intermediate Frequency Amplifier (IFA), active filters at the transmitter and the receiver, Analogue-to-Digital Converter (ADC) and frequency synthesizer whose values and a detailed block diagram are given in (Cui et. al., 2004; Nguyen et. al., 2007). Therefore, the total energy consumption per bit E_{bt} for the SISO system can be obtained as:

$$E_{bt} = \frac{(P_{pa} + P_c)}{R_b} \quad (3)$$

when $M = N = 1$. Equations (1) and (2) can be used to model the cooperative BF, STBC and SM systems with an arbitrary number of M and N . For the traditional Carrier Sense Multiple Access with Collision Avoidance (CSMA-CA) protocol, the energy consumed for an unsuccessful transmission attempt is given as:

$$E_{u_siso} = E_{rts} + E_{cts} + E_{data_siso} \quad (4)$$

and that for a successful attempt is given as:

$$E_{s_siso} = E_{rts} + E_{cts} + E_{data_siso} + E_{ack} \quad (5)$$

where E_{rts} , E_{cts} , E_{data_siso} , E_{ack} are the energy consumed while sending Ready-to-Send (RTS), Clear-to-Send (CTS), SISO data and Acknowledgment (ACK). Given the size of each packet as N_{rts} , N_{cts} , N_{data_siso} and N_{ack} , Equations (4) and (5) can be rewritten as:

$$E_{u_siso} = E_{bt} (N_{rts} + N_{cts} + N_{data_siso}) \quad (6)$$

$$E_{s_siso} = E_{bt} (N_{rts} + N_{cts} + N_{data_siso} + N_{ack}). \quad (7)$$

The expected total energy consumption is given as:

$$E_{siso} = \left(\frac{P_{psiso}}{1 - P_{psiso}} \right) E_{u_siso} + E_{s_siso} \quad (8)$$

where P_{psiso} is the packet error probability of the SISO system which can be obtained in (Ahmad et. al., 2008)

3.2 Cooperative MIMO System

In this sub-section, we consider two scenarios where the first scenario involves transmission from M cooperating transmitting nodes to 1 destination node with a local exchange of information at the transmitting side. This scenario applies to the cooperative MISO BF and STBC schemes. The second scenario deals with transmissions from M cooperating transmitting nodes to N receiving nodes with local exchanges at both the transmitting and receiving sides. This scenario applies to the cooperative MIMO SM scheme.

To model transmission energy for the first scenario, we start with the power consumed by the power amplifier, P_{paBs} during a local exchange between the source node and its cooperating neighbours. P_{paBs} is dependent on the local exchange transmitted power P_{tm} and can be approximated as:

$$P_{paBs} = (1 + \alpha) P_{tm}. \quad (9)$$

The total circuit power for the local exchange is given by:

$$P_{cBs} \approx P_{ct} + (M - 1) \times P_{cr}. \quad (10)$$

Therefore the total energy consumption per bit E_{btBs} for the local exchange can be obtained as:

$$E_{btBs} = \frac{(P_{paBs} + P_{cBs})}{R_b}. \quad (11)$$

The energy consumed for an unsuccessful BF and STBC transmissions attempt is given as:

$$E_{u_M} = E_{rts} + E_{cts} + E_{Bs} + M \cdot E_{data_M} \quad (12)$$

and that for a successful attempt is given as:

$$E_{s_M} = E_{u_M} + E_{ack} \quad (13)$$

where E_{Bs} and E_{data_M} are the amounts of energy consumed during packet broadcasting from the source node to its neighbours and the energy consumed for Cooperative BF or STBC data transmission. Given the size of each packet as N_{rts} , N_{cts} , N_{Bs} , N_{data_M} and N_{ack} , Equations (12) and (13) can be rewritten as:

$$E_{u_M} = E_{bt}(N_{rts} + N_{cts}) + E_{btBs}N_{Bs} + M \cdot E_{btdata_M}N_{data_M} \quad (14)$$

$$E_{s_M} = E_{u_M} + E_{bt} \times N_{ack}. \quad (15)$$

The expected total energy consumption is given as:

$$E_M = \left(\frac{P_{pM}}{1 - P_{pM}} \right) E_{u_M} + E_{s_M} \quad (16)$$

where P_{pM} is the packet error probability for BF or STBC which can be obtained in (Ahmad et. al., 2008). To model transmission energy for the second scenario, we start with the power consumed by the power amplifier, P_{paBr} from the destination node to its cooperating receiving nodes and P_{paCol} from $N-1$ receiving nodes to the destination node. P_{paBr} and P_{paCol} are dependent on the local exchange transmit power P_{tm} and can be approximated as:

$$P_{paBr} = (1 + \alpha)P_{tm} \quad (17)$$

$$P_{paCol} = (1 + \alpha)P_{tm}(N - 1). \quad (18)$$

The total circuit power for the former case is given by:

$$P_{cBr} \approx P_{ct} + (N - 1) \times P_{cr} \quad (19)$$

and the total circuit power for the latter case is given by:

$$P_{cCol} \approx (N-1) \times P_{ct} + P_{cr}. \quad (20)$$

Therefore the total energy consumption per bit E_{btBr} and E_{btCol} for both cases can be obtained as:

$$E_{btBr} = \frac{(P_{paBr} + P_{cBr})}{R_b} \quad (21)$$

$$E_{btCol} = \frac{(P_{paCol} + P_{cCol})}{R_b}. \quad (22)$$

The energy consumed for an unsuccessful SM transmission attempt is given as:

$$E_{u_SM} = E_{rts} + E_{Br} + E_{cts} + E_{Bs} + M \cdot E_{data_SM} + (N-1) \cdot E_{Col} \quad (23)$$

and that for a successful attempt is given as:

$$E_{s_SM} = E_{u_SM} + E_{ack} \quad (24)$$

where E_{Br} , E_{Col} and E_{data_SM} are the energy consumed during packet broadcasting from the destination node to its neighbours, the energy consumed by $N-1$ cooperating receiving nodes to the destination node and the energy consumed for the cooperative SM data transmission. Given the size of each packet as N_{rts} , N_{cts} , N_{Bs} , N_{data_SM} and N_{ack} , Equations (23) and (24) can be rewritten as:

$$E_{u_SM} = E_{bt}(N_{rts} + N_{cts}) + E_{btBr}N_{Br} + E_{btBs}N_{Bs} + M \cdot E_{btdata_SM}N_{data_SM} + (N-1)E_{btCol}N_{Col} \quad (25)$$

$$E_{s_SM} = E_{u_SM} + (E_{bt} \times N_{ack}). \quad (26)$$

The expected total energy consumption is given as:

$$E_{SM} = \left(\frac{P_{pSM}}{1 - P_{pSM}} \right) E_{u_SM} + E_{s_SM} \quad (27)$$

where P_{pSM} is the packet error probability of cooperative MIMO with spatial multiplexing which can be obtained in (Yang et. al., 2007). The values of the system parameters used in Figures 4 to 7 are listed in Table 1 (Cui et. al., 2004; Yang et. al., 2007).

Symbol	Quantity
N_{rts}	65 bits
N_{cts}	55 bits
N_{ack}	54 bits
N_{Bs}	1300 bits
N_{Br}	120 bits
$N_{data} = N_{Col}$	1024 bits
P_{mix}	30.3mW
P_{syn}	50mW
$P_{filt} = P_{fltr}$	2.5mW
P_{ADC}	9.85mW
P_{DAC}	15.48mW
P_{LNA}	20mW
P_{IFA}	3mW

Table 1. system parameter for energy consumption modeling

4. Packet Latency Performance Analysis

As we noted earlier, each packet transmission in cooperative transmission requires more steps which introduces more overhead. These steps may increase packet delays. However, the reduction of PER as the diversity gain increases from the cooperative MIMO exploitation can reduce the retransmissions rates which in turn can reduce packet latency. Sub-section 4.1 models packet latency performance for the non-cooperative SISO system. Comparison is then made with the models developed for the cooperative MIMO systems in Sub-section 4.2. The performance results are discussed in Section 5.

4.1 SISO System

For SISO communication, T_{rts} , T_{cts} , T_{data} and T_{ack} are the transmission periods for the RTS, CTS, DATA and ACK packets. The period with a successful transmission attempt is given as:

$$T_{s_siso} = T_{rts} + T_{cts} + T_{data} + T_{ack} \quad (28)$$

and the period with an unsuccessful transmission attempt is given as:

$$T_{u_siso} = T_{rts} + T_{cts} + T_{data} + T_{wait} \quad (29)$$

where T_{wait} is the duration for which the sender waits for an ACK packet. The packet transmission delay is then given as:

$$T_{d_SISO} = \left(\frac{P_{psiso}}{1 - P_{psiso}} \right) T_{u_siso} + T_{s_siso} \quad (30)$$

4.2 Cooperative MIMO System

In addition to the delay incurred as calculated in the previous section, the broadcast packet transmission from the source node to its neighbours introduces a broadcast transmission period, T_{Bs} in cooperative BF, STBC and SM transmissions. The transmission period of cooperative BF, STBC and SM data packets is the same as that for the SISO system due to the fact that the packet size and the modulation scheme are the same. The duration of a successful transmission attempt is given as:

$$T_{s_M} = T_{rts} + T_{cts} + T_{Bs} + T_{data} + T_{ack} \quad (31)$$

and the period with an unsuccessful transmission attempt is given as:

$$T_{u_M} = T_{rts} + T_{cts} + T_{Bs} + T_{data} + T_{wait} \cdot \quad (32)$$

The expected packet transmission delay is then given by:

$$T_{d_M} = \left(\frac{P_{pM}}{1 - P_{pM}} \right) T_{u_M} + T_{s_M} \cdot \quad (33)$$

For the case of cooperative MIMO SM, we introduce the delay for the broadcast transmission time of a recruitment message sent by the destination node, T_{Br} and the delay for the time required by the cooperating receiving nodes ($N-1$) to send the data to the destination, T_{col} . The duration of a successful transmission attempt is given as:

$$T_{s_SM} = T_{s_M} + T_{Br} + T_{col} \quad (34)$$

and the period with an unsuccessful transmission attempt is given as:

$$T_{u_SM} = T_{s_SM} + T_{wait} - T_{ack} \cdot \quad (35)$$

The expected packet transmission delay is then given by:

$$T_{d_SM} = \left(\frac{P_{pSM}}{1 - P_{pSM}} \right) T_{u_SM} + T_{s_SM} \cdot \quad (36)$$

The values of the system parameters used in Figures 8 to 11 are as follows: $T_{rts} = 0.52\text{ms}$, $T_{cts} = 0.44\text{ms}$, $T_{ack} = 0.432\text{ms}$, $T_{Bs} = 10.4\text{ms}$, $T_{Br} = 0.96\text{ms}$, $T_{data} = 8.192\text{ms}$, $T_{col} = 22.3\text{ms}$ (Nguyen et. al., 2007), and $T_{wait} = 70\text{ms}$ (Yang et. al., 2007).

5. Performance Results and Discussions

As shown in Figure 4, SISO is more energy efficient than the cooperative schemes at transmission powers above 100mW with any number of M and N nodes. The cooperative SM scheme suffers more in terms of energy efficiency because the total energy consumption is increasing as the diversity gain and the number of nodes M increases. The cooperative BF and STBC schemes suffer only with the increasing of the diversity gain.

As we noted earlier the cooperative schemes are more energy efficient when the transmission power is below 100mW. We can see in Figure 5, that the cooperative BF and STBC schemes outperform the cooperative SM scheme and that the cooperative BF scheme is more energy efficient than the cooperative STBC scheme with two transmitting nodes.

For imperfect synchronisation scenarios, as shown in Figure 6, in the case of equal diversity gain for all the schemes, the cooperative BF scheme is more energy efficient than the other schemes. However, as the diversity gain of the cooperative SM scheme is increased, as shown in Figure 7, cooperative SM outperforms the other schemes in terms of energy efficiency at and above $0.8T_b$ in the region of operating transmission power for WSNs (common operating transmission power is between 20mW to 60mW (Polastre et. al., 2004; Kohvakka et. Al., 2006; Kuorilehto et. al., 2007)). These results indicate that if we allow some delays to occur within a particular range during transmission, the cooperative SM scheme can achieve a significant energy saving. However, by relaxing the synchronisation algorithm with $0.4T_b$ jitters tolerance, the cooperative BF scheme can achieve the highest energy saving among the other schemes.

As shown in Figures 8 and 9, the SISO scheme outperforms the cooperative schemes at the transmission power region above 800mW. At the lower transmission power region, the three cooperative schemes outperform the SISO scheme. The cooperative SM scheme enjoys a lower transmission delay when the diversity gain is increasing with any arbitrary number of transmitting nodes with one condition that the number of cooperative SM receiving N nodes must be greater than the number of M nodes in cooperative BF and STBC. It also important to note that cooperative BF outperforms cooperative STBC when $M = 2$.

For imperfect synchronisation scenarios, as shown in Figures 10 and 11, at the lower transmission power region, the three cooperative schemes outperform the SISO scheme. The cooperative BF scheme enjoys lower packet latency and outperforms the other schemes even when the diversity gain of the cooperative SM scheme is increased.

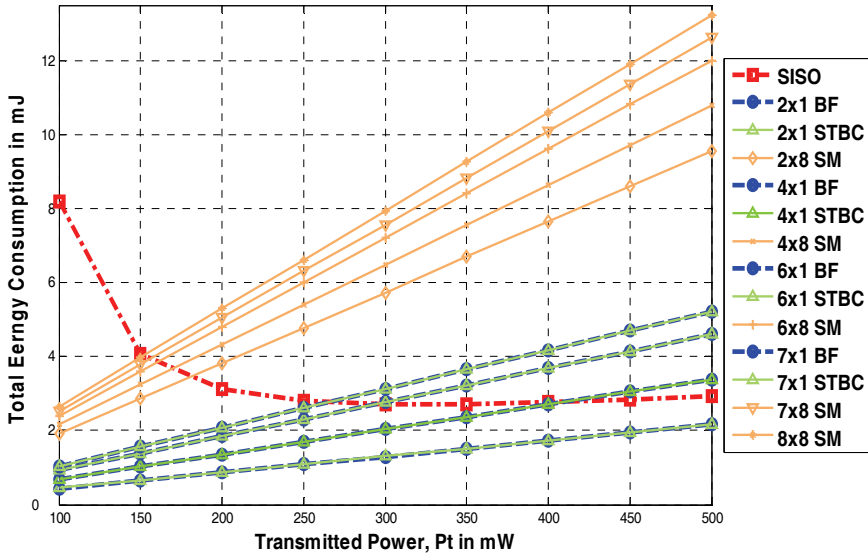


Fig. 4. Total energy consumption vs. transmission power for various schemes with $M = 2, 4, 6, 7, 8$ and $N = 1$ (Cooperative BF and STBC) and $N = 8$ (Cooperative SM).

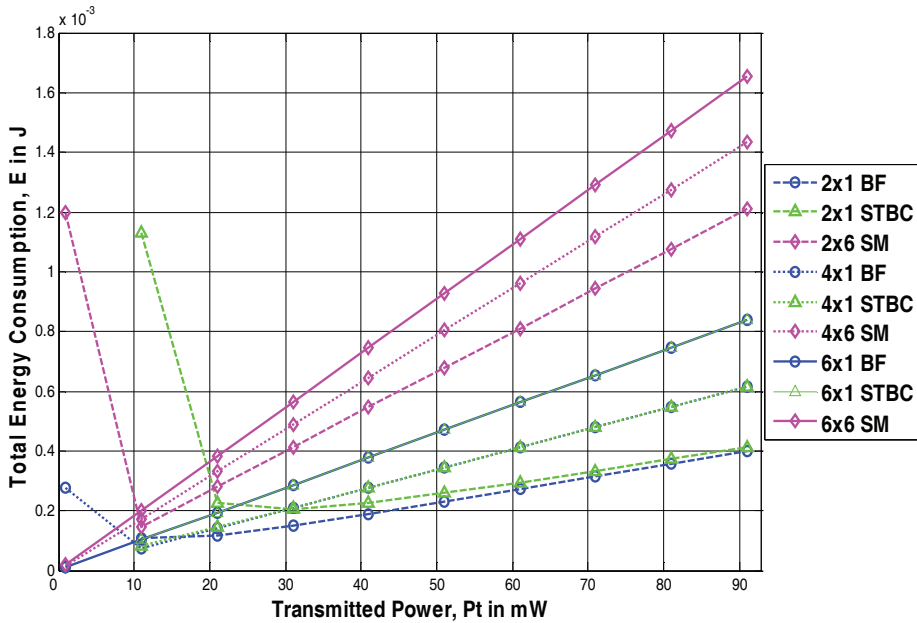


Fig. 5. Total energy consumption vs. transmission power for various schemes with $M = 2, 4, 6$ and $N = 1$ (Cooperative BF and STBC) and $N = 6$ (Cooperative SM).

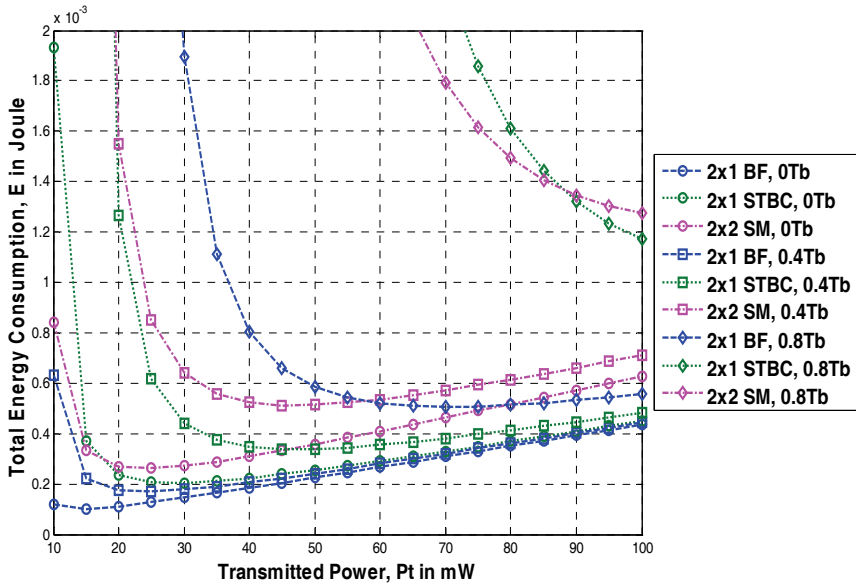


Fig. 6. Total energy consumption vs. transmission power (lower region) for various imperfect synchronisation schemes with $M = 2$ and $N = 1$ (Cooperative BF and STBC) and $N = 2$ (Cooperative SM).

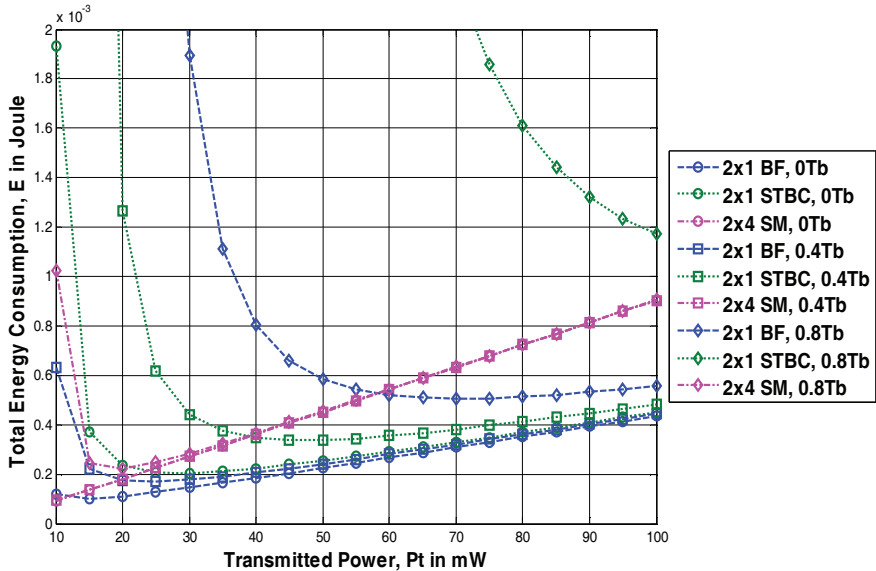


Fig. 7. Total energy consumption vs. transmission power (lower region) for various imperfect synchronisation schemes with $M = 2$ and $N = 1$ (Cooperative BF and STBC) and $N = 4$ (Cooperative SM).

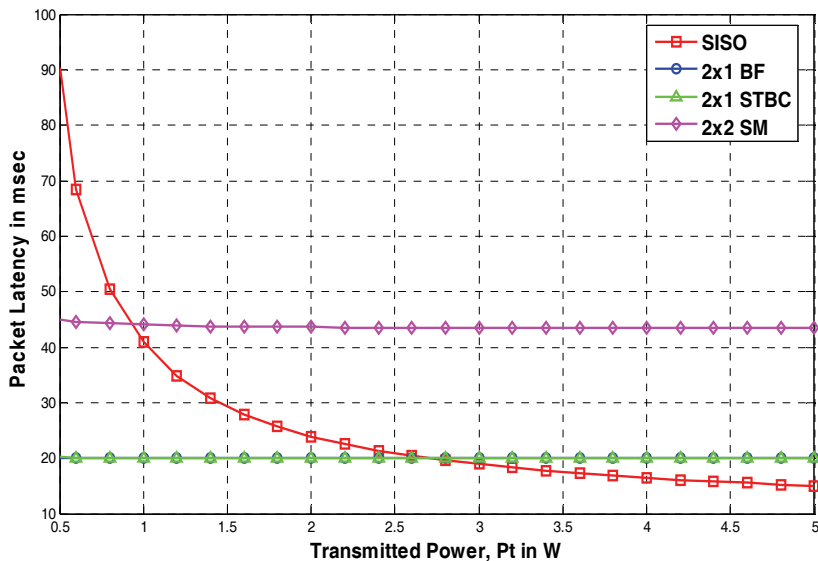


Fig. 8. Packet latency vs. transmission power for various schemes with $M = 2$ and $N = 1$ (Cooperative BF and STBC) and $N = 2$ (Cooperative SM).

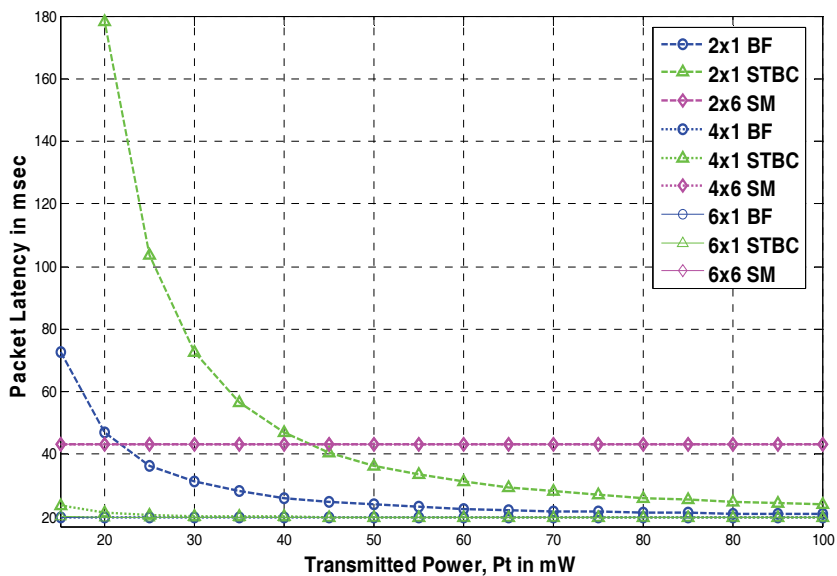


Fig. 9. Packet latency vs. transmission power for various schemes with $M = 2, 4$, and 6 and $N = 1$ (Cooperative BF and STBC) and $N = 6$ (Cooperative SM).

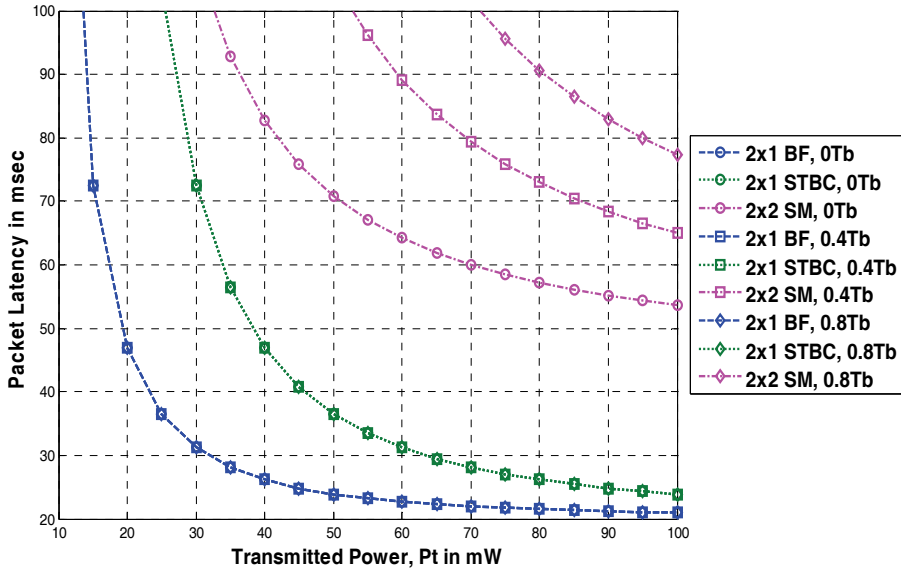


Fig. 10. Packet latency vs. transmission power (lower region) for various imperfect synchronisation schemes with $M = 2$ and $N = 1$ (Cooperative BF and STBC) and $N = 2$ (Cooperative SM).

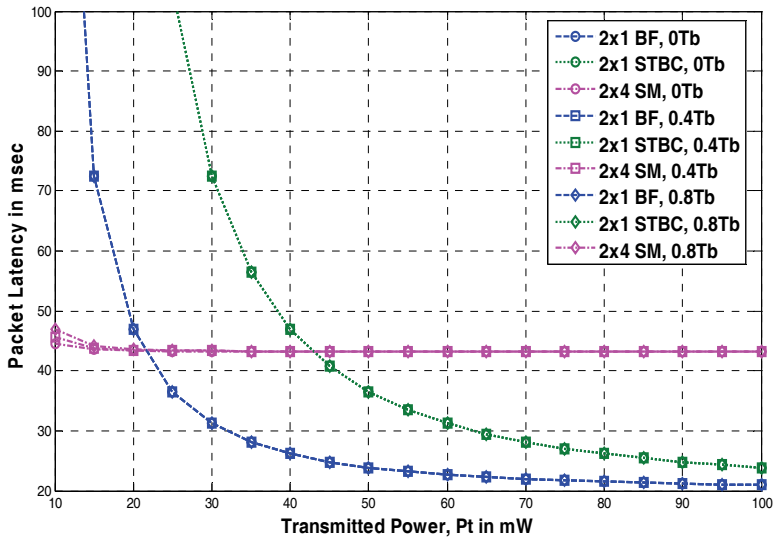


Fig. 11. Packet latency vs. transmission power (lower region) for various imperfect synchronisation schemes with $M = 2$ and $N = 1$ (Cooperative BF and STBC) and $N = 4$ (Cooperative SM).

6. Conclusion

This chapter presents a comparison study of three cooperative MIMO schemes in WSNs. We have developed analytical models for BER and PER to estimate retransmission rates from PER in (Ahmad et. al., 2008) and these are used to evaluate the total energy consumption and packet latency of the cooperative systems in this chapter. We show that the SISO scheme is more energy efficient and has lower latency at higher regions of transmission power while the three cooperative MIMO schemes are more energy efficient and outperform the SISO scheme at lower regions. Clearly, at the higher transmission power region, the SISO scheme enjoys lower transceiver circuit energy consumption and no energy cost at all on establishing a cooperative mechanism compared to the cooperative MIMO schemes. These results provide a constraint on the optimal transmission power or equivalently the optimal distance that should be used when implementing cooperative MIMO transmission in WSNs.

From the analysis we can conclude that at the lower transmission power region, the cooperative optimal BF scheme outperforms both the cooperative SM and STBC schemes in terms of energy efficiency and packet latency for both perfect and imperfect synchronisation scenarios. Also we note that the cooperative BF scheme with $M = 2$ nodes is an efficient cooperative system. Further work will involve development of MAC protocols optimised for the cooperative transmission schemes and with the aim of creating an optimal cooperative transmission mechanism for use in distributed WSNs.

7. References

- Ahmad, M.R.; Dutkiewicz, E.; & Huang, X. (2008). Performance Analysis of Cooperative MIMO Transmission Schemes in WSN, to be presented at the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Cannes, France.
- Cui, S.; Goldsmith, A.J.; & Bahai, A. (2004). Energy-efficient of MIMO and Cooperative MIMO Techniques in Sensor Networks, *IEEE Journal on Selected Areas in Communications*, vol. 22, issue 6, pp. 1089-1098.
- Kohvakka, M.; Kuorilehto, M.; Hannikainen, M. & Hamalainen, T.D. (2006). Performance Analysis of IEEE 802.15.4 and Zigbee for Large-scale Wireless Sensor Network Applications, *Proceedings of ACM International Workshop on Performance Evaluation of Wireless Ad hoc, Sensor, and Ubiquitous Networks*, pp. 1-6, Malaga, Spain.
- Kuorilehto, M.; Kohvakka, M.; Suhonen, J.; Hamalainen, P.; Hannikainen, M. & Hamalainen, T.D. (2007). MAC Protocols, In: *Ultra-Low Energy Wireless Sensor Networks in Practice*, pp. 73-88, John Wiley & Sons, 978-0-470-05786-5, West Sussex, England.
- Nguyen, T.D.; Berder, O.; & Sentieys, O. (2007). Cooperative MIMO Schemes Optimal Selection for Wireless Sensor Networks, presented at IEEE Vehicular Technology Conference (VTC2007), Baltimore, MD, USA.
- Polastre, J.; Hill, J.; & Culler, D. (2004). Versatile Low Power Media Access for Wireless Sensor Networks, presented at The ACM Conference on Embedded Networked Sensor Systems (Sensys), Baltimore, Maryland, USA.
- Yang, H.; Shen, H.-Y. & Sikdar, B. (2007). A MAC Protocol for Cooperative MIMO Transmissions in Sensor Networks, *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM)*, pp. 636-640, Washington, USA.

Single/Multi-User MIMO Differential Capacity

Daniel Castanheira and Atilio Gameiro

*University of Aveiro
(Instituto de Telecomunicações)
Portugal*

1. Introduction

This chapter will be structured around two contributions by the authors on the topic, (Castanheira & Gameiro 2008) and (Castanheira & Gameiro 2009).

The provision of broadband services to everyone is considered one of the key components for enabling the so-called information society. It is more or less consensual that to attain the high-rates envisioned by IMT-2000 (I. R. R. M. M1645 2003) of providing around 1Gbit/s for pedestrian and 100Mbit/s for high mobility, will require the use of multiple antennas at the transceivers, to exploit the scattering properties of the wireless medium. Nevertheless, due to the physical size limitations of the transceivers, the number of antennas cannot be high and the space between them is limited, which implies that the degree of channel independence achieved is not sufficient to attain the high capacities envisioned, in most scenarios. One possible solution to cope with this problem is to have the mobiles simultaneously communicating with a group of geographically distributed antennas with perfect cooperation between them. The key to achieve perfect cooperation is to have the radio signals transparently transmitted / received to / from a central unit (CU), where all the signal processing is done (FUTON 2008). Considering the high capacities envisioned optical fiber, due to its low attenuation and enormous bandwidth is the obvious technology of choice to build these transparent interconnections. However the joint processing of a group of antennas and the remote transmission of their signals to the CU will require additional processing power and will imply additional costs to the overall network. It is also expected, by the law of diminishing returns, that as more and more antennas are jointly processed the improvement in throughput will not increase linearly with the added complexity. Thus a tradeoff must be made between the costs/complexity and the number of antennas deployed. One possible way to ease the complexity problem is to use low complexity/sub-optimal schemes, like Zero-Forcing (ZF) (Caire & Shamai 2003) or Block-Diagonalization (BD) (Seijoon Shim et al. 2008) at the CU. Following this line of thought, in (Jindal 2005) and (Juyul Lee & Jindal 2007), the authors study the incurred losses, in terms of power/rate offsets, between ZF/BD and the optimal scheme, which is well known to be Dirty Paper Coding (DPC). The authors conclude that the losses are higher when the number of transmit antennas is close to the number of aggregate receive antennas. Thereafter the analysis of the gains introduced by the connection of the system users to more transmit antennas is of special importance, either to estimate a “reasonable” number of

antennas that should be jointly processed, either to give a measure for the network to check if, in a distributed antenna system (DAS), it is worthwhile to provide an additional connection to the mobiles, or to provide some guidelines for the deployment of the distributed antennas or even to compare sub-optimal schemes to the optimal scheme. In that context we define Differential CAPacity (DCAP) as the increase in ergodic sum-capacity when one additional transmit antenna is connected to the system users, to quantify the gains provided by the processing of one additional transmit antenna at the CU and analyze its behavior.

This chapter is organized as follows: section 2 presents the channel model; section 3 describes the single-user and multi-user channel capacities. In section 4 and 5 the DCAP, for the single-user and multi-user scenarios, is studied. In section 6 numerical results are provided and finally in section 7 some conclusions are drawn.

2. Channel Model

Throughout this chapter a Distributed Multiple-Input Multiple-Output (MIMO) Broadcast channel with K users, each with N receive antennas, and M transmit antennas is considered. A broadcast channel (BC) is a communication channel in which there is one sender and two or more receivers (Cover 1972). For a broadcast channel, the user k received signal, $\mathbf{y}_k (\in \mathbb{C}^{N \times 1})$ can be modeled by:

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x} + \mathbf{n}_k, \quad k = 1, \dots, K \quad (1)$$

where $\mathbf{H}_k (\in \mathbb{C}^{N \times M})$ is the user k channel matrix, $\mathbf{x} (\in \mathbb{C}^{M \times 1})$ is the transmitted signal vector with power constraint $\mathbb{E}[\mathbf{x}\mathbf{x}^H] \leq P$ and $\mathbf{n}_k (\in \mathbb{C}^{N \times 1})$ is complex white Gaussian noise with zero mean and unit variance per vector component ($\tilde{\mathcal{N}}_N(\mathbf{0}, \mathbf{I})$). $\mathbf{H}^H = [\mathbf{H}_1^H, \dots, \mathbf{H}_K^H]$ denotes the concatenation of all channels and is matrix-variate complex Gaussian distributed (Shin & Lee 2003), with zero mean and covariance $\Sigma (\in \mathbb{C}^{KNM \times KNM})$. Through the chapter it is assumed that the receiver has perfect knowledge of its own channel and the transmitter has perfect knowledge of all channels, for each channel realization. Since a DAS is considered all channel gains are independent (Σ is diagonal).

Notations: Boldface letter denote matrix-vector quantities. The operation $tr(\cdot)$, $(\cdot)^H$ and $|\cdot|$ represents the trace, the Hermitian transpose and the determinant of a matrix, respectively. By $\mathbf{A} > 0$ and \mathbf{I}_n we denote that \mathbf{A} is positive definite and an $n \times n$ identity matrix. The notation $\mathbf{a} \sim \tilde{\mathcal{N}}_p(\mathbf{0}, \Sigma)$ and $\mathbf{B} \sim \tilde{\mathcal{W}}_p(n, \Gamma)$ is used to denote that the column vector \mathbf{a} is distributed as p -variate complex Gaussian, with zero mean and covariance $\Sigma \in \mathbb{C}^{p \times p}$, and to denote that matrix $\mathbf{B} \in \mathbb{C}^{p \times p}$ is complex Wishart distributed, with n degrees of freedom and with mean Γ . The notation $\mathbf{A} \sim \mathbf{B}$ denotes that matrix \mathbf{A} and \mathbf{B} are identically distributed.

3. Single/Multi-User Sum-Capacity

3.1 Single-User case

For the single-user case, if only one receive antenna is considered, the system reduces to the classical multiple input single output (MISO) system. As a consequence, the resulting capacity expression are much simpler to analyze than for the general case of a multi-user system. Thus for this case, the system capacity and the corresponding DCAP is described in more detail, either to introduce the reader to the topic, either to provide some baseline work to be used in the multi-user case.

According to (Liu & Li 2005a) and (Goldsmith 2005), if only one user with one receive antenna is considered ($K = N = 1$), the ergodic channel capacity can be expressed by:¹

$$C = \mathbb{E}_H \left[\log \left| \mathbf{I}_1 + \frac{\mathbf{h}\mathbf{Q}\mathbf{h}^H}{\sigma^2} \right| \right] \quad (2)$$

where $\mathbf{h} = [h_1, h_2, \dots, h_M]$ is the channel matrix, \mathbf{Q} is the transmit signal covariance matrix and σ^2 is the noise variance. Subject to a power constrain P , $\text{Tr}(\mathbf{Q}) = P$, \mathbf{x} must be circularly symmetric complex Gaussian (Telatar 1999) and its correlation matrix, \mathbf{Q} , must be diagonal² (Visotsky & Madhow 2001), which is equivalent to independent transmit signals, for the ergodic channel capacity, C , to be maximal.

Taking into account that \mathbf{Q} must be a diagonal matrix, $\mathbf{Q} = \text{diag}([P_1, P_2, \dots, P_M])$, where P_i is the signal x_i mean transmit power, and knowing that a single antenna user is considered, $N = 1$, the ergodic capacity formula reduce to:

$$C_M = \mathbb{E}_H \left[\log \left(1 + \sum_{i=1}^M \frac{|h_i|^2 P_i}{\sigma^2} \right) \right] \quad (3)$$

From the previous expression it is easy to identify $\frac{|h_i|^2 P_i}{\sigma^2}$ as the link i SNR (γ_i), then the capacity expression can be put into the following equivalent format:³

$$C_M = \mathbb{E}_{\Gamma_M} [\log(1 + \gamma)] = \int_0^\infty \log(1 + \gamma) f_{\Gamma_M}(\gamma) d\gamma \quad (4)$$

where Γ_M is a random variable (RV) corresponding to the sum of M exponential distributed RVs with mean λ_i^{-1} each, which follow the pdf:

$$f_{\Gamma_M}(\gamma) = \sum_{i=1}^L \sum_{n=1}^{K_i} \frac{a_{in}}{(n-1)!} \gamma^{n-1} e^{-\lambda_i \gamma} \quad (5)$$

¹ We consider for now on that the capacity units are nats/s/Hz, when omitted.

² The covariance matrix of the channel gains is diagonal, because the channel gains are independent. So their unitary singular value decomposition matrices are equal to the identity matrix.

³ According to (Goldsmith 2005) the $|h_i|^2$ random variable is exponential distributed and consequently γ_i .

where L is the number of different mean SNR's, λ_i^{-1} is the mean SNR of link i , K_i is the number of antennas with SNR λ_i^{-1} , a_{in} are constants related to the partial fraction expansion (Ghosh 2005) of the product of L Erlang distribution characteristic functions (Gubner 2006) and $M = \sum_{i=1}^L K_i$. Finally, after the integral evaluation, (Dohler et al. 2006) or (Gradshteyn & Ryzhik 1994), the ergodic channel capacity is given by:

$$C_M = \sum_{i=1}^L \sum_{n=1}^{K_i} \sum_{k=0}^{n-1} \frac{a_{in}}{\lambda_i^n} C_i(\lambda_i, k) \quad (6)$$

where

$$C_i(\lambda_i, k) = \frac{(-\lambda_i)^k}{k!} \left[C_0(\lambda_i) + u(k-1) \sum_{p=1}^k \frac{(p-1)!}{(-\lambda_i)^p} \right]$$

$$C_0(\lambda_i) = e^{\lambda_i} E_1(\lambda_i)$$

$C_0(\lambda_i)$ is equal to the capacity of the link associated with a single transmit antenna with SNR λ_i^{-1} and is also equal to $C_i(\lambda_i, 0)$. $E_1(x)$ denotes the exponential integral function, given by $E_1(x) = \int_x^\infty e^{-t}/t dt$ and $u(n)$ is the unit step function.

3.2 Multi-User case

In this section we briefly describe the broadcast channel sum-capacity and an affine approximation to this sum-capacity, in the high SNR regime. This approximation will be used in the following sections to analyze the DCAP for the multi-user case.

The broadcast channel capacity was over a large number of years an active area of research, which has culminated with an article by Weingarten et al., (Weingarten et al. 2006), where the authors have shown that the DPC rate region is the capacity region of the Gaussian MIMO BC. For the multi-user case the capacity expression is not as easy to analyze as the one for the single-user case, due to some additional constraints in the transmit covariance matrix and from the fact that there is no known closed form solution for this matrix that maximizes the channel capacity (Jindal et al. 2005).

From the MIMO BC-MAC duality, (Vishwanath et al. 2003), (Viswanath & Tse 2003), (Wei Yu & Cioffi 2004), the DPC sum rate with a total power constraint P , ($\sum_k \text{tr}(\mathbf{Q}_k) \leq P$), can be expressed, by:

$$C_{DPC}^{(M,K,N)} = \max_{\sum_k \text{tr}(\mathbf{Q}_k) \leq P} \log \left| \mathbf{I}_M + \sum_{k=1}^K \mathbf{H}_k^H \mathbf{Q}_k \mathbf{H}_k \right| = \max_{\text{tr}(\mathbf{Q}) \leq P} \log |\mathbf{I}_M + \mathbf{H}^H \mathbf{Q} \mathbf{H}| \quad (7)$$

where $\mathbf{Q}_k (\in \mathbb{C}^{N \times N})$ is the transmit covariance matrix in the dual MAC channel and $\mathbf{Q} = \text{diag}(\mathbf{Q}_1, \dots, \mathbf{Q}_K)$ is a block-diagonal matrix. In the high SNR regime and for $KN \leq M$, equal power allocation, $\mathbf{Q}_k = \frac{P}{KN} \mathbf{I}_N$, is asymptotically optimal (Juyul Lee & Jindal 2007). As a byproduct of that, the following affine approximation to the sum-capacity can be made (Juyul Lee & Jindal 2007), (Shamai & Verdú 2001):

$$\mathbf{C}_{DPC}^{(M,K,N)} \approx KN \log P - KN \log KN + \log |\mathbf{H}\mathbf{H}^H| \quad (8)$$

From this affine approximation it is easy to see that the sum-capacity scales linearly with the number of aggregate receive antennas and that this scaling factor is not dependent on the individual values of K or N , in the high SNR regime. However, for the case of more aggregate receive than transmit antennas, $KN > M$, equal power allocation ceases to be asymptotic optimal, since even for the standard degraded BC ($M = 1$) the throughput is maximized by transmitting only to the best user (Caire & Shamai 2003).

4. Single-User DCAP

4.1 Exact Expression

In this sub-section we obtain an exact expression for the DCAP considering the single-user scenario. During our analysis of the DCAP behavior it will be observed that the DCAP tends to have a small decay relatively to the previous DCAP value, when the new added antenna has the same mean SNR of a previously connected antenna. In that context we also derive an expression for the DCAP that contains explicitly the value of the previous DCAP to, latter on, do an analysis of the sensitivity of the DCAP values versus the mean SNR of the new connected antenna, as more and more antennas are connected to the system.

From the definition of the Γ_M RV it is possible to prove that:

$$f_{\Gamma_M}(\gamma) = \lambda_M e^{-\lambda_M \gamma} \int_0^\gamma f_{\Gamma_{M-1}}(x) e^{\lambda_M x} dx \quad (9)$$

with which a recursive algorithm for the calculation of the a_{in} coefficients, which are central to the calculation of the capacity values, equation (6), can be obtained, and the following formula can be derived:

$$-\frac{df_{\Gamma_M}(\gamma)}{d\gamma} = \lambda_M [f_{\Gamma_M}(\gamma) - f_{\Gamma_{M-1}}(\gamma)] \quad (10)$$

From equation (10), after multiplying by $\log(1 + \gamma)$ and integrating from zero to plus infinity the γ variable, at both sides of the equation, the DCAP can be expressed by:

$$\begin{aligned} \Delta C_{M-1}^M &= C_M - C_{M-1} = -\lambda_M^{-1} \int_0^\infty \frac{df_{\Gamma_M}(\gamma)}{d\gamma} \log(1 + \gamma) d\gamma \\ &= \lambda_M^{-1} \int_0^\infty \frac{f_{\Gamma_M}(\gamma)}{1 + \gamma} d\gamma = \lambda_M^{-1} \sum_{i=1}^L \sum_{n=1}^{K_i} \frac{a_{in}}{\lambda_i^{n-1}} C_i(\lambda_i, n-1) \end{aligned} \quad (11)$$

With some more mathematical manipulations, the differential capacity can also be equivalently expressed by:

$$\begin{aligned} \Delta C_{M-1}^M &= \frac{\lambda_{M-1}}{\lambda_M} \Delta C_{M-2}^{M-1} \\ &\quad + \frac{1}{\lambda_M^2} \left[C_0(\lambda_M) f_{\Gamma_M}(0) - \int_0^\infty \frac{f_{\Gamma_M}(\gamma)}{(1 + \gamma)^2} d\gamma \right] \end{aligned} \quad (12)$$

where $f_{\Gamma_M}(0)$ is equal to 0 for $M > 1$ and equal to λ_1 for $M = 1$. This expression will be used in the following sections to analyze the DCAP sensitivity with the number of transmit antennas and the mean SNR's of the new connected antennas.

4.2 DCAP Bounds

Even if the previous expressions obtained for the DCAP, namely equation (11) and (12), are exact they are not easy to analyze, mostly because of the presence of the a_{in} coefficients. Therefore the derivation of some simple upper/lower bounds to the DCAP is an important step to analyze its behavior. For doing that, we will rely mostly on the second integral definition of the DCAP, from equation (11). From that expression one can easily see that if some bounds for equation $g(\gamma) = 1/(1 + \gamma)$ are available, they can be easily used to bound also the DCAP. One simple bound for equation $g(\gamma)$ is $g(\gamma) \leq 1$, for all γ in $[0, \infty[$. As a consequence the DCAP can be upper bounded by:

$$\Delta C_{M-1}^M \leq \frac{1}{\lambda_M} \int_0^\infty f_{\Gamma_M}(\gamma) d\gamma = \lambda_M^{-1} \quad (13)$$

and the channel capacity can be upper bounded by:

$$C_M \leq \sum_{k=1}^M \lambda_k^{-1} = \sum_{i=1}^L K_i \lambda_i^{-1} \quad (14)$$

Another possible bound for equation $g(\gamma)$ can be obtained from $1 + x < e^x$, for all x in $[0, \infty[$, by the Taylor series expansion of the exponential function around zero. Thus $g(\gamma) \leq e^{-\gamma}$, for all γ in $[0, \infty[$, with a maximum difference of $(M_d \approx 0.204)^4$, so:

$$e^{-\gamma} \leq g(\gamma) \leq e^{-\gamma} + M_d \quad (15)$$

and as a byproduct:

$$\frac{\varphi_{\Gamma_M}(-1)}{\lambda_M} \leq \Delta C_{M-1}^M \leq \frac{\varphi_{\Gamma_M}(-1)}{\lambda_M} + \frac{M_d}{\lambda_M} \quad (16)$$

where $\varphi_{\Gamma_M}(s)$ is the Γ_M RV moment generating function.

From bound (13) and (16) follows that in the low SNR regime the DCAP can be approximated by λ_M^{-1} and the channel capacity by:

$$C_M \approx \sum_{i=1}^L K_i \lambda_i^{-1} \quad (17)$$

The previous bounds although simple can provide some interesting information on the behavior of the DCAP, as will be seen in section 6.1. However they do not give any idea on the SNR vector that attain the maximum of the DCAP. This vector can give information on how the distributed antennas should be geographically positioned to attain most of the gains provided by the connection of additional antennas to the system. In the following paragraphs we answer this question, and prove that the SNR vector that attains the highest

⁴ Obtained numerically, and knowing that this maximum is global.

DCAP value is the SNR vector where all elements are equal. In other words the highest DCAP is obtained when all transmit antennas are co-located, have the same link SNR

From equation (11) one can see that $d\Delta C_{M-1}^M/d\lambda_i \geq 0$ for $M > 1$ and $i \neq M$. Therefore if we assume that the new connected antenna is always the one with the highest mean SNR, the DCAP will be upper bounded by the one where all mean SNRs are equal to the mean SNR of the new connected antenna:

$$\begin{aligned} \Delta C_{M-1}^M(\lambda_1, \dots, \lambda_M) &\leq \Delta C_{M-1}^M(\lambda_M, \dots, \lambda_M) \\ \lambda_M &\geq \lambda_n, \forall n \in [1, 2, \dots, M-1] \end{aligned} \quad (18)$$

When the mean SNR tends to infinity the maximum of the DCAP is obtained and is equal to:

$$\lim_{\lambda_M \rightarrow \infty} \Delta C_{M-1}^M(\lambda_M, \dots, \lambda_M) = \frac{1}{M-1} \quad (19)$$

Thus the DCAP in general is upper bounded by the co-located transmit antennas case and its maximal value is only dependent on the number of transmit antennas and not on the actual values of the mean SNR of the links. However this limit is not attainable in practice. But, how far is it from a real scenario? It can be shown that for a moderate SNR of 17 dB, for all links, the difference is lower than 0.1 bps/Hz when we pass from one connected antenna to two and even lower for the other connected antennas.

Since the previous bound is a limit bound it can be not very tight. However a tighter bound can be developed considering only a slight higher degree of information:

$$\Delta C_{M-1}^M \leq \frac{1}{\lambda_M + M - 1} \leq \frac{1}{M - 1} \quad (20)$$

From the previous expression one can also upper bound the capacity increase by the connection of J new antennas, considering that the user is actually only connected to one antenna and that antenna has the greatest SNR of all of them, by:

$$\Delta C_1^{J+1} \leq \sum_{n=2}^{J+1} \frac{1}{n-1} = \sum_{n=1}^J \frac{1}{n} \approx \gamma + \log(J + \beta) \quad (21)$$

where γ is the Euler constant (0.577215667) and $\beta = e^{1-\gamma} - 1$. Consequently as the number of connected antennas increase the DCAP decreases. However, if the number of antennas grows to infinity so does the capacity, but of course, in practice, one can only have a limited number of antennas. Thus a tradeoff between the cost and capacity gains must be made when we choose the number of antennas to be deployed in a real system.

5. Multi-User DCAP

In the previous section we have analyzed the DCAP for the single-user case and have provided an exact closed form expression for it, that is valid for all mean SNR's. We have started with the single user case due to its simplicity and also because the analysis of that simpler case can give some insight for the analysis of the more difficult case of a multi-user

system. However for the more general case of a multi-user system the analysis of the DCAP over all mean SNR's is intractable. Thus, for this case, we consider only the DCAP in the high SNR regime. In this section we first define the DCAP and after derive a closed form expression for it, for the co-located antenna case and for $M \geq KN$. For $KN > M$ the DCAP is analyzed numerically.

According to equation (8), for $M \geq KN$, the DCAP can be expressed by:

$$\begin{aligned} \Delta C_{M,K,N}^{M+1,K,N} &= \mathbb{E}[\mathbf{C}_{DPC}^{(M+1,K,N)} - \mathbf{C}_{DPC}^{(M,K,N)}] \\ &= \mathbb{E}[\log|\bar{\mathbf{H}}_{M+1}\bar{\mathbf{H}}_{M+1}^H|] - \mathbb{E}[\log|\bar{\mathbf{H}}_M\bar{\mathbf{H}}_M^H|] \end{aligned} \quad (22)$$

where $\bar{\mathbf{H}}_i$, ($i = M, M+1$) is the channel matrix for i transmit antennas. Thus for the multiuser scenario the DCAP is only dependent on the matrix \mathbf{H} distribution. On the other hand, for $KN > M$, the DCAP can be expressed by:

$$\begin{aligned} \Delta C_{M,K,N}^{M+1,K,N} &= \mathbb{E}[\mathbf{C}_{DPC}^{(M+1,K,N)} - \mathbf{C}_{DPC}^{(M,K,N)}] \\ &= \mathbb{E}[\mathbf{C}_{DPC}^{(M+1,M+1,1)} + \Delta C_{M+1,M+1,1}^{M+1,K,N} - (\mathbf{C}_{DPC}^{(M,M,1)} + \Delta C_{M,M,1}^{M,K,N})] \\ &= \Delta C_{M,M,1}^{M+1,M+1,1} + \Delta C_{M+1,M+1,1}^{M+1,K,N} - \Delta C_{M,M,1}^{M,K,N} \\ &= \log(P) + \Delta \end{aligned} \quad (23)$$

From which we can see that the connection of the system users to a new transmit antenna is equal to $\log(P)$, the multiplexing gain, plus a Δ factor, that is equal to $\Delta C_{M+1,M+1,1}^{M+1,K,N} - \Delta C_{M,M,1}^{M,K,N}$ plus the power offset gain provided when we pass from M to $M+1$ transmit antennas and users⁵, which can be closely approximated by $-\log(e)$, for any number of transmit antennas. Hence, the significance of the Δ factor versus the multiplexing gain must be analyzed to have a clear picture of the full obtained gain, please refer to section 5.2 and 6.2.

5.1 DCAP for $M \geq KN$

In this sub-section the DCAP for the optimal scheme, DPC, and for two sub-optimal schemes, namely ZF and BD is analyzed. A closed form expression and an approximation for it are also derived. Unfortunately, the general case of a distributed antenna system is difficult to analyze mathematically, thus for this case we will rely on numerical analysis, which are provided in section 6.2. In the analysis, contained in this sub-section, we consider that all transmit antennas are co-located.

For DPC, as shown in appendix (sub-section 9.1), the DCAP when we pass from M to $M+1$ transmit antennas is given by:

$$\begin{aligned} {}_5 \Delta C_{M,M,1}^{M+1,M+1,1} &\approx \log(P) + \log(M + \beta) + M \log M - (M+1) \log(M+1) \\ &\approx \log(P) - \log(e) \end{aligned}$$

$$\Delta C_{M,K,N}^{M+1,K,N} = \sum_{n=M-KN+1}^M \frac{1}{n} = \sum_{n=M-KN+1}^M \Delta C_{n,1,1}^{n+1,1,1} \approx \log \left(1 + \frac{KN}{M-KN+\beta} \right) \quad (24)$$

where $\beta = e^{1-\gamma} - 1$ and $\gamma = 0.577215665$ is the Euler constant. For the linear precoding schemes a similar expression can be obtained, by noting that for BD the DCAP is equal to K times the DPC one, for a $(M - (K - 1)N) \times N$ channel, from the BC equivalent point to point MIMO interpretation, in the high SNR regime (Juyul Lee & Jindal 2007). For that reason, the DCAP for ZF/BD can be show to be equal to:

$$\overline{\Delta C}_{M,K,N}^{M+1,K,N} \approx K \log \left(1 + \frac{N}{M-KN+\beta} \right) \quad (25)$$

From the previous expression one can see that asymptotically in the number of transmit antennas, M , the DCAP behaves like $\log_2(e)KN/(M - KN + \beta)$, for both DPC and ZF/BD, implying that the difference between the gains of the optimal scheme and the sub-optimal schemes is very small, when a high number of transmit antennas is considered. Nevertheless for a finite number of transmit antennas the DCAP gains provided by DPC are lower and the difference increases with K , as can be seen by equations (24) and (25). But how it scales with K , the number of users? It can be seen that for $M - KN$ a constant the DCAP increases logarithmically with K for DPC and linearly with K for ZF/BD. Thus one can conclude that even if the optimal scheme has a higher complexity, it will need a much lower number of transmit antennas than the sub-optimal schemes, which require a lower complexity, for the same target capacity. Thus the number of antennas to be jointly processed must be carefully chosen, taking into account the tradeoff between complexity, network costs and obtained benefits.

5.2 DCAP for $KN > M$

In this sub-section the DCAP is analyzed for the case of more aggregate receive than transmit antennas. For this case, since equal power allocation stop to be asymptotically optimal and no closed form solution exists for the power allocation problem (Jindal et al. 2005), a mathematical study to attain a closed form expression for the DCAP is not possible. Nevertheless an approximation for this DCAP can be obtained. But how? It is know that for the case of only one transmit antenna, $M = 1$, it is optimal to transmit only to the best user (Caire & Shamai 2003) and as proven in the appendix (sub-section 9.2), for this case the maximum increase in capacity, when one additional user is connected to the system, is attained when all users are co-located i.e. when they have equal mean link SNR's. For $M = 1$ and $N = 1$ and for the high SNR regime this maximum is equal to, see appendix (sub-section 9.2):

$$\Delta C_{1,K,1}^{1,K+1,1} = \sum_{n=0}^K \binom{K}{n} (-1)^{K+1} \log(K+1) \quad (26)$$

At this point we have only obtained an expression for the capacity increase by the connection of one additional system user, for the case of only one transmit antenna. But to

obtain an approximation for the DCAP, for $KN > M$, we need also to know that for $M > 1$, see equation (23). For $M > 1$ the increase in capacity obtained by the connection of a new user to the system, can be approximated taking into account the DCAP expression for the $M > KN$ case, equation (24) line 1, and making an extrapolation of equation (26). From those assumptions the capacity increase by the connection of one additional user to the system can be approximated by:

$$\Delta C_{M,K,1}^{M,K+1,1} \approx \sum_{n=K-M+1}^K \Delta C_{1,n,1}^{1,n+1,1} \quad (27)$$

and as can be confirmed in sub-section 6.2 (Fig. 3(a)), this approximation is close to the values obtained by numerical simulations and is better for a high number of users.

6. Numerical Results

6.1 Single-User DCAP analysis

In this sub-section, the DCAP, for a grid antenna placement, as show in Fig. 3 (a), is analyzed. First the exact DCAP values are compared to the obtained bounds, to access their tightness. Next we analyze the sensitivity of the DCAP to the variation of the mean SNR and finally we perform a DCAP analysis of a representative area covered by the antennas.

In all numerical analysis presented in this chapter we consider that the mean SNR is only dependent on the signal path loss (Simplified Path Loss Model, (Goldsmith 2005)), that⁶ $P_t K d_0^\gamma / \sigma_n^2 = 1$ and that $\gamma = 3$. It is also considered that the new connected antenna is always the one with highest mean SNR from the group of unconnected antennas and that $\Delta d_x = \Delta d_y = \Delta d$. By circular ring we mean the group of antennas with the same SNR.

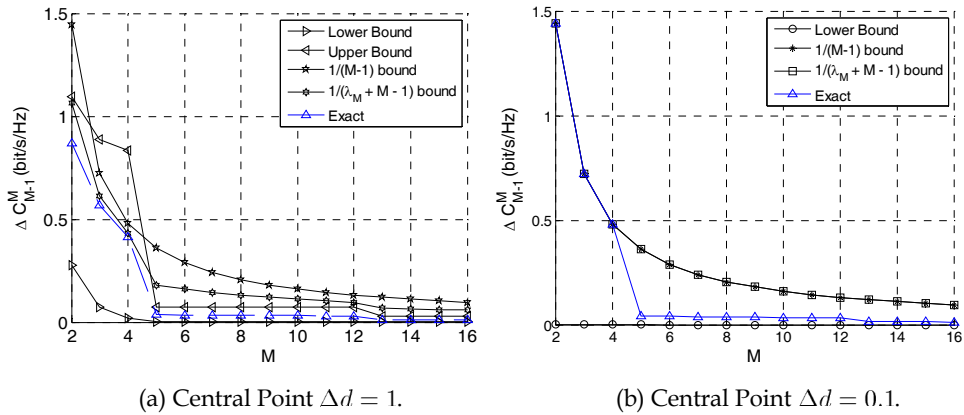


Fig. 1. Differential capacity by the connection to one more antenna, exact, upper and lower bounds.

⁶ K is a unit less constant which depends on the antenna characteristics and on the average channel attenuation, d_0 is a reference distance for the antenna far-field, γ is the path loss exponent and P_t is the transmitted power at a distance d_0 .

In Fig. 1 we plot the exact DCAP values and respective bounds for the central point of the grid antenna placement and do that for two different inter-antenna distances, namely $\Delta d = 1$ and $\Delta d = 0.1$. From this figure one can see that if the new connected antenna is in the same circular ring as a previously connected antenna then its DCAP value, ΔC_{M-1}^{M-1} , will be approximately the same as the previous one, ΔC_{M-2}^{M-1} . On the other hand if the new connected antenna is far away from the circular ring of a previously connected antenna then the DCAP value decreases a lot. This approximation is better when a high number of transmit antennas is considered. Thus one can conclude intuitively that if all antennas are in the same circular ring, have the same mean SNR, their DCAP values will be maximum. Indeed this is true as shown in section 4.2. This behavior can be explained by the fact that ΔC_{M-1}^M is dependent on ΔC_{M-2}^{M-1} by a factor of λ_{M-1}/λ_M , an approximation that due to its importance will be analyzed in the following paragraphs. It can also be explained by the fact that in equation (16), $\varphi_{\Gamma_M}(-1)$ tends to zero as we connect to more antennas and as a consequence the bound becomes independent of all SNR's except the new one, which for a circular ring is constant. Thus, since the central point in the grid antenna placement is the one with highest symmetry we can say that if a target increase in capacity is pre-established then the user will be connected to 4 or 12 or 16 or 24 or 32 or . . . antennas, depending on the target increase in capacity defined.

Concerning the bounds tightness we can see from that figure that for low SNR's the upper bound provided by equation (16) is the better one, but for high SNR's the bound from equation (20) is more accurate. Although, the user will only connect to a small number of antennas in a real system, due to the diminishing returns expected as the number of antennas increase. Thus, taking into account this fact, the most important bounds will be the ones that are tighter for a low number of transmit antennas, and in this case the winner is the bound from equation (20).

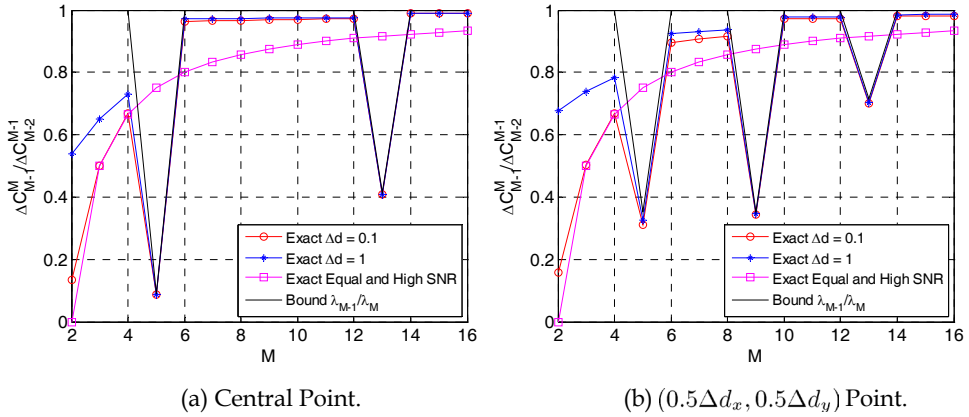


Fig. 2. Differential capacity sensitivity with respect to M and for different inter-antenna distances.

From Fig. 1 (a) and Fig. 1 (b) one can also see that the exact values of the DCAP, the dashed and solid blue lines, converge to the same value, for a high number of transmit antennas in the case of SNR vectors that are multiple among themselves. This fact can be easily proven

together with the fact that as the SNR's get higher this convergence occurs at a smaller M value. In the case of high SNR's the DCAP cannot be higher than a given value, having as critical value $1/(M-1)$, in the case of equal SNR's. As a consequence, if all transmit antennas are put closer to the user terminal, by a given factor, the only antennas for which the DCAP increases are the ones near the terminal.

As previously verified the DCAP in a circular ring tends to be constant. But how close to the previous value and when this approximation can be made? In the following paragraphs we propose to analyze this fact using a bound for the ratio $\Delta C_{M-1}^M / \Delta C_{M-2}^{M-1}$. If the ratio is close to one then the ΔC_{M-1}^M value will be close to ΔC_{M-2}^{M-1} . In Fig. 2 (a) and Fig. 2 (b) we plot this ratio for two different points in the grid antenna placement, namely for the central point and for $(0.5\Delta d_x, 0.5\Delta d_y)$. In each figure the exact ratio value for two inter-antenna distances ($\Delta d = 1$ and $\Delta d = 0.1$), the bound from equation (12) and the exact value for equal and high SNR's are plotted. From those results one can view that as the number of connected transmit antennas increase the ratio sensitivity to the SNR vector variation decreases and that its exact value becomes closer to the bound. One can also see that when the new connected antenna is not from the same ring as the previously connected antenna, the bound is indeed very close to the actual value of the respective ratio and the ratio value as a step decrease that is lower as more connected antennas are considered. As a consequence the DCAP tends to be constant in a circular ring, as previously observed, and the approximation is better when a high number of transmit antennas is considered.

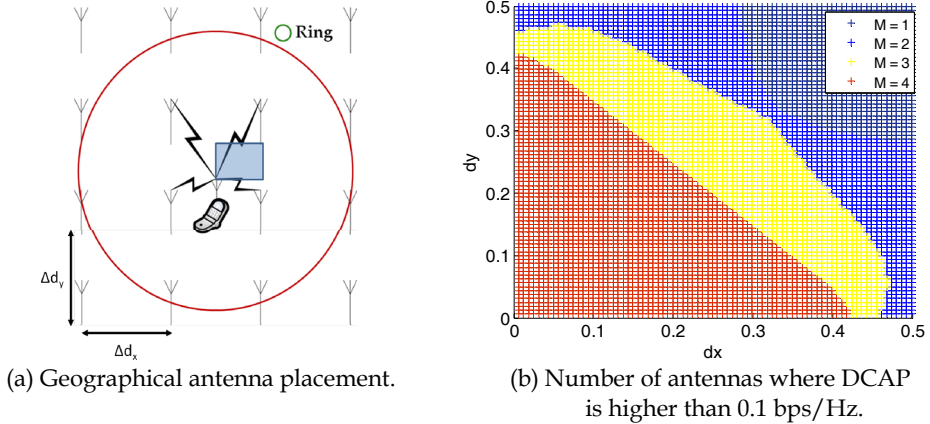


Fig. 3. Single-User differential capacity analysis, for a regular antenna placement.

Thus far, we have only analyzed the DCAP values for a single point, but how it will behave if we consider a given area. The analysis of the DCAP values for a given area cannot be easily visualized in a two dimensional plot, thus we have relied on a different measure to analyze its behavior. The metric considered was the maximum number of transmit antennas that achieve a target DCAP value or more. In this study we have considered a target DCAP value of 0.1 bps/Hz. The geographical area considered was the one shown in figure Fig. 3 (a) in light blue, and we have also considered $\Delta d_x = \Delta d_y = 1$. The results of this study are presented on figure Fig. 3 (b). To explain the information contained into this figure it is

easier to give an example. Thus, let's assume, for example, the point (0.4, 0.4). For this point $M = 4$, thus when we connect the second, third and fourth antennas the DCAP value is higher than the target, but departing from that number of transmit antennas, 4, the DCAP will be lower than the target.

For the considered scenario and for a target DCAP of 0.1bps/Hz only the first four antennas will be connected to the user, or equivalent the antennas presented in the first ring. This figure also shows a circular pattern in the number of necessary antennas. This is related to the fact that at a given distance the link mean SNR is only dependent on the distance from the user to the considered transmit antenna.

6.2 Multi-User DCAP analysis

In this sub-section the DCAP of DPC is compared to the one of ZF/BD, for the co-located antennas scenario. For the DAS, as stated before, we rely on numerical simulations to access the capacity gains provided by the connection of additional transmit antennas. For this numerical analysis we consider a scenario with 4 transmit antennas and 2 users, each with only one receive antenna. Finally for $KN > M$ we analyze the DCAP for the equal mean SNR's case. But to do that we first analyze the capacity gain provided by the connection of an additional user to the system and after that we investigate the gain provided by Δ in comparison to $\log(P)$, see equation (23).

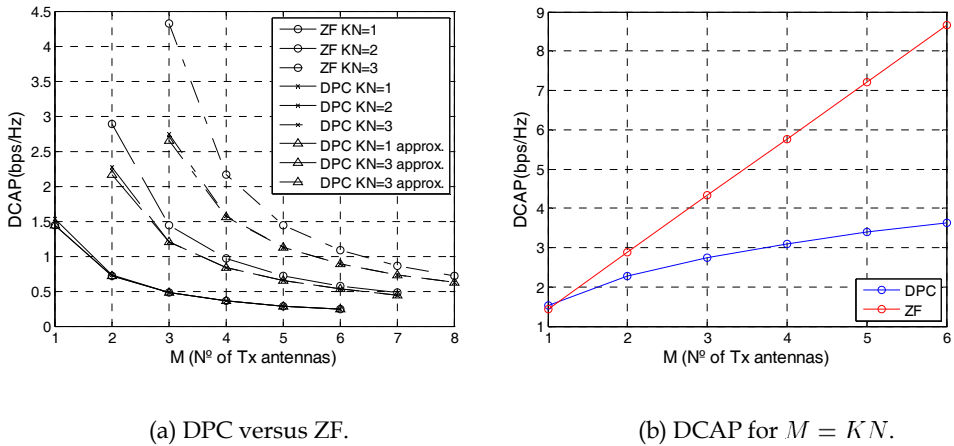
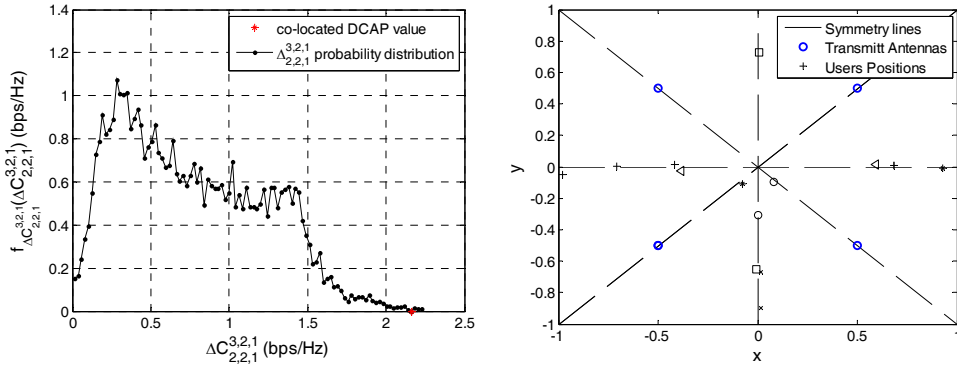


Fig. 4. Differential capacity for different number of user's and receive antennas.

As can be seen in Fig. 4 (a), were we present a plot of the DPC and ZF DCAP versus the number of connected transmit antennas, for the co-located scenario, the logarithm approximation although simple is very tight. Concerning the benefits of the connection of new transmit antennas, ZF has higher gains than DPC and the difference increase with KN . To see how this difference scales with K we plot in Fig. 4 (b) the DCAP values versus M for $M = KN$. From this figure one can view that the DCAP scales logarithmically with the number of users for DPC and linearly for ZF, as already seen from equations (24) and (25). Thus even if the implementation complexity of DPC is much higher than the one for ZF, the sub-optimal scheme, ZF, will need a higher number of transmit antennas than the optimal

scheme for a target system capacity. Thus a tradeoff must be made between the cost of additional processing power and the cost of additional physical resources, namely transmit antennas and respective connection to the central unit. As a result of this analysis, as a scheme approaches the optimal one, the gains obtained by the connection of additional antennas at the transmitter side decrease. But, remember that the optimal scheme will have always a higher sum-capacity. As seen before, from equations (24) and (25), the DCAP for the two precoding schemes converge to the same value for a high number of transmit antennas and the convergence point increases with KN .



(a) DCAP distribution for $M = 2$.

(b) User's position, where distributed DCAP is higher than co-located DCAP.

Fig. 5. Differential capacity distribution for a uniform user distribution and user's positions where the distributed DCAP is higher than the co-located DCAP. Each pair of equal black markers represent the positions of each user.

For the DAS we will only analyze the DCAP for DPC and that analysis will rely on numerical simulations. For this simulation a scenario with 4 transmit antennas and two users, each with only one receive antenna, is considered. To model the channel we have considered the Simplified Path Loss Model again, (Goldsmith 2005), with a path loss exponent equal to 3, and have considered Rayleigh multipath fading. The transmit antennas were positioned as shown in Fig. 5 (b) and the user positions were randomly generated with an uniform distribution in $[-1, 1] \times [-1, 1]$. Ten thousand user's positions were generated and the DCAP was averaged over 10000 trials. The new connected antennas were always chosen to be the one that will imply the highest DCAP from the group of unconnected antennas. As a result of this simulation, we have Fig. 5, where the co-located DCAP, equation (24), value is represented by a red star. From that figure we see that most of the user's positions have a smaller DCAP than the co-located case. However for a small elite of users, 7 in this specific case, the DCAP is higher than the one obtained by the co-located scenario. Thus one can conclude that even if the DCAP for a DAS can be higher than the one for the co-located case the difference will not be high and the number of user positions in that condition will be low. To have a better understanding of the positions that attain the maximum DCAP gains for a DAS, in Fig. 5 (b), we show the 7 positions that have a higher DCAP than the co-located case. From that figure it is possible to infer that these positions

are all very close to the system symmetry lines defined by the four transmit antennas and in that way, as in the co-located case, the new connected antenna will have the same distribution as a previous one. Thus system symmetry plays an important role in obtaining most of the DCAP gains, either for the single-user either for the multi-user cases.

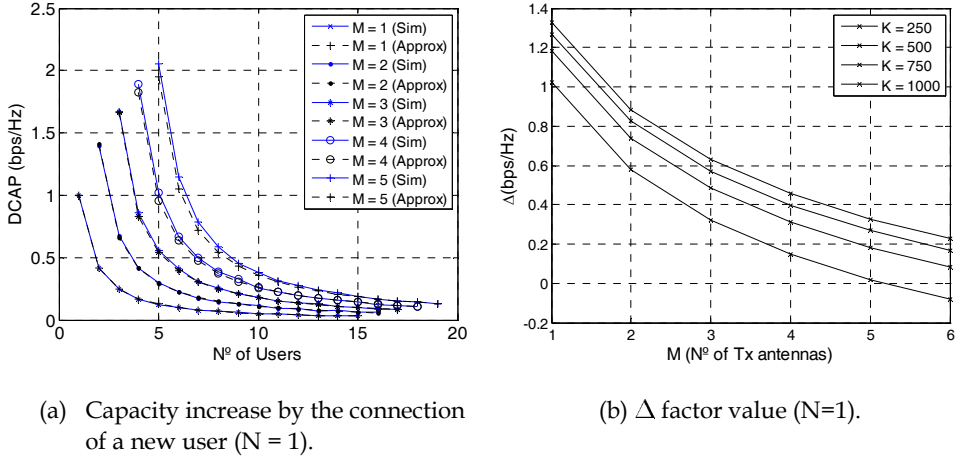


Fig. 6. Capacity increase by the connection of one more transmit antenna to the system, for $KN \geq M$.

When more aggregate receive than transmit antennas are considered it is important to analyze the Δ factor in equation (23), to see if it has some significance in relation to $\log_2(P)$. Thus, we must evaluate the capacity increase obtained by the connection of an additional user to the system. To do that, we have relied on numerical simulations, considering that $N = 1$ and that each column of \mathbf{H} is $\mathcal{N}_{KN}(0, \mathbf{I})$ distributed. Since for this case no closed form expression exists for the power allocation we have relied on the algorithm presented in (Jindal et al. 2005) to optimally allocate power among users. As a result Fig. 6 (a) the ergodic capacity increase, when a new user is connected to the system, was obtained... In Fig. 6 (a) it is also shown the equation (27) approximation, which is very close to the obtained numerical results. Taking into account that approximation in our analysis, in Fig. 6 (b), we plot the Δ factor values for different number of users. From those values one can see that even for 1000 users this factor is small, not higher than 1.4 bps/Hz. Thus it has no significance in relation to $\log(P)$, since we are considering the high SNR regime. From that figure we can also confirm that Δ increases fewer as K increases. Consequently, to obtain a Δ value with the same order of magnitude of $\log(P)$ a very high number of users should be jointly processed, which will result in a very high processing complexity at the CU, which must be avoided.

7. Conclusion

In this chapter we have analyzed the benefits of the connection of additional transmit antennas to the system users of a BC. The single-user case has been analyzed in more detail, for all SNR's and the multi-user case has been analyzed only in the high SNR regime. For

the single user case we have obtained a closed-form expression for DCAP that is valid for all SNR's and have analyzed its variation/sensitivity with respect to M . From that analysis we have seen that symmetry is the most important system property to obtain most of the gains provided by the connection of additional transmit antennas. It was also shown that the maximum increase in capacity is obtained when all links have the same mean SNR's and not when they are different. Thus showing that the DCAP maximum is obtained when all antennas are co-located and not distributed.

For the multi-user case a closed form expression for the co-located transmit antennas DCAP, for $M > KN$, has been obtained. For the general case of a DAS the DCAP was analyzed numerically, considering a scenario with 4 transmit antennas and 2 users, each with only one receive antenna. For this case even if the DCAP can be superior to the co-located case it is not much higher and most of the user's positions will have a lower DCAP than the co-located case. For the positions that attain a higher DCAP than the co-located case we have shown that this positions are very close to the system symmetry lines, defined by the transmit antennas, and in that way symmetry plays again a important role in the multi-user case. For the case of more aggregate receive than transmit antennas we have verified that the DCAP will be power dependent and will be given by the multiplexing gain plus a Δ factor. However this Δ factor is much smaller than the multiplexing gain even for one thousand users. Thus for this case the most significant gain is the multiplexing gain, $\log(P)$.

8. References

- Caire, G. & Shamai, S., 2003. On the achievable throughput of a multiantenna Gaussian broadcast channel. *Information Theory, IEEE Transactions on*, 49(7), 1691-1706.
- Castanheira, D. & Gameiro, A., 2008. Distributed MISO system capacity over Rayleigh flat fading channels'. In *Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium on*. pp. 1-5.
- Castanheira, D. & Gameiro, A., 2009. High SNR Broadcast Channel Differential Capacity. In *ICT-MobileSummit 2009*. Santander, Spain, pp. 1-5.
- Cover, T., 1972. Broadcast channels. *Information Theory, IEEE Transactions on*, 18(1), 2-14.
- Dohler, M., Gkelias, A. & Aghvami, A., 2006. Capacity of distributed PHY-layer sensor networks. *Vehicular Technology, IEEE Transactions on*, 55(2), 622-639.
- FUTON, 2008. FUTON - Fibre-Optic Networks for Distributed Extendible Heterogeneous Radio Architectures and Service Provisioning. Available at: <http://www.ict-futon.eu/>.
- Ghosh, S., 2005. *Network Theory*, PHI Learning Pvt. Ltd.
- Goldsmith, A., 2005. *Wireless communications*, Cambridge University Press.
- Gradshteyn, I.S. & Ryzhik, I.M., 1994. *Table of Integrals, Series, and Products* 5th ed., Academic Press.
- Gubner, J.A., 2006. *Probability and Random Processes for Electrical and Computer Engineers* 1st ed., Cambridge University Press.
- I. R. R. M. M1645, 2003. M.1645: Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000. Available at: <http://www.itu.int/rec/R-REC-M.1645/e>.
- Jindal, N., 2005. High SNR analysis of MIMO broadcast channels. In *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*. pp. 2310-2314.

- Jindal, N. et al., 2005. Sum power iterative water-filling for multi-antenna Gaussian broadcast channels. *Information Theory, IEEE Transactions on*, 51(4), 1570-1580.
- Juyul Lee & Jindal, N., 2007. High SNR Analysis for MIMO Broadcast Channels: Dirty Paper Coding Versus Linear Precoding. *Information Theory, IEEE Transactions on*, 53(12), 4787-4792.
- Liu, H. & Li, G., 2005a. *OFDM-Based Broadband Wireless Networks: Design and Optimization*, Wiley-Interscience.
- Liu, H. & Li, G., 2005b. *OFDM-Based Broadband Wireless Networks: Design and Optimization*, Wiley-Interscience.
- Seijoon Shim et al., 2008. Block diagonalization for multi-user MIMO with other-cell interference. *Wireless Communications, IEEE Transactions on*, 7(7), 2671-2681.
- Shamai, S. & Verdú, S., 2001. The impact of frequency-flat fading on the spectral efficiency of CDMA. *Information Theory, IEEE Transactions on*, 47(4), 1302-1327.
- Shin, H. & Lee, J.H., 2003. Capacity of multiple-antenna fading channels: spatial fading correlation, double scattering, and keyhole. *Information Theory, IEEE Transactions on*, 49(10), 2636-2647.
- Telatar, I.E., 1999. Capacity of multi-antenna Gaussian channels. Available at: <http://eprints.kfupm.edu.sa/29242/>.
- Vishwanath, S., Jindal, N. & Goldsmith, A., 2003. Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels. *Information Theory, IEEE Transactions on*, 49(10), 2658-2668.
- Visotsky, E. & Madhow, U., 2001. Space-time transmit precoding with imperfect feedback. *Information Theory, IEEE Transactions on*, 47(6), 2632-2639.
- Viswanath, P. & Tse, D., 2003. Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality. *Information Theory, IEEE Transactions on*, 49(8), 1912-1921.
- Wei Yu & Cioffi, J., 2004. Sum capacity of Gaussian vector broadcast channels. *Information Theory, IEEE Transactions on*, 50(9), 1875-1892.
- Weingarten, H., Steinberg, Y. & Shamai, S., 2006. The Capacity Region of the Gaussian Multiple-Input Multiple-Output Broadcast Channel. *Information Theory, IEEE Transactions on*, 52(9), 3936-3964.

9. Appendix

9.1 DCAP derivation

For the co-located transmit antennas case $\bar{\mathbf{H}}_i \bar{\mathbf{H}}_i^H$, ($i = M, M+1$), is $\tilde{\mathcal{W}}_{KN}(i, \Xi)$ distributed. But since $\bar{\mathbf{H}}_i \bar{\mathbf{H}}_i^H \sim \Xi^{1/2} \mathbf{A}_i \Xi^{1/2}$, where \mathbf{A}_i is $\tilde{\mathcal{W}}_{KN}(i, \mathbf{I}_{KN})$ distributed, the DCAP can be expressed by:

$$\begin{aligned}
 \Delta C_{M,K,N}^{M+1,K,N} &= \mathbb{E}[\log|\bar{\mathbf{H}}_{M+1} \bar{\mathbf{H}}_{M+1}^H|] - \mathbb{E}[\log|\bar{\mathbf{H}}_M \bar{\mathbf{H}}_M^H|] \\
 &= \mathbb{E}[\log|\mathbf{A}_{M+1}|] - \mathbb{E}[\log|\mathbf{A}_M|] \\
 &= \sum_{l=0}^{KN-1} \Psi(M+1-l) + \sum_{l=0}^{KN-1} \Psi(M-l) \\
 &= \sum_{n=M-KN+1}^M \frac{1}{n} \approx \log \left(1 + \frac{KN}{M-KN+\beta} \right)
 \end{aligned} \tag{28}$$

where $\Psi(l) = -\gamma + \sum_{n=1}^{l-1} \frac{1}{n}$, is the Euler's digamma function. The result from the second line comes from the determinant property ($|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$) and the result from the third line from lemma 1 of (Juyul Lee & Jindal 2007) ($\mathbb{E}[\log|\mathbf{A}_i|] = \sum_{l=0}^{i-1} \Psi(i-l)$). To obtain an approximation to the partial sum of the harmonic series we have used $\sum_{n=1}^M \frac{1}{n} = \gamma + \log(M + \beta)$ which has a maximum error of 4.1×10^{-3} .

9.1 Capacity Increase by the connection of one more user, for $N = 1$ and $M = 1$

For $N = 1$ and $M = 1$ the broadcast channel ergodic sum-capacity, equation (7), in the high SNR regime, simplifies to:

$$\overline{C}_{DPC}^{1,K,1} = \mathbb{E} \left[\log \left(P \max_{k=1 \dots K} (\xi_k^{1/2} h_k^H h_k \xi_k^{1/2}) \right) \right] = \mathbb{E}[\log(Px)] \quad (29)$$

where ξ_k is the average SNR of user k , $h_k \sim \tilde{\mathcal{N}}(0, 1)$ and $x = \max_k (\xi_k h_k^H h_k)$. It is easy to show that $d\Delta C_{1,K,1}^{1,K+1,1}/d\xi_k \leq 0$ for $k \neq K+1$ and that the opposite happens for $k = K+1$. Thus the maximum of $\Delta C_{1,K,1}^{1,K+1,1}$ is obtained when $\xi_k = \xi$ for all users, if the new connected user is always the one which implies the highest increase in the ergodic sum-capacity. It can also be proven that the same value for $\Delta C_{1,K,1}^{1,K+1,1}$ is obtained for any ξ . Hence, for simplicity we consider that $\xi = 1$. Thus x has cumulative distribution function $F_X(x, K) = F_Y(y)^K$, (Liu & Li 2005b), where $y \sim \xi_k h_k^H h_k$ is exponential distributed with mean 1. Thus the capacity increase by the connection of an additional user to the system can be expressed by⁷:

$$\Delta C_{1,K,1}^{1,K+1,1} = \int_0^\infty \log(x) dF_X(x, K+1) - \int_0^\infty \log(x) dF_X(x, K) \quad (30)$$

which after the integral evaluation gives the result shown in equation (26).

⁷ The P term cancel out from the expression since it appears in $\overline{C}_{DPC}^{1,K+1,1}$ and $\overline{C}_{DPC}^{1,K,1}$.

Low Dimensional MIMO Systems with Finite Sized Constellation Inputs

Rizwan Ghaffar and Raymond Knopp

Eurecom
FRANCE

1. Introduction

The seminal works in (Foschini & Gans, 1998) and (Telatar, 1999) on multiple antenna elements at the transmitter and the receiver show a huge increase in the throughput of this point-to-point channel referred to also as multiple input multiple output (MIMO) system. These promising results of high spectral efficiency and enhanced reliability shifted the focus of research on multi antenna communications and motivated the introduction of multiple antenna elements in the future communication systems. Researchers persist to strive for finding space time codes (STC) with reduced decoding complexity. These codes take into account both the spatial and temporal dimensions of the MIMO channel. Orthogonal Space-Time Block Codes (OSTBCs) (Larsson & Stoica, 2003) are widely used because they are easy to encode and decode. For the case of two transmit antennas, the OSTBC is known as Alamouti code (Alamouti, 1998). OSTBCs are repetition codes that only provide diversity gain. In order to approach the capacity limit they have to be used in concatenation with an outer code. Remarkable coding gains can be obtained if a capacity achieving temporal encoder, such as turbo or Low-Density Parity Check (LDPC) code is used in concatenation with a STC (Gonzalez-Lopez et al., 2006). Recently it has been shown for the ergodic channels that the complex concatenation of the STC and the outer codes can be replaced with temporally coded and spatially multiplexed streams (coded spatial streams) for nearing capacity (Ghaffar & Knopp, 2008a). Each spatial stream can also be independently coded using temporal encoders as convolutional, turbo or LDPC codes whereas at the receiver, standard off-the-shelf decoders are used after the demodulator. To combat the frequency selectivity of MIMO wireless channels with low complexity equalization at the receivers, MIMO OFDM is the appropriate alternative. To contest the inherent fading of MIMO OFDM wireless channels, improved code diversity of bit interleaved coded modulation (BICM) for fading channels is rendering it the preferred option. Consequently the future wireless systems shall be based on BICM MIMO OFDM systems. However the requisite antenna spacing combined with the complexity constraints at the receiver are restricting the future MIMO based communication systems to the maximum of 4 spatial streams whereas it is reduced to 2 spatial streams in most scenarios. The existing and forthcoming standards as IEEE 802.11n (802.11n, 2006), IEEE 802.16m (802.16m, 2007) and Third Generation Partnership Project Long Term Evolution (3GPP LTE) (LTE, 2006) substantiate this argument. This chapter therefore focuses on low dimensional spatially multiplexed time coded BICM MIMO OFDM systems with first part being devoted to the transmission strategies and corresponding receiver structures for such systems in the broadcast scenario while second part

deliberates on interference suppression in such systems in the cellular scenario. This chapter particularly takes into account the finite sized constellation inputs and departs from the customary idealistic Gaussian assumption for the codewords. Each part is also accompanied by relevant information theoretic analysis and by simulation results under the settings of upcoming wireless standards.

2. MIMO Broadcast scenario

This part deliberates on the broadcast scenario of BICM MIMO OFDM system though the discussion also remains valid for the point-to-point MIMO systems. We consider the transmission strategy in which each spatial stream is independently encoded and modulated. We focus on the case of uniform power and nonuniform rate spatial streams (Ghaffar & Knopp, 2008a) and the case of uniform rate and nonuniform power distribution (Ghaffar & Knopp, 2008b) between these spatial streams. In such a broadcast scenario, receiver consequently views a multiple access channel (MAC). Shamaï (Shamaï & Steiner, 2003) termed the approach of single code layer at each transmit antenna as *MAC-outage approach*. The reception is consequently based on successive interference cancellation (SIC) i.e. sequential decoding and subtraction (stripping) of spatial streams which introduces unequal error protection (UEP). This can be coarsely regarded as MMSE DFE as described in (Varanasi & Guess, 1997). The idea of multiple data streams with UEP adds flexibility to the system which can be exploited for having prioritized users or advanced services in MIMO broadcast systems and in multimedia broadcast multicast services (MBMS). For instance it can be the broadcast of multimedia streams with different rates (quality) of the same data and the users decoding the stream depending on the received SNR. It can also be the broadcast of low and high rate streams (as audio and video) with prioritized or high SNR users decoding both audio and video streams while low SNR users decoding only the low rate audio stream. It is also applicable to high-definition TV (HDTV) scenario where low priority/quality users are able to receive standard-definition TV (SDTV) transmission while high priority/quality users access HDTV. This idea has limited similarity to superposition codes (Liu et al., 2002) whose signal space has a cloud/satellite topology. Cloud centers because of relatively higher distance amongst them carry information for low quality receivers whereas better receivers having larger noise tolerance can resolve up to the actual transmitted satellite symbol within the cloud.

For coded spatial streams (also for the STC), the well-known data model after appropriate filtering and sampling is $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}$ (to be made precise in the subsequent sections) where \mathbf{y} is the received data, \mathbf{H} is the channel matrix, \mathbf{x} is the symbol vector with the elements from finite constellations and \mathbf{z} is the noise. The problem is then to detect some or all elements of \mathbf{x} from \mathbf{y} . Essentially the same problem occurs in multiuser detection for CDMA (Verdu, 1998) and for single-carrier transmission over channels that induce intersymbol interference. In these cases, the matrix \mathbf{H} usually has a specific structure.

The problem of detection of \mathbf{x} from \mathbf{y} has stimulated a large body of research (Verdu, 1998) and references therein. One can easily show that if the noise \mathbf{z} is Gaussian then obtaining the maximum-likelihood (ML) solution for some or all elements of \mathbf{x} is equivalent to minimizing the Euclidean distance $\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2$ with respect to \mathbf{x} over the finite set spanned by all possible combinations of constellation points that can constitute the vector \mathbf{x} . For ML soft MIMO detection, the demodulator calculates the log-likelihood ratios (LLRs) for all bits that constitute the desired elements of \mathbf{x} by summing the Euclidean distances for the values of \mathbf{x} for which that particular bit of the desired element of \mathbf{x} is one and zero thereby amounting to $2\log(M_1) + \dots + \log(M_{n_r})$ terms where M_k is the modulation alphabet of the k -th spatial stream and

n_r is the total number of spatial streams (Larsson & Jaldén, 2008). In many cases of practical interest, one resorts to the approximation of replacing the sums with the largest term which is equivalent to minimizing the Euclidean distance and is termed as max log MAP approach. Unfortunately this problem is NP-hard for general \mathbf{H} and \mathbf{y} (Verdu, 1989) which implies that there are no known efficient (i.e. polynomial-time) solutions. Many sophisticated methods as lattice reduction and sphere decoding (Hochwald & Brink, 2003) exist which find the ML solution with high probability, but these methods are in general still computationally complex. This is true also in an average sense if \mathbf{H} is random (i.e. for a fading channel). The popular "sphere decoding" method is much more efficient than a brute-force search, but it still admits an average complexity that is exponential in the dimension of \mathbf{x} .

Naive solutions, like neglecting the integer constraint coupled with the Gaussian assumption for the alphabets and then subsequently projecting the so-obtained solution onto the finite set of permissible \mathbf{x} [linear receivers as LMMSE and zero forcing ZF], in general work poorly especially at lower SNRs. Standard linear detection approaches are further based on ignoring the spatial color at the output of linear detectors which results in the decoupling of spatial streams thereby fundamentally reducing the complexity of detection. These disregards proliferate the suboptimality of linear receivers which exhibit degraded performance especially at lower SNRs.

Standard receiver solutions for spatially multiplexed broadcast schemes including V-BLAST (Wolniansky et al., 1998) (Golden et al., 1999) use stripping decoders which incorporate sub-optimal linear minimum mean square error (MMSE) filters (Medvedev et al., 2006) against the yet undecoded streams at each successive cancellation stage. MMSE because of its relative improved performance in the family of linear detectors is the preferred choice. Its optimality for power constrained Gaussian alphabets is well known but it is suboptimal for finite size constellations. Gaussian assumption of the post detection interference is open to discussion. Its behavior is close to Gaussian under various asymptotic conditions which include large SNRs and large number of transmit and receive antennas (Poor & Verdu, 1997). But the fidelity of Gaussian assumption in a low dimensional system at moderate SNRs is questionable. Degradation of the performance due to the suboptimality combined with the complexity in the calculation of linear equalizers at each frequency tone (in OFDM based system) renders their real-time implementation debatable especially in fast fading wideband environments.

2.1 System Model

Before deliberating further on these receiver structures, we discuss the system model. As the overall system is based on BICM MIMO OFDM, it is imperative to first understand the significance and the implication of using BICM.

2.1.1 BICM SISO System

BICM because of its improved code diversity for fading channels and its flexibility to variable transmission rates, is a likely choice for future wireless systems as IEEE 802.11n (802.11n, 2006), IEEE 802.16m (802.16m, 2007) and 3GPP LTE (LTE, 2006). The landmark paper of Caire (Caire et al., 1998) on BICM showed that on some channels, the separation of demodulation and decoding is beneficial, provided that the encoder output is interleaved bit wise and a suitable soft decision metric is used in the Viterbi decoder. Code diversity, and therefore the reliability of coded modulation over a Rayleigh channel, can be improved this way. The code diversity in this case is equal to the smallest number of distinct bits along any error event. This leads to a better coding gain over a fading channel when compared to other coded mod-

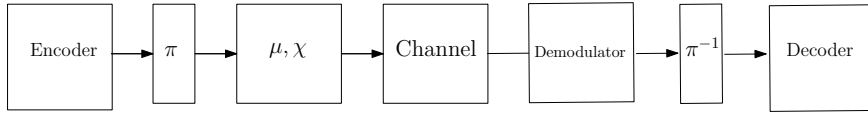


Fig. 1. Block Diagram of BICM system. π denotes denotes a bit interleaver.

ulation schemes as Trellis Coded Modulation (TCM). BICM increases considerably Hamming distance while reducing (often marginally) Euclidean distance so BICM outperforms TCM over Rayleigh fading channel while suffering a moderate loss of performance over AWGN channel. If the channel model is nonstationary, in the sense that the propagation environment changes during transmission, then BICM provides a robust coding scheme.

The main idea of BICM is therefore to transform the channel generated by the multilevel constellation χ into parallel and independent binary channels. For transmission of complex modulation, channel is not binary but after bit interleaving, any transmission of a multilevel signal from χ with $|\chi| = 2^m$, can actually be thought of as taking place over m parallel channels, each carrying one binary symbol from the signal label. However, these channels are generally not independent, due to the constellation structure. To make them independent, binary symbols are interleaved over infinite length before being used as signal labels. The maximum-likelihood decoding (MLD) of BICM requires combined demodulation/decoding, which is often too complicated to implement. As a result, MLD is separated at the receiver, concatenating soft-metrics computation, deinterleaving and decoding. BICM block diagram is shown in fig. 1 which is the concatenation of an encoder for a code C with an interleaver π followed by a modulator (μ, χ) . In the decoder, the metrics reflect the fact of bits separation. Suppose that the code word to be transmitted is \underline{c} . After interleaving and modulation, we transmit the codeword

$$\mathbf{x} = (x_1, x_2, \dots, x_n)$$

and we receive \mathbf{y} at the output of a stationary memoryless channel. With symbol interleaving, we decode by maximizing the metric

$$\log p(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^n \log p(y_k|x_k) \quad (1)$$

with respect to \mathbf{x} .

The bit interleaver can be seen as a one-to-one correspondence $\pi : k' \rightarrow (k, i)$, where k' denotes the original ordering of the coded bits $c_{k'}$, k denotes the time ordering of the signals x_k transmitted, and i indicates the position of the bit $c_{k'}$ in the symbol x_k . Let χ_b^i denote the subset of all signals $x \in \chi$ whose label has the value $b \in \{0, 1\}$ in position i . Then the ML bit metric is given as

$$\lambda^i(y_k, c_{k'}) = \log \sum_{x \in \chi_b^i} p(y_k|x) \quad \text{where } c_{k'} \in [0, 1] \quad \text{and } i = 1, 2, \dots, \log |\chi| \quad (2)$$

So in case of BICM, it is the summation of bit metrics $\lambda^i(y_k, c_{k'})$ instead of the symbol metrics $\log p(y_k|x_k)$ for decoding. i.e.

$$\hat{\underline{c}} = \arg \max_{\underline{c} \in C} \sum_{k'} \lambda^i(y_k, c_{k'}) \quad (3)$$

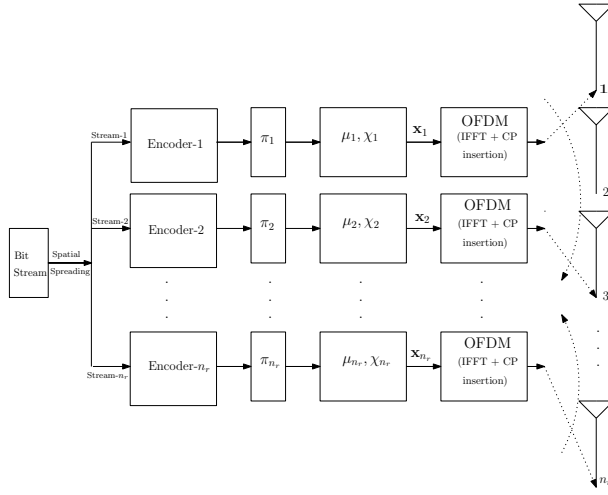


Fig. 2. Block diagram of Transmitter of $n_t \times n_r$ BICM MIMO OFDM system. π_1 denotes random interleaver, μ_1 labeling map, χ_1 signal set and \mathbf{x}_1 complex symbols vector for stream-1.

The bit metrics (2) may be computationally too complex for implementation. Suboptimal simplified branch metric can be obtained by the log-sum approximation $\log \sum_j z_j \approx \max_j \log z_j$. This yields

$$\lambda^i(y_k, c_{k'}) = \max_{x \in \chi_{c_{k'}}^i} \log p(y_k | x) = \min_{x \in \chi_{c_{k'}}^i} |y_k - h_k x|^2 \quad (4)$$

where h_k denotes the Rayleigh coefficient.

2.1.2 BICM MIMO OFDM System

We consider a MIMO broadcast system (without CSIT) which is a $n_t \times n_r$ ($n_t \geq n_r$) BICM MIMO OFDM system with n_r spatial streams as shown in figs. 2 and 3. We effectively reduce this to $n_r \times n_r$ system by antenna cycling at the transmitter (Foschini & Gans, 1998) with each stream being transmitted by one antenna in any dimension. The antenna used by a particular stream is randomly assigned per dimension so that each stream sees all degrees of freedom of the channel. Let the spatial streams be $\mathbf{x}_1, \dots, \mathbf{x}_{n_r}$. x_l is the symbol of \mathbf{x}_l over a signal set $\chi_l \subseteq \mathcal{C}$ with a Gray labeling map $\mu_l : \{0, 1\}^{\log_2 |\chi_l|} \rightarrow \chi_l$. During the transmission of l -th spatial stream, the code sequence \mathbf{c}_l is interleaved by π_l and then is mapped onto the signal sequence $\mathbf{x}_l \in \chi_l$. Bit interleaver for the l -th stream can be modeled as $\pi_l : k' \rightarrow (k, i)$ where k' denotes the original ordering of the coded bits $c_{k'}$ of the l -th stream, k denotes the time ordering of the signal $x_{l,k}$ and i indicates the position of the bit $c_{k'}$ in the symbol $x_{l,k}$.

We assume that the frequency reuse factor is one and cyclic prefix (CP) of appropriate length is added to the OFDM symbols. Cascading IFFT at the transmitter and FFT at the receiver

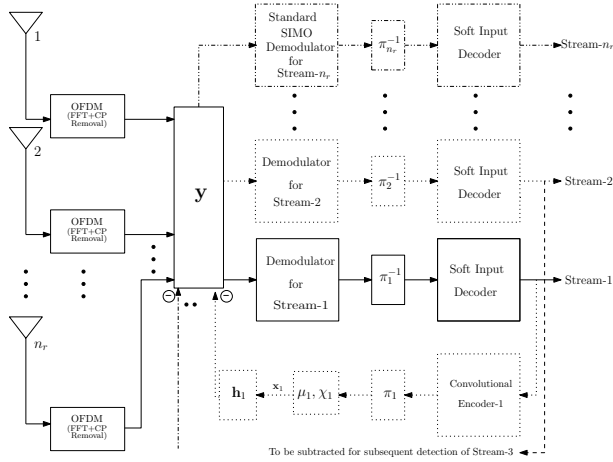


Fig. 3. Block diagram of SIC Receiver of BICM MIMO OFDM system. π_1^{-1} denotes deinter-leaver and \mathbf{h}_1 denotes the channel seen by stream-1.

with CP extension, transmission at the k -th frequency tone can be expressed as:-

$$\begin{aligned} \mathbf{y}_k &= \mathbf{h}_{1,k}x_1 + \mathbf{h}_{2,k}x_2 + \cdots + \mathbf{h}_{n_r,k}x_{n_r} + \mathbf{z}_k, \quad k = 1, 2, \dots, T \\ &= \mathbf{H}_k \mathbf{x}_k + \mathbf{z}_k \end{aligned} \quad (5)$$

where $\mathbf{H}_k = [\mathbf{h}_{1,k} \cdots \mathbf{h}_{n_r,k}]$ i.e. the channel at the k -th frequency tone, $\mathbf{x}_k = [x_{1,k}, \dots, x_{n_r,k}]^T$ and $(\cdot)^T$ indicates the transpose operation. Each subcarrier corresponds to a symbol \mathbf{x} from a constellation map $\chi_1, \dots, \chi_{n_r}$. $\mathbf{y}_k, \mathbf{z}_k \in \mathbb{C}^{n_r}$ are the vectors of received symbols and circularly symmetric complex white Gaussian noise of double-sided power spectral density $N_0/2$ at the n_r receive antennas. $\mathbf{h}_{l,k} \in \mathbb{C}^{n_r}$ is the vector characterizing flat fading channel response from l -th transmitting antenna to n_r receive antennas at k -th subcarrier. This vector has complex-valued multivariate Gaussian distribution with $E[\mathbf{h}_{l,k}] = \mathbf{0}$ and $E[\mathbf{h}_{l,k}\mathbf{h}_{l,k}^\dagger] = \mathbf{I}$. The antennas at the transmitter are also assumed to be sufficiently spaced and therefore are uncorrelated. The complex symbols $x_{1,k}, \dots, x_{n_r,k}$ of the spatial streams are assumed to be independent with variances $\sigma_1^2, \dots, \sigma_{n_r}^2$ respectively. The channels at different subcarriers are also assumed to be independent. Bit metric for the bit $c_{k'}$ at the i -th location of the symbol $x_{l,k}$ is given as

$$\lambda_l^i(\mathbf{y}_k, c_{k'}) = \log \sum_{x_1 \in \chi_1} \cdots \sum_{x_l \in \chi_{l,c_{k'}}} \cdots \sum_{x_{n_r} \in \chi_{n_r}} \exp \left[-\frac{1}{N_0} \|\mathbf{y}_k - \mathbf{H}_k \mathbf{x}\|^2 \right]$$

Applying log-sum approximation we have:-

$$\lambda_l^i(\mathbf{y}_k, c_{k'}) \approx \min_{x_1 \in \chi_1 \cdots x_l \in \chi_{l,c_{k'}} \cdots x_{n_r} \in \chi_{n_r}} \left[\|\mathbf{y}_k - \mathbf{H}_k \mathbf{x}\|^2 \right] \quad (6)$$

2.2 Information Theoretic View

We now calculate the mutual information of this system for the cases of Gaussian and finite sized constellation inputs.

2.2.1 Gaussian Inputs

The system equation ignoring the frequency index takes the form:-

$$\mathbf{y} = \mathbf{h}_1 x_1 + \mathbf{h}_2 x_2 + \cdots + \mathbf{h}_{n_r} x_{n_r} + \mathbf{z} \quad (7)$$

Since the receiver knows the realization of \mathbf{H} , the channel output is the pair $(\mathbf{y}; \mathbf{H}) = (\mathbf{H}\mathbf{x} + \mathbf{z}; \mathbf{H})$. The mutual information between input and output is then Telatar (1999)

$$\begin{aligned} I(\mathbf{x}; (\mathbf{y}, \mathbf{H})) &= I(\mathbf{x}; \mathbf{H}) + I(\mathbf{x}; \mathbf{y} | \mathbf{H}) \\ &= I(\mathbf{x}; \mathbf{y} | \mathbf{H}) \\ &= E_{\mathbf{H}} I(\mathbf{x}; \mathbf{y} | \mathbf{H} = \mathbf{H}) \end{aligned}$$

For the Gaussian inputs, we consider the following two cases:-

1. Spatial streams of uniform power and non-uniform rate.
2. Spatial streams of uniform rate and non-uniform power.

For Gaussian inputs, channel capacity of the system as per the chain rule (Foschini & Gans, 1998) is

$$I(x_1, x_2 \cdots x_{n_r}; \mathbf{y}) = I(x_1; \mathbf{y}) + I(x_2; \mathbf{y} | x_1) + \cdots + I(x_{n_r}; \mathbf{y} | x_1, x_2 \cdots x_{n_r-1})$$

The terms in the summation represent the channel capacities of each spatial stream once they are detected in the successive subtractive cancellation way. Conditioned on the channel, these terms can be written as:-

$$\begin{aligned} I(x_1; \mathbf{y} | \mathbf{H}) &= \log_2 \left[\det \left\{ \mathbf{I} + \sigma_1^2 \mathbf{h}_1 \mathbf{h}_1^\dagger \left(N_0 \mathbf{I} + \sigma_2^2 \mathbf{h}_2 \mathbf{h}_2^\dagger + \cdots + \sigma_{n_r}^2 \mathbf{h}_{n_r} \mathbf{h}_{n_r}^\dagger \right)^{-1} \right\} \right] \\ I(x_2; \mathbf{y} | \mathbf{H}, x_1) &= \log_2 \left[\det \left\{ \mathbf{I} + \sigma_2^2 \mathbf{h}_2 \mathbf{h}_2^\dagger \left(N_0 \mathbf{I} + \sigma_3^2 \mathbf{h}_3 \mathbf{h}_3^\dagger + \cdots + \sigma_{n_r}^2 \mathbf{h}_{n_r} \mathbf{h}_{n_r}^\dagger \right)^{-1} \right\} \right] \end{aligned}$$

and

$$I(x_{n_r}; \mathbf{y} | \mathbf{H}, x_1, x_2 \cdots, x_{n_r-1}) = \log_2 \left(1 + \frac{\sigma_{n_r}^2}{N_0} \|\mathbf{h}_{n_r}\|^2 \right)$$

where $\mathbf{H} = [\mathbf{h}_1 \mathbf{h}_2 \cdots \mathbf{h}_{n_r}]$ is the channel matrix. Fig. 4 shows the ergodic capacity for the case of 2×2 system with spatial streams of uniform power and nonuniform rate. Note that SNR is the received SNR per antenna i.e. $\text{SNR} = \frac{\sigma_1^2 + \sigma_2^2}{N_0}$. It is evident that the stream to be detected first has lower capacity as compared to the stream to be detected last which enjoys higher diversity.

Fig. 5 compares two cases of spatial streams with uniform power and nonuniform rate and spatial streams with uniform rate and nonuniform power for 2×2 , 3×3 and 4×4 systems. Key to the optimality of stripping is the use of Gaussian inputs as long as the stripping decoders incorporate MMSE filters against yet undecoded streams at each successive cancellation stage. Successive stripping requires that each stream must be transmitted at a different

rate with uniform power. We investigate a slightly suboptimal solution where we guarantee equal rate with nonuniform powers on each stream. Numerical optimization revealed that uniform rate and nonuniform power distribution leads to negligible suboptimality as shown in fig. 5.

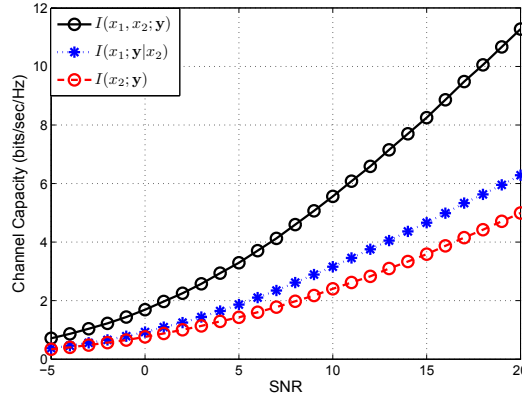


Fig. 4. Capacity of 2×2 system for Gaussian alphabets for the case of uniform power and nonuniform rate spatial streams.

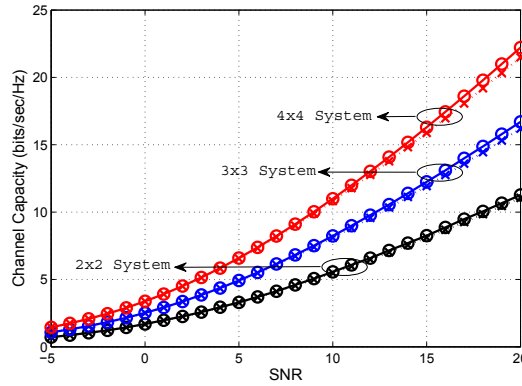


Fig. 5. Capacity of 2×2 , 3×3 and 4×4 systems for Gaussian alphabets for the cases of spatial streams of uniform power and nonuniform rate and the spatial streams of uniform rate and nonuniform power. Note that the circles indicate the case of uniform power and nonuniform rate spatial streams while crosses indicate the case of uniform rate and nonuniform power spatial streams.

2.2.2 Finite Sized Constellation Inputs

To reduce the complexity and enhance the understanding of mutual information for finite sized constellation inputs, we restrict to the case of dual stream transmission. The system equation ignoring the frequency index takes the form:-

$$\mathbf{y} = \mathbf{h}_1 x_1 + \mathbf{h}_2 x_2 + \mathbf{z} \quad (8)$$

Mutual information expression for the dual streams from the chain rule (Foschini & Gans, 1998) is given as

$$I(\mathbf{y}; x_1, x_2) = I(\mathbf{y}; x_1) + I(\mathbf{y}; x_2 | x_1) \quad (9)$$

For equal power distribution, $I(\mathbf{y}; x_1) < I(\mathbf{y}; x_2 | x_1)$ dictating rate of first stream being less than rate of second stream ($R_1 < R_2$). For finite size QAM constellation with $x_1 \in M_1$ and $x_2 \in M_2$, the mutual information expression conditioned on the channel takes the form (Ghaffar & Knopp, 2008a)

$$\begin{aligned} I(\mathbf{y}; x_1 | \mathbf{H}) &= \mathcal{H}(x_1 | \mathbf{H}) - \mathcal{H}(x_1 | \mathbf{y}, \mathbf{H}) \\ &= \log M_1 - \mathcal{H}(x_1 | \mathbf{y}, \mathbf{H}) \end{aligned} \quad (10)$$

where $\mathcal{H}(\cdot) = -E \log p(\cdot)$ is the entropy function. Second term of eq. (10) is given as:-

$$\begin{aligned} \mathcal{H}(x_1 | \mathbf{y}, \mathbf{H}) &= \sum_{x_1} \int_{\mathbf{y}} \int_{\mathbf{H}} p(x_1, \mathbf{y}, \mathbf{H}) \log \frac{1}{p(x_1 | \mathbf{y}, \mathbf{H})} d\mathbf{y} d\mathbf{H} \\ &= \sum_{x_1} \int_{\mathbf{y}} \int_{\mathbf{H}} p(x_1, \mathbf{y}, \mathbf{H}) \log \frac{p(\mathbf{y}, \mathbf{H})}{p(x_1, \mathbf{y}, \mathbf{H})} d\mathbf{y} d\mathbf{H} \\ &= \sum_{x_1} \sum_{x_2} \int_{\mathbf{y}} \int_{\mathbf{H}} p(x_1, x_2, \mathbf{y}, \mathbf{H}) \log \frac{\sum_{x'_1} \sum_{x'_2} p(\mathbf{y} | x'_1, x'_2, \mathbf{H})}{\sum_{x'_2} p(\mathbf{y} | x_1, x'_2, \mathbf{H})} d\mathbf{y} d\mathbf{H} \end{aligned} \quad (11)$$

For our purposes, it suffices to note that for each choice of x_1 and x_2 , there are two sources of randomness in the choices of channel and noise. The above quantities can be easily approximated numerically using sampling (Monte-Carlo) methods with N_z realizations of noise and N_H realizations of the channel i.e.

$$\begin{aligned} \mathcal{H}(x_1 | \mathbf{y}, \mathbf{H}) &= \frac{1}{M_1 M_2 N_z N_H} \sum_{x_1} \sum_{x_2} \sum_{\mathbf{H}} \sum_{\mathbf{z}} \log \frac{\sum_{x'_1} \sum_{x'_2} \exp \left[-\frac{1}{N_0} \left\| \mathbf{y} - \mathbf{h}_1 x'_1 - \mathbf{h}_2 x'_2 \right\|^2 \right]}{\sum_{x'_2} \exp \left[-\frac{1}{N_0} \left\| \mathbf{y} - \mathbf{h}_1 x_1 - \mathbf{h}_2 x'_2 \right\|^2 \right]} \\ &= \frac{1}{M_1 M_2 N_z N_H} \sum_{x_1} \sum_{x_2} \sum_{\mathbf{H}} \sum_{\mathbf{z}} \log \frac{\sum_{x'_1} \sum_{x'_2} \exp \left[-\frac{1}{N_0} \left\| \mathbf{h}_1 x_1 + \mathbf{h}_2 x_2 + \mathbf{z} - \mathbf{h}_1 x'_1 - \mathbf{h}_2 x'_2 \right\|^2 \right]}{\sum_{x'_2} \exp \left[-\frac{1}{N_0} \left\| \mathbf{h}_2 x_2 + \mathbf{z} - \mathbf{h}_2 x'_2 \right\|^2 \right]} \end{aligned} \quad (12)$$

Similarly the mutual information of second stream conditioned on the channel when first stream has been detected is given by:-

$$\begin{aligned}
 I(\mathbf{y}; x_2 | x_1, \mathbf{H}) &= \mathcal{H}(x_2 | x_1, \mathbf{H}) - \mathcal{H}(x_2 | \mathbf{y}, x_1, \mathbf{H}) \\
 &= \log M_2 - \sum_{x_1} \sum_{x_2} \int_{\mathbf{y}} \int_{\mathbf{H}} p(x_1, x_2, \mathbf{y}, \mathbf{H}) \log \frac{1}{p(x_2 | \mathbf{y}, x_1, \mathbf{H})} d\mathbf{y} d\mathbf{H} \\
 &= \log M_2 - \sum_{x_1} \sum_{x_2} \int_{\mathbf{y}} \int_{\mathbf{H}} p(x_1, x_2, \mathbf{y}, \mathbf{H}) \log \frac{p(\mathbf{y}, x_1, \mathbf{H})}{p(x_1, x_2, \mathbf{y}, \mathbf{H})} d\mathbf{y} d\mathbf{H} \\
 &= \log M_2 - \sum_{x_1} \sum_{x_2} \int_{\mathbf{y}} \int_{\mathbf{H}} p(x_1, x_2, \mathbf{y}, \mathbf{H}) \log \frac{\sum_{x'_2} p(\mathbf{y} | x_1, x'_2, \mathbf{H})}{p(\mathbf{y} | x_1, x_2, \mathbf{H})} d\mathbf{y} d\mathbf{H} \quad (13)
 \end{aligned}$$

Estimation of this quantity using Monte-Carlo simulation

$$\begin{aligned}
 I(\mathbf{y}; x_2 | x_1, \mathbf{H}) &= \log M_2 - \frac{1}{M_1 M_2 N_z N_H} \sum_{x_1} \sum_{x_2} \sum_{\mathbf{H}} \sum_{\mathbf{z}} \log \frac{\sum_{x'_2} \exp \left[-\frac{1}{N_0} \|\mathbf{y} - \mathbf{h}_1 x_1 - \mathbf{h}_2 x'_2\|^2 \right]}{\exp \left[-\frac{1}{N_0} \|\mathbf{y} - \mathbf{h}_1 x_1 - \mathbf{h}_2 x_2\|^2 \right]} \\
 &= \log M_2 - \frac{1}{M_1 M_2 N_z N_H} \sum_{x_1} \sum_{x_2} \sum_{\mathbf{H}} \sum_{\mathbf{z}} \log \frac{\sum_{x'_2} \exp \left[-\frac{1}{N_0} \|\mathbf{h}_2 x_2 + \mathbf{z} - \mathbf{h}_2 x'_2\|^2 \right]}{\exp \left[-\frac{1}{N_0} \|\mathbf{z}\|^2 \right]} \quad (14)
 \end{aligned}$$

Fig. 6 shows the capacity of first stream once second stream is not yet decoded for different combinations of finite constellation alphabets. For moderate values of SNR, the capacity of first stream is a function of the yet undetected second stream and this capacity decreases as the rate (constellation size) of second stream increases. This degradation is not observed at low and high values of SNR as at low SNR, two streams are orthogonal while at high SNR, second stream can be perfectly stripped off leading to detection of first stream. Rate of first stream being a function of the rate of second stream leads to nonuniform rates in uniform power dual stream scenario and this leads to the following proposed broadcast strategy.

2.3 Broadcast Strategy

We restrict to dual stream scenario for the broadcast case. The broadcast approach in dual stream scenario based on UEP (*MAC-outage* (Shamai & Steiner, 2003)) is motivated by the capacity of a Gaussian broadcast channel with two users i.e.

$$C = I(x_1; y_1) + I(x_2; y_2 | x_1) \quad (15)$$

where user 2 sees a better channel and so is able to decode and strip off the interference.

The broadcast strategy (Ghaffar & Knopp, 2008a) incorporates the transmission of two spatial streams of uniform power and nonuniform rate and incorporates two levels of performance. The reliably decoded information rate depends on the state of the channel which is determined by monitoring the received SNR being above or below a certain threshold. Transmitter is operating at a constant power and data rate but the limited adaptability of the system helps receivers to gear up to a higher data rate as the channel conditions improve.

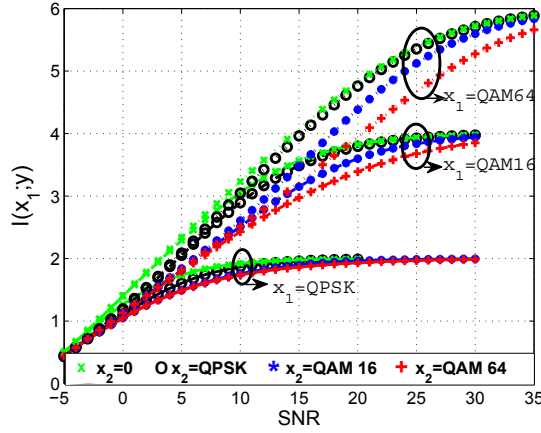


Fig. 6. Capacity of first stream in dual-stream broadcast approach for finite size alphabets once second stream is not known. Both streams have equal power. $x_2 = 0$ indicates the special case when second stream has been decoded and stripped off. Note that SNR includes power of both streams.

Low priority/quality users are able to decode low rate stream x_1 while high priority/quality users are able to decode both low and high rate streams i.e. x_1 and x_2 by successive stripping. The rates of two streams are

$$R_1 \leq I(\mathbf{y}; x_1) \quad (16)$$

and

$$R_2 \leq I(\mathbf{y}; x_2 | x_1) \quad (17)$$

The notion of priority/quality is typically the received SNR and/or stream decoupling. The users are divided into two groups i.e. near-in users and far-out users based on their received SNR. The lower rate stream x_1 is designed for a lower value of SNR i.e. SNR_1 while the higher rate stream x_2 is designed for higher value of SNR i.e. SNR_2 . The received SNR of a particular user dictates two decoding options.

1. If $\text{SNR}_2 > \text{SNR} \geq \text{SNR}_1$, the user decodes x_1 .
2. If $\text{SNR} \geq \text{SNR}_2$, the user decodes both streams i.e. x_1 and x_2 . The user first decodes low rate stream x_1 , strips it out and then decodes high rate stream x_2 .

This leads us to SIC detection based MIMO broadcast scenario with uniform power and nonuniform rate spatial streams. We now discuss the detectors for such broadcast scenario.

2.4 Detectors

The detectors discussed in this section are valid not only for spatially multiplexed MIMO systems but may be extended to other types of STC systems. We discuss two types of detectors as MMSE detector and low complexity max log MAP detector (Ghaffar & Knopp, 2009b).

2.4.1 MMSE

The frequency domain MMSE filter for $x_{1,k}$ is given as

$$\mathbf{h}_{1,k}^{MMSE} = \left(\mathbf{h}_{1,k}^\dagger \mathbf{R}_{1,k}^{-1} \mathbf{h}_{1,k} + \sigma_1^{-2} \right)^{-1} \mathbf{h}_{1,k}^\dagger \mathbf{R}_{1,k}^{-1} \quad (18)$$

where $\mathbf{R}_{1,k} = \sigma_2^2 \mathbf{h}_{2,k} \mathbf{h}_{2,k}^\dagger + \sigma_3^2 \mathbf{h}_{3,k} \mathbf{h}_{3,k}^\dagger + \dots + \sigma_{n_r}^2 \mathbf{h}_{n_r,k} \mathbf{h}_{n_r,k}^\dagger + N_0 \mathbf{I}$. After the application of MMSE filter we get

$$y_k = \alpha_k x_{1,k} + z_k \quad (19)$$

where z_k is assumed to be zero mean complex Gaussian random variable with variance $N_k = \mathbf{h}_{1,k}^{MMSE} \mathbf{R}_{1,k} \mathbf{h}_{1,k}^{MMSE^\dagger}$ and $\alpha_k = \mathbf{h}_{1,k}^{MMSE} \mathbf{h}_{1,k}$. Gaussianity has been assumed for post detection interference which increases the suboptimality of MMSE in the case of less number of interferers. Bit metric for the bit $c_{k'}$ on first stream is given as:-

$$\lambda_1^i(\mathbf{y}_k, c_{k'}) \approx \min_{x_1 \in \chi_{1,c_{k'}}^i} \left[\frac{1}{N_k} |y_k - \alpha_k x_1|^2 \right] \quad (20)$$

where $\chi_{1,c_{k'}}^i$ denotes the subset of the signal set $x_1 \in \chi_1$ whose labels have the value $c_{k'} \in \{0, 1\}$ in the position i . This metric has computational complexity $\mathcal{O}(|\chi_1|)$.

2.4.2 Low complexity max log MAP Detector

The max log MAP bit metric as per (6) is given as

$$\lambda_1^i(\mathbf{y}_k, c_{k'}) \approx \min_{x_1 \in \chi_{1,c_{k'}}^i, x_2 \in \chi_2, \dots, x_{n_r} \in \chi_{n_r}} \|\mathbf{y}_k - \mathbf{h}_{1,k} x_1 - \dots - \mathbf{h}_{n_r,k} x_{n_r}\|^2 \quad (21)$$

which has computational complexity $\mathcal{O}(|\chi_1| \dots |\chi_{n_r}|)$. For brevity we drop the frequency index k and the bit position index k' i.e.

$$\begin{aligned} \lambda_1^i(\mathbf{y}, c) &\approx \min_{x_1 \in \chi_{1,c}^i, x_2 \in \chi_2, \dots, x_{n_r} \in \chi_{n_r}} \|\mathbf{y} - \mathbf{h}_1 x_1 - \dots - \mathbf{h}_{n_r} x_{n_r}\|^2 \\ &= \min_{x_1 \in \chi_{1,c}^i, x_2 \in \chi_2, \dots, x_{n_r} \in \chi_{n_r}} \left\{ \|\mathbf{y}\|^2 + \sum_{j=1}^{n_r} \|\mathbf{h}_j x_j\|^2 + 2\Re \sum_{j=1}^{n_r-1} \sum_{l=j+1}^{n_r} (\mathbf{h}_j x_j)^\dagger (\mathbf{h}_l x_l) - 2\Re \sum_{j=1}^{n_r} (\mathbf{h}_j^\dagger \mathbf{y}) x_j \right\} \\ &= \min_{x_1 \in \chi_{1,c}^i, x_2 \in \chi_2, \dots, x_{n_r} \in \chi_{n_r}} \left\{ \|\mathbf{y}\|^2 + \sum_{j=1}^{n_r-1} \|\mathbf{h}_j x_j\|^2 + 2\Re \sum_{j=1}^{n_r-1} \sum_{l=j+1}^{n_r} p_{jl} x_j^* x_l - 2\Re \sum_{j=1}^{n_r-1} y_j x_j^* \right. \\ &\quad \left. + 2\Re \sum_{j=1}^{n_r-1} p_{jn_r} x_j^* x_{n_r} - 2\Re y_{n_r} x_{n_r}^* + \|\mathbf{h}_{n_r} x_{n_r}\|^2 \right\} \quad (22) \end{aligned}$$

where $y_k = \mathbf{h}_k^\dagger \mathbf{y}$ be the matched filter (MF) output for k -th stream and $p_{km} = \mathbf{h}_k^\dagger \mathbf{h}_m$ be the cross correlation between k -th and m -th channel. Breaking some of the terms in their real and

imaginary parts with subscripts $(\cdot)_R$ and $(\cdot)_I$ indicating real and imaginary parts of a complex number, we have

$$\begin{aligned} \lambda_1^i(\mathbf{y}, c) = \min_{x_1 \in \mathcal{X}_{1,c}^i \cdots x_{n_r} \in \mathcal{X}_{n_r}} & \left\{ \sum_{j=1}^{n_r-1} \|\mathbf{h}_j x_j\|^2 + 2\Re \sum_{j=1}^{n_r-1} \sum_{l=j+1}^{n_r-1} p_{jl} x_j^* x_l - 2\Re \sum_{j=1}^{n_r-1} y_j x_j^* \right. \\ & + \left(2 \sum_{j=1}^{n_r-1} (p_{jn_r,R} x_{j,R} + p_{jn_r,I} x_{j,I}) - 2y_{n_r,R} \right) x_{n_r,R} + \|\mathbf{h}_{n_r}\|^2 x_{n_r,R}^2 \\ & \left. + \left(2 \sum_{j=1}^{n_r-1} (p_{jn_r,R} x_{j,I} - p_{jn_r,I} x_{j,R}) - 2y_{n_r,I} \right) x_{n_r,I} + \|\mathbf{h}_{n_r}\|^2 x_{n_r,I}^2 \right\} \quad (23) \end{aligned}$$

This equation reduces one complex dimension of the system. For x_{n_r} belonging to equal energy alphabets, the bit metric is written as

$$\begin{aligned} \lambda_1^i(\mathbf{y}, c) = \min_{x_1 \in \mathcal{X}_{1,c}^i \cdots x_{n_r-1} \in \mathcal{X}_{n_r-1}} & \left\{ \sum_{j=1}^{n_r-1} \|\mathbf{h}_j x_j\|^2 + 2\Re \sum_{j=1}^{n_r-1} \sum_{l=j+1}^{n_r-1} p_{jl} x_j^* x_l - 2\Re \sum_{j=1}^{n_r-1} y_j x_j^* \right. \\ & \left. - \left| 2 \sum_{j=1}^{n_r-1} (p_{jn_r,R} x_{j,R} + p_{jn_r,I} x_{j,I}) - 2y_{n_r,R} \right| |x_{n_r,R}| - \left| 2 \sum_{j=1}^{n_r-1} (p_{jn_r,R} x_{j,I} - p_{jn_r,I} x_{j,R}) - 2y_{n_r,I} \right| |x_{n_r,I}| \right\} \end{aligned}$$

For x_{n_r} belonging to non-equal energy alphabets, it's real and imaginary part which minimizes (23) are given as

$$\begin{aligned} x_{n_r,R} & \rightarrow - \frac{\sum_{j=1}^{n_r-1} (p_{jn_r,R} x_{j,R} + p_{jn_r,I} x_{j,I}) - y_{n_r,R}}{\|\mathbf{h}_{n_r}\|^2} \\ x_{n_r,I} & \rightarrow - \frac{\sum_{j=1}^{n_r-1} (p_{jn_r,R} x_{j,I} - p_{jn_r,I} x_{j,R}) - y_{n_r,I}}{\|\mathbf{h}_{n_r}\|^2} \quad (24) \end{aligned}$$

where \rightarrow indicates the quantization process in which amongst the finite available points, the point closest to the calculated continuous value is selected.

This bit metric implies reduction in the complexity to $\mathcal{O}(|\mathcal{X}_1| \cdots |\mathcal{X}_{n_r-1}|)$. Reduction of one complex dimension without any additional processing is a fundamental result of significant importance for lower dimensional systems. Additionally this bit metric is based on MF outputs and channel correlations and is therefore simpler for fixed point implementations. The intricacy in the practical implementation of a higher dimensional MIMO system due to space (requisite antenna spacing) and technology constraints underlines the significance of complexity reduction algorithms for lower dimensional systems. MMSE based demodulators involve computationally complex operations of matrix inversions which are very hard for fixed point implementations. Moreover MMSE demodulator additionally needs the knowledge of noise variance.

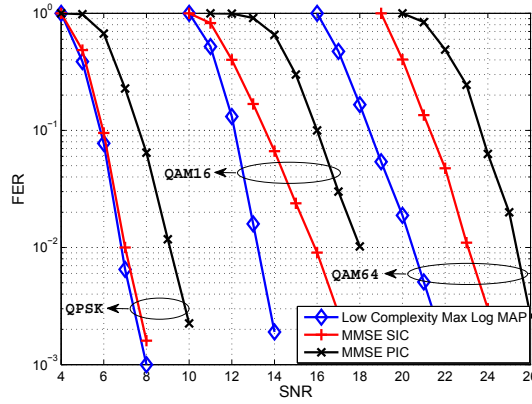


Fig. 7. 2×2 system with uniform rate and nonuniform power spatial streams. For QPSK $\sigma_1^2 = 0.63P_T, \sigma_2^2 = 0.37P_T$, for QAM 16 $\sigma_1^2 = 0.67P_T, \sigma_2^2 = 0.33P_T$ while for QAM64 $\sigma_1^2 = 0.70P_T, \sigma_2^2 = 0.30P_T$.

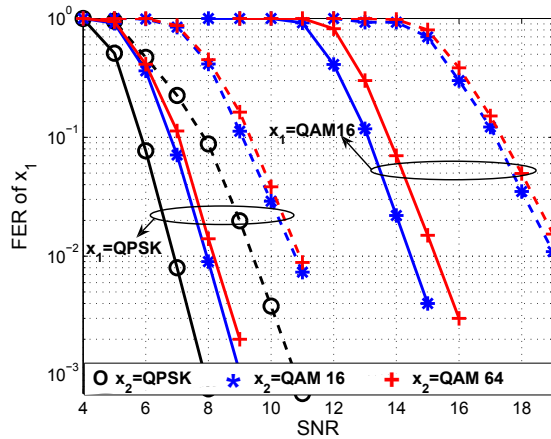


Fig. 8. Performance of lower rate stream in 2×2 BICM MIMO OFDM system using 802.11n convolutional code. Continuous lines indicate low complexity max log MAP detector while dashed lines indicate MMSE detector.

2.5 Simulations

We consider a 2×2 BICM MIMO OFDM system using the *de facto* standard, 64 state rate-1/2 convolutional encoder of 802.11n standard (802.11n, 2006) and rate-1/2 punctured turbo code

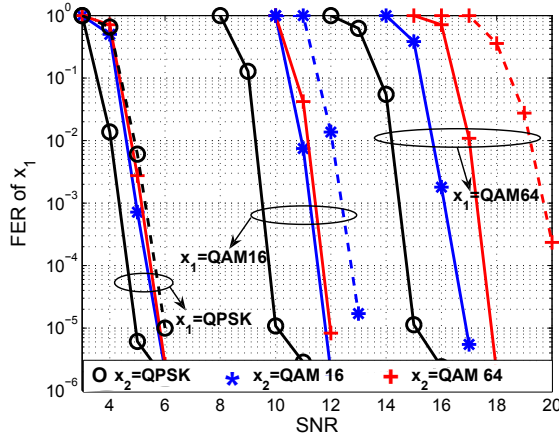


Fig. 9. Performance of lower rate stream in 2×2 BICM MIMO OFDM system using 3GPP LTE turbo code. Continuous lines indicate low complexity max log MAP detector while dashed lines indicate MMSE detector. Block length of the lower rate stream is 1296 bits while number of decoding iterations are 5.

proposed for 3GPP LTE (LTE, 2006)¹. MIMO channel has iid Gaussian matrix entries with unit variance. The channel is independently generated for each time instant and perfect CSI at the receiver is assumed. Furthermore, all mappings of coded bits to QAM symbols use Gray encoding. We consider MMSE and low complexity max log MAP detector. There are two scenarios.

In first scenario, spatial streams of uniform rate and nonuniform power are transmitted in 2×2 MIMO broadcast system. The upcoming WLAN standard 802.11n (802.11n, 2006) supports the codeword sizes of 648, 1296, and 1944 bits. For our purposes, we selected the codeword size of 1296 bits and coding scheme of convolutional coding. We focus on the frame error rates (FER) of the system. We consider the low complexity max log MAP and MMSE SIC approach in which the higher power stream is detected first and is subsequently stripped off leading to the detection of lower power stream. With P_T being the total power available, the power distribution between two streams is optimized to equate their rates in the desired SNR region where SNR is defined as the received SNR per antenna i.e. $\frac{P_T}{N_0}$. As a reference, MMSE parallel interference cancellation (PIC) has also been simulated in which two streams are independently detected using MMSE filters and two streams have equal power. Fig. 7 shows the improved performance of low complexity max log MAP approach with respect to both MMSE SIC and MMSE PIC approach. The gap widens as the constellation proliferates i.e. QAM 16 and QAM64 which is attributed to the higher suboptimality of MMSE for larger sized constellations.

In second scenario, spatial streams of uniform power and nonuniform rate are transmitted in 2×2 MIMO broadcast system. We focus on the FER of first stream (lower rate) as subsequent

¹ LTE turbo decoder design was performed using the coded modulation library www.iterativesolutions.com

to stripping, the detection of second stream (higher rate) is trivial (using SIMO detectors). The frame length of first stream is fixed to 1296 information bits as per 802.11n (802.11n, 2006). Figs. 8 and 9 compare the performance of low complexity max log MAP detector with MMSE detector. The max log MAP detector performs significantly better than the MMSE detector. Degradation of the performance for first stream as the rate (constellation size) of second stream increases confirms the earlier result of sec. 2.2.2 that rate of first stream is a function of the rate of second stream.

3. Interference Suppression for future Wireless Systems

To cope with the ever-increasing demands on the higher spectral efficiency, appendage of spatial dimension (MIMO) needs to be coupled with a tight frequency reuse as is advocated in the future wireless communication systems as 3GPP LTE (LTE, 2006) and LTE-Advanced (LTE-A, 2008). Adaptive modulation and coding schemes will be supported in the next generation wireless systems which combined with the diversified data services will lead to variable transmission rate streams. These system characteristics will overall lead to an interference-limited system. Most state-of-the-art wireless systems deal with the interference either by orthogonalizing the communication links in time or frequency (Gesbert et al., 2007) or allow the communication links to share the same degrees of freedom but model the interference as additive Gaussian random process (Russell & Stuber, 1995). Both of these approaches may be suboptimal as first approach entails an *a priori* loss of the degrees of freedom in both links, independent of the interference strength while second approach treats the interference as pure noise while it actually carries information and has the structure that can be potentially exploited in mitigating its effect.

3GPP LTE (LTE, 2006) has chosen orthogonal frequency division multiple access (OFDMA) technology for the downlink in order to provide multiple access and eliminate the intracell interference. However frequency reuse factor being 1 will lead to intercell interference impairments among neighboring cells. Intercell interference coordination techniques (Gesbert et al., 2007) are studied to minimize the interference level while spatial interference cancellation filters are the focus of attention to cancel the interferers which will be 1 in most cases (near cell boundaries) and 2 in rare cases (near cell corners). Different spatial interference cancellation techniques involving equalization and subtractive cancellation (Bladsjö et al., 1999) (Debbah et al., 2000) have been proposed in the literature. Amongst them, MMSE linear detectors are being considered as likely candidates for 3GPP LTE (Dahlman et al., 2006). The suboptimality of MMSE for non Gaussian alphabets in low dimensional systems (less number of interferers) has already been discussed and simulated in the previous sections and moreover MMSE detection being based on interference attenuation is void of exploiting the interference structure in mitigating its effect. Though not optimal, but their low complexity still makes them attractive for practical systems.

Optimal strategy for treating the interference in the regime of very strong (Carleial, 1975) and very weak interference is well known however if the interference is in the moderate region, no optimal strategy is known but partial decoding of interference can significantly improve performance (Han & Kobayashi, 1981). This part of the chapter discusses a low complexity spatial interference cancellation algorithm for single frequency reuse synchronized cellular networks in the presence of one strong interferer. This algorithm is based on the low complexity max log MAP detector and benefits from its ability to exploit interference structure in mitigating its effect. The algorithm encompasses two strategies for interference mitigation i.e. interference suppression and interference cancellation and their selection in the receiver is dictated

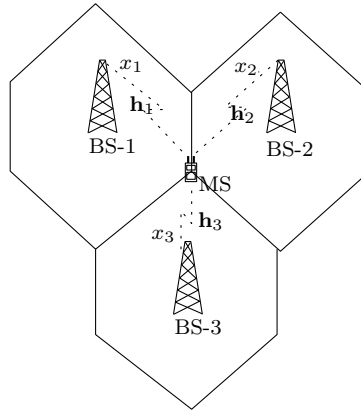


Fig. 10. Interference cancellation in single frequency cellular network. x_1 is the desired signal while x_2 and x_3 are the interference signals.

by the relative strength and the rate of interfering stream. In the scenario of interfering stream being weak or of higher rate relative to the desired stream, thereby making it unfeasible to be decoded, the mobile station (MS) resorts to the strategy of interference suppression. It can be interpreted as partial decoding of the interference which is the recommended strategy in the regime of moderate interference (Han & Kobayashi, 1981). When the interfering stream is relatively stronger or is of lower rate thereby making it feasible to be decoded, MS adopts interference cancellation strategy (subtractive cancellation) which is the optimal strategy in the case of strong interference (Carleial, 1975).

3.1 System Model

The system model as shown in fig. 10 remains same as described in the sec.2.1.2 with 3 spatial streams. However these streams arriving at the receiver (MS) are now from three different base stations (BS) thereby ensuring independent channels. The MS has receive diversity with n_r receive antennas. All the BSs are assumed to be synchronous.

3.2 Information Theoretic view

For better understanding of the effect of strength and rate (alphabet size) of interference, the case of one strong interference is considered in this section. The focus is on the mutual information of the desired stream in the presence of one strong interferer (Ghaffar & Knopp, 2009b).

$$I(\mathbf{y}; x_1) = \log M_1 - \frac{1}{M_1} \sum_{x_1} \int_{\mathbf{y}} p(\mathbf{y}|x_1) \log \frac{\sum_{x_1} p(\mathbf{y}|x_1)}{p(\mathbf{y}|x_1)} d\mathbf{y} \quad (25)$$

Fig. 11 shows the mutual information of the desired stream in the presence of the interference stream. We define the term $\alpha = \sigma_2^2 / \sigma_1^2$. Mutual information of the desired stream is a function of the rate as well as the strength of the interference stream. For moderate values of α and when the interference has a lower rate (smaller constellation size) relative to the desired stream, as the interference strength increases, the mutual information of the desired stream

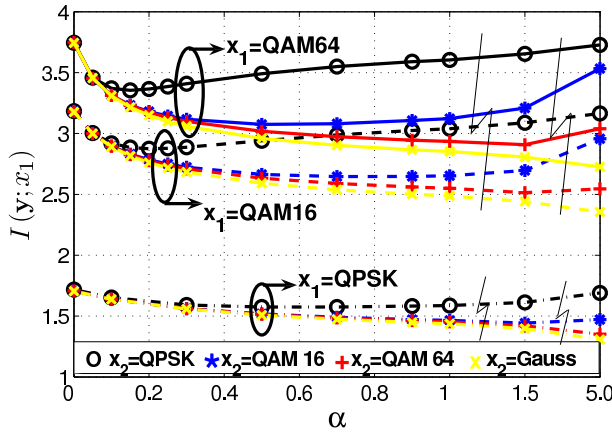


Fig. 11. Mutual information of the desired stream x_1 in the presence of the interference stream x_2 for different constellations. SNR is 4.5 dB for x_1 =QPSK, 11 dB for x_1 =QAM16 and 13 dB for x_1 =QAM64. Note that the flash sign indicates a discontinuity of abscissa.

increases. However when the interference stream has a higher rate as compared to the rate of the desired stream, this behavior is observed for higher values of α . This can be interpreted as the decoding capability of the MS of the interference in the presence of the desired stream. Once the interference strength and its rate relative to the strength and the rate of the desired stream permits the decoding of the interference, we observe an increase in the mutual information of the desired stream with the increase of α . Fig. 11 also authenticates the well known result of Gaussian being the worst case interference however the gap decreases as the rate of the interference stream increases. This diminution of gap may be related to the proximity of the behavior of large size constellations to Gaussianity as both are characterized by high peak to average power ratios.

3.3 Interference Mitigation Strategies

Based on the low complexity max log MAP detector, an interference mitigation strategy (Ghafar & Knopp, 2009a) is discussed which is based on the partial decoding of the interference in the regime when interference because of its relative rate or strength is undecodable and subtractive cancellation when the interference is quite strong and is decodable. This strategy is based on exploiting the structure of the interference in mitigating its effect once subtractive cancellation is not possible and resorting to subtractive cancellation otherwise. So there are two options for interference mitigation.

1. In the regime when interference has higher rate or is weaker in strength relative to the desired stream thereby rendering the absolute decoding of interference unfeasible, target stream is decoded using the low complexity max log MAP detector which takes into account the effect of interference and can be termed as the partial decoding of interference or partial joint decoding. This approach is termed as interference suppression.
2. In the regime when interference has lower rate or is stronger in strength relative to the desired stream thereby rendering the absolute decoding of the interference feasible, the

interference stream is decoded using low complexity max log MAP detector, stripping it off and then decoding the desired stream. This approach is termed as interference cancellation.

The factors that will decide the strategy to be adopted will be the relative rate and the strength of the interference stream comparative to the desired stream. The requisites for this algorithm are the knowledge of interference channel and the modulation and coding scheme (MCS) of interfering stream. The BSs need to be synchronous with pilot signals from the adjacent BSs to be orthogonal to meet these requisites.

3.4 Performance Analysis

This section deliberates on the performance analysis of two detectors for detecting the desired stream in the presence of interfering stream (Ghaffar & Knopp, 2009c).

3.4.1 PEP Analysis - Max Log MAP Detector

The conditional PEP i.e. $P(\underline{\mathbf{c}}_1 \rightarrow \hat{\underline{\mathbf{c}}}_1 | \mathbf{H})$ of max log MAP detector is given as

$$P(\underline{\mathbf{c}}_1 \rightarrow \hat{\underline{\mathbf{c}}}_1 | \mathbf{H}) = P \left(\sum_{k'} \min_{x_1 \in \mathcal{X}_{1,c_{k'}}, x_2 \in \mathcal{X}_2} \frac{1}{N_0} \|\mathbf{y}_k - \mathbf{h}_{1,k}x_1 - \mathbf{h}_{2,k}x_2\|^2 \geq \sum_{\bar{k}} \min_{x_1 \in \mathcal{X}_{1,\bar{c}_{\bar{k}}}, x_2 \in \mathcal{X}_2} \frac{1}{N_0} \|\mathbf{y}_k - \mathbf{h}_{1,k}x_1 - \mathbf{h}_{2,k}x_2\|^2 \right) \quad (26)$$

where $\mathbf{H} = [\mathbf{H}_1 \cdots \mathbf{H}_K]$ i.e. the complete channel for the transmission of the codeword $\underline{\mathbf{c}}_1$ and $\mathbf{H}_k = [\mathbf{h}_{1,k} \ \mathbf{h}_{2,k}]$ i.e. the channel at k -th frequency tone. For the worst case scenario once $d(\underline{\mathbf{c}}_1 - \hat{\underline{\mathbf{c}}}_1) = d_{free}$, the inequality on the right hand side of (26) shares the same terms on all but d_{free} summation points for which $\hat{c}_{k'} = \bar{c}_{k'}$ where $(\bar{\cdot})$ denotes the binary complement. Let

$$\begin{aligned} \tilde{x}_{1,k}, \tilde{x}_{2,k} &= \arg \min_{x_1 \in \mathcal{X}_{1,\bar{c}_{k'}}, x_2 \in \mathcal{X}_2} \frac{1}{N_0} \|\mathbf{y}_k - \mathbf{h}_{1,k}x_1 - \mathbf{h}_{2,k}x_2\|^2 \\ \hat{x}_{1,k}, \hat{x}_{2,k} &= \arg \min_{x_1 \in \mathcal{X}_{1,c_{k'}}, x_2 \in \mathcal{X}_2} \frac{1}{N_0} \|\mathbf{y}_k - \mathbf{h}_{1,k}x_1 - \mathbf{h}_{2,k}x_2\|^2 \end{aligned} \quad (27)$$

As $x_{1,k}$ and $x_{2,k}$ are the transmitted symbols so $\|\mathbf{y}_k - \mathbf{h}_{1,k}x_{1,k} - \mathbf{h}_{2,k}x_{2,k}\|^2 \geq \|\mathbf{y}_k - \mathbf{h}_{1,k}\tilde{x}_{1,k} - \mathbf{h}_{2,k}\tilde{x}_{2,k}\|^2$. The conditional PEP is given as

$$\begin{aligned} P(\underline{\mathbf{c}}_1 \rightarrow \hat{\underline{\mathbf{c}}}_1 | \mathbf{H}) &\leq P \left(\sum_{k,d_{free}} \frac{1}{N_0} \|\mathbf{y}_k - \mathbf{h}_{1,k}x_{1,k} - \mathbf{h}_{2,k}x_{2,k}\|^2 \geq \sum_{k,d_{free}} \frac{1}{N_0} \|\mathbf{y}_k - \mathbf{h}_{1,k}\hat{x}_{1,k} - \mathbf{h}_{2,k}\hat{x}_{2,k}\|^2 \right) \\ &= Q \left(\sqrt{\sum_{k,d_{free}} \frac{1}{2N_0} \|\mathbf{H}_k(\hat{\mathbf{x}}_k - \mathbf{x}_k)\|^2} \right) \\ &= Q \left(\sqrt{\frac{1}{2N_0} \text{vec}(\bar{\mathbf{H}}^\dagger)^\dagger \Delta \text{vec}(\bar{\mathbf{H}}^\dagger)} \right) \end{aligned} \quad (28)$$

where $\bar{\mathbf{H}} = [\mathbf{H}_1 \cdots \mathbf{H}_{k,d_{free}}]$, $\hat{\mathbf{x}}_k = [\hat{x}_{1,k} \hat{x}_{2,k}]^T$ and $\Delta = \mathbf{I}_{n_r} \otimes \mathbf{D}\mathbf{D}^\dagger$ while $\mathbf{D} = \text{diag} \left\{ \hat{\mathbf{x}}_1 - \mathbf{x}_1, \hat{\mathbf{x}}_2 - \mathbf{x}_2, \cdots, \hat{\mathbf{x}}_{k,d_{free}} - \mathbf{x}_{k,d_{free}} \right\}$. Q is the Gaussian Q-function i.e. $Q(y) = \frac{1}{\sqrt{2\pi}} \int_y^\infty e^{-x^2/2} dx$ and vec indicates vectorization of a matrix. For a Hermitian quadratic form in complex Gaussian random variable $q = \mathbf{m}^\dagger \mathbf{A} \mathbf{m}$ where \mathbf{A} is a Hermitian matrix and column vector \mathbf{m} is a circularly symmetric complex Gaussian vector i.e. $\mathbf{m} \sim \mathcal{NC}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = E[\mathbf{m}]$ and $\boldsymbol{\Sigma} = E[\mathbf{m}\mathbf{m}^\dagger] - \boldsymbol{\mu}\boldsymbol{\mu}^\dagger$, the moment generating function (MGF) is

$$E \left[\exp \left(-t \mathbf{m}^\dagger \mathbf{A} \mathbf{m} \right) \right] = \frac{\exp \left[-t \boldsymbol{\mu}^\dagger \mathbf{A} (\mathbf{I} + t \boldsymbol{\Sigma} \mathbf{A})^{-1} \boldsymbol{\mu} \right]}{\det (\mathbf{I} + t \boldsymbol{\Sigma} \mathbf{A})} \quad (29)$$

Using Chernoff bound $Q(x) \leq \frac{1}{2} \exp \left(-\frac{x^2}{2} \right)$ and the MGF, PEP is upper bounded as

$$\begin{aligned} P(\mathbf{c}_1 \rightarrow \hat{\mathbf{c}}_1) &\leq \frac{1}{2 \det \left(\mathbf{I} + \frac{1}{4N_0} \mathbf{I} \Delta \right)} \\ &= \frac{1}{2 \prod_{k=1}^{d_{free}} \left(1 + \frac{1}{4N_0} \|\hat{\mathbf{x}}_k - \mathbf{x}_k\|^2 \right)^{n_r}} \end{aligned} \quad (30)$$

$\|\hat{\mathbf{x}}_k - \mathbf{x}_k\|^2 \geq d_{1,\min}^2 + d_{2,\min}^2$ if $\hat{x}_{2,k} \neq x_{2,k}$ and $\|\hat{\mathbf{x}}_k - \mathbf{x}_k\|^2 \geq d_{1,\min}^2$ if $\hat{x}_{2,k} = x_{2,k}$. There exists $2^{d_{free}}$ possible vectors of $[\hat{x}_{2,1}, \cdots, \hat{x}_{2,d_{free}}]^T$ basing on the binary criteria that $\hat{x}_{2,k}$ is equal or not equal to $x_{2,k}$. Taking into account all these cases combined with their corresponding probabilities, the PEP is upper bounded as

$$P(\mathbf{c}_1 \rightarrow \hat{\mathbf{c}}_1 | \mathbf{H}) \leq \frac{1}{2} \left(\frac{4N_0}{\sigma_1^2 d_{1,\min}^2} \right)^{n_r d_{free}} \left(\sum_{j=0}^{d_{free}} C_j^{d_{free}} \frac{(P(\hat{x}_{2,k} \neq x_{2,k}))^j (1 - P(\hat{x}_{2,k} \neq x_{2,k}))^{d_{free}-j}}{\left(1 + \frac{\sigma_2^2 d_{2,\min}^2}{\sigma_1^2 d_{1,\min}^2} \right)^{j n_r}} \right) \quad (31)$$

where $d_{j,\min}^2 = \sigma_j^2 d_{j,\min}^2$ with $d_{j,\min}^2$ being the normalized minimum distance of the constellation χ_j for $j = \{1, 2\}$ and $C_j^{d_{free}}$ is the binomial coefficient. $P(\hat{x}_{2,k} \neq x_{2,k})$ has been derived in the following section.

3.4.2 $P(\hat{x}_{2,k} \neq x_{2,k})$

Considering (27), $P(\hat{x}_{2,k} \neq x_{2,k} | \mathbf{h}_{1,k}, \mathbf{h}_{2,k}, x_{1,k})$ is

$$\begin{aligned} &P(\hat{x}_{2,k} \neq x_{2,k} | \mathbf{h}_{1,k}, \mathbf{h}_{2,k}, x_{1,k}) \\ &= P \left(-2\Re \left((\mathbf{h}_{1,k} (x_{1,k} - x_1) + \mathbf{z}_k)^\dagger \mathbf{h}_{2,k} (x_{2,k} - x_2) \right) \geq \|\mathbf{h}_{2,k} (x_{2,k} - x_2)\|^2 | \mathbf{H}_k, x_{1,k} \right) \\ &= Q \left(\sqrt{\frac{\|\mathbf{h}_{2,k} (x_{2,k} - x_2)\|^2}{2N_0}} + \sqrt{\frac{2}{N_0}} \Re \left(\frac{(\mathbf{h}_{1,k} (x_{1,k} - x_1) + \mathbf{z}_k)^\dagger \mathbf{h}_{2,k} (x_{2,k} - x_2)}{\sqrt{\|\mathbf{h}_{2,k} (x_{2,k} - x_2)\|^2}} \right) \right) \end{aligned}$$

Using the relation $Q(a+b) \leq Q(a_{\min} - |b_{\max}|)$ and $\Re(\mathbf{a}^\dagger \hat{\mathbf{b}}) \leq \|\mathbf{a}\|$ where $\hat{\mathbf{b}}$ is the unit vector we get

$$P(\hat{x}_{2,k} \neq x_{2,k} | \mathbf{h}_{1,k}, \mathbf{h}_{2,k}) \leq \frac{1}{2} \exp \left(-\frac{\|\mathbf{h}_{2,k}\|^2 d_{2,\min}^2}{4N_0} - \frac{\|\mathbf{h}_{1,k}\|^2 d_{1,\max}^2}{N_0} + \frac{\|\mathbf{h}_{2,k}\| \|\mathbf{h}_{1,k}\| d_{2,\min} d_{1,\max}}{N_0} \right) \quad (32)$$

Conditioned on the norm of $\mathbf{h}_{1,k}$ we make two non-overlapping regions as $(\|\mathbf{h}_{2,k}\| \geq \|\mathbf{h}_{1,k}\| \|\mathbf{h}_{1,k}\|)$ and $(\|\mathbf{h}_{2,k}\| < \|\mathbf{h}_{1,k}\| \|\mathbf{h}_{1,k}\|)$ with the corresponding probabilities as $\mathcal{P}_{\mathbf{h}_1}^<$ and $\mathcal{P}_{\mathbf{h}_1}^>$. Note that in first region $\|\mathbf{h}_{2,k}\| \|\mathbf{h}_{1,k}\| \leq \|\mathbf{h}_{2,k}\|^2$ while for second region $\|\mathbf{h}_{2,k}\| \|\mathbf{h}_{1,k}\| < \|\mathbf{h}_{1,k}\|^2$. So

$$\begin{aligned} P(\hat{x}_{2,k} \neq x_{2,k}) &\leq \frac{1}{2} E_{\mathbf{h}_1} \left[\left(\frac{4N_0}{d_{2,\min}^2 - 4d_{2,\min} d_{1,\max}} \right)^{n_r} \exp \left(-\frac{\|\mathbf{h}_{1,k}\|^2 d_{1,\max}^2}{N_0} \right) E_{\mathbf{h}_2 | \mathbf{h}_1}(\mathcal{P}_{\mathbf{h}_1}^<) \right. \\ &\quad \left. + \left(\frac{4N_0}{d_{2,\min}^2} \right)^{n_r} \exp \left(-\|\mathbf{h}_{1,k}\|^2 \frac{d_{1,\max}^2 - d_{2,\min} d_{1,\max}}{N_0} \right) E_{\mathbf{h}_2 | \mathbf{h}_1}(\mathcal{P}_{\mathbf{h}_1}^>) \right] \\ &\leq \frac{1}{2} \left(\frac{4N_0}{\sigma_2^2 d_{2,\min}^2} \right)^{n_r} \left(\frac{N_0}{\sigma_1^2 d_{1,\max}^2} \right)^{n_r} \left(\frac{1}{\left(1 - \frac{4\sigma_1 d_{1,\max}}{\sigma_2 d_{2,\min}}\right)^{n_r}} + \frac{1}{\left(1 - \frac{\sigma_2 d_{2,\min}}{\sigma_1 d_{1,\max}}\right)^{n_r}} \right) \end{aligned} \quad (33)$$

where we upper bound $E_{\mathbf{h}_2 | \mathbf{h}_1}(\mathcal{P}_{\mathbf{h}_1}^<)$ and $E_{\mathbf{h}_2 | \mathbf{h}_1}(\mathcal{P}_{\mathbf{h}_1}^>)$ by 1.

$P(\hat{x}_{2,k} \neq x_{2,k}) \rightarrow 0$ as $\sigma_2^2 \rightarrow \infty$ while $P(\hat{x}_{2,k} \neq x_{2,k})$ increases as σ_2^2 increases. Eq. (31) demonstrates a significant result of achieving full diversity by the low complexity max log MAP detector and converging to the performance of single stream using maximum ratio combining in the case of very weak and strong interference. In the moderate region, as the strength of interference increases, $P(\hat{x}_{2,k} \neq x_{2,k})$ reduces and there is a coding gain for the detection of desired stream contrary to the case of MMSE where there is a coding loss as the interference gets stronger (shown in the next section).

3.4.3 PEP Analysis - MMSE Detector

3.5 Gaussian Assumption

Conditional PEP for MMSE basing on Gaussian assumption of post detection interference (20) is given as

$$P(\underline{\mathbf{c}}_1 \rightarrow \hat{\mathbf{c}}_1 | \mathbf{H}) = P \left(\sum_{k'} \min_{x_1 \in \mathcal{X}_{1,c_{k'}}} \frac{|y_k - \alpha_k x_1|^2}{N_k} \geq \sum_{k'} \min_{x_1 \in \mathcal{X}_{1,\hat{c}_{k'}}} \frac{|y_k - \alpha_k x_1|^2}{N_k} \right) \quad (34)$$

Let

$$\tilde{x}_{1,k} = \arg \min_{x_1 \in \mathcal{X}_{1,c_{k'}}} \frac{|y_k - \alpha_k x_1|^2}{N_k}, \quad \hat{x}_{1,k} = \arg \min_{x_1 \in \mathcal{X}_{1,\hat{c}_{k'}}} \frac{|y_k - \alpha_k x_1|^2}{N_k}$$

Considering the worst case scenario $d(\underline{\mathbf{c}}_1 - \hat{\underline{\mathbf{c}}}_1) = d_{free}$ and using the fact that $\frac{1}{N_k} |y_k - \alpha_k x_{1,k}|^2 \geq \frac{1}{N_k} |y_k - \alpha_k \tilde{x}_{1,k}|^2$, the conditional PEP is upper bounded as

$$P(\underline{\mathbf{c}}_1 \rightarrow \hat{\underline{\mathbf{c}}}_1 | \mathbf{H}) \leq Q \left(\sqrt{\sum_{k, d_{free}} \frac{\alpha_k^2}{2N_k} |\hat{x}_{1,k} - x_{1,k}|^2} \right) \quad (35)$$

Bounding $|\hat{x}_{1,k} - x_{1,k}|^2 \geq d_{1,\min}^2$ and using the Chernoff bound

$$P(\underline{\mathbf{c}}_1 \rightarrow \hat{\underline{\mathbf{c}}}_1 | \mathbf{H}) \leq \frac{1}{2} \exp \left(-\frac{d_{1,\min}^2}{4} \sum_{k, d_{free}} \mathbf{h}_{1,k}^\dagger \mathbf{R}_{2,k}^{-1} \mathbf{h}_{1,k} \right) \quad (36)$$

where the summation in (36) can be written as

$$\sum_{k, d_{free}} \mathbf{h}_{1,k}^\dagger \mathbf{R}_{2,k}^{-1} \mathbf{h}_{1,k} = [\mathbf{h}_{1,1}^\dagger, \dots, \mathbf{h}_{1,d_{free}}^\dagger] \text{diag} [\mathbf{R}_{2,1}^{-1}, \dots, \mathbf{R}_{2,d_{free}}^{-1}] [\mathbf{h}_{1,1}^T, \dots, \mathbf{h}_{1,d_{free}}^T]^T$$

The eigenvalues of $\mathbf{R}_{2,k}^{-1}$ are

$$\lambda_l = \begin{cases} (\sigma_2^2 \|\mathbf{h}_{2,k}\|^2 + N_0)^{-1}, & l = 1 \\ N_0^{-1}, & l = 2, \dots, n_r \end{cases} \quad (37)$$

Using the MGF (29), PEP conditioned on $\bar{\mathbf{h}}_2 = [\mathbf{h}_{2,1}, \dots, \mathbf{h}_{2,d_{free}}]$ is upper bounded as

$$P(\underline{\mathbf{c}}_1 \rightarrow \hat{\underline{\mathbf{c}}}_1 | \bar{\mathbf{h}}_2) \leq \frac{1}{2} \left(\frac{4N_0}{d_{1,\min}^2} \right)^{d_{free}(n_r-1)} \left(\frac{4}{d_{1,\min}^2} \right)^{d_{free}} \prod_{l=1}^{d_{free}} (\sigma_2^2 \|\mathbf{h}_{2,l}\|^2 + N_0)$$

Channel independence at each subcarrier yields

$$P(\underline{\mathbf{c}}_1 \rightarrow \hat{\underline{\mathbf{c}}}_1) \leq \frac{1}{2} \left(\frac{4N_0}{\sigma_1^2 d_{1,\min}^2} \right)^{d_{free}(n_r-1)} \left(\frac{4}{\sigma_1^2 d_{1,\min}^2} \right)^{d_{free}} (n_r \sigma_2^2 + N_0)^{d_{free}} \quad (38)$$

which not only demonstrates the well known result of the loss of one diversity order in MMSE in the presence of one interferer (Winters, 1984) but also exhibits a coding loss as interference gets stronger.

3.6 Simulation Results

Moderate and high SNR regime in the interference-limited scenario demands more attention as when the noise is small, interference will have a significant impact on the performance. Low SNR regime is less interesting since here the performance is noise-limited and interference is not having a significant effect. For simulations, we have restricted ourselves to the case of one strong interference. These simulations have been performed in moderate and high SNR region while the interference strength is being varied.

We consider 2 BSs each using BICM OFDM system for downlink transmission using the *de facto* standard, 64 state (133, 171) rate-1/2 convolutional encoder of 802.11n standard (802.11n,

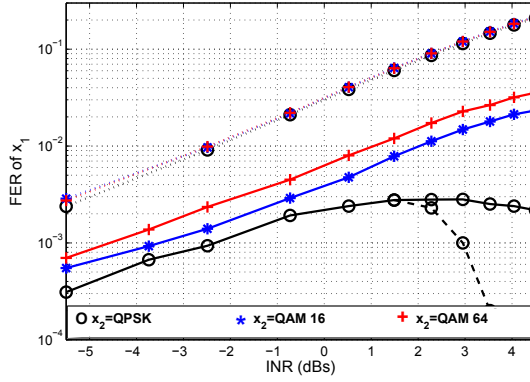


Fig. 12. Desired stream x_1 is QPSK while interference stream x_2 is from QPSK, QAM16 and QAM64. SNR is 4.5 dB. Continuous lines indicate interference suppression while dashed lines indicate interference cancellation. Dotted lines indicates detection of x_1 by MMSE detector. 64-state, rate 1/2 Convolutional Code is used. Note that SNR is with respect to the desired stream

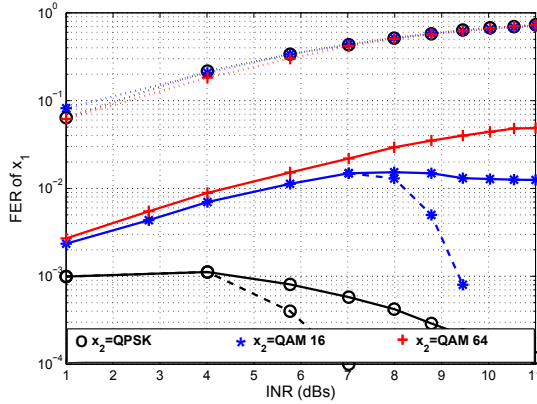


Fig. 13. Desired stream x_1 is QAM 16 while interference stream x_2 is from QPSK, QAM16 and QAM64. SNR is 11 dB. Continuous lines indicate interference suppression while dashed lines indicate interference cancellation. Dotted lines indicates detection of x_1 by MMSE detector. 64-state, rate 1/2 Convolutional Code is used.

2006) and the punctured rate 1/2 turbo code of 3GPP LTE (LTE, 2006)². Each BS has multiple antennas and employs antenna cycling. MS has two antennas. We consider an ideal OFDM

² The LTE turbo decoder design was performed using the coded modulation library www.iterativesolutions.com

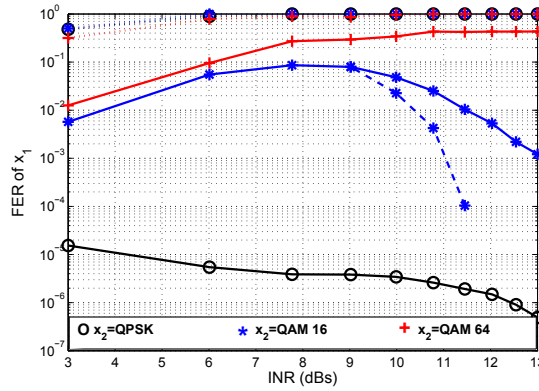


Fig. 14. Desired stream x_1 is QAM 64 while interference stream x_2 is from QPSK, QAM16 and QAM64. SNR is 13 dB. Continuous lines indicate interference suppression while dashed lines indicate interference cancellation. Dotted lines indicates detection of x_1 by MMSE detector. Punctured rate 1/2 3GPP turbo code is used with 5 decoding iterations.

based system (no ISI) and analyze the system in frequency domain. Due to bit interleaving followed by OFDM, this can be termed as frequency interleaving. Therefore SIMO channel at each sub carrier from BS to MS has iid Gaussian matrix entries with unit variance. Perfect CSI is assumed at the receiver. Furthermore, all mappings of coded bits to QAM symbols use Gray encoding. We consider interference suppression and interference cancellation approaches using low complexity max log MAP detector. For comparison we also consider interference suppression using MMSE detector.

Figs. 12, 13 and 14 show the FERs of target stream in the presence of one interference stream. These simulation results show that the dependence of the performance for MMSE detection is insignificant on the rate of the interference stream but its dependence on interference strength is substantial. This can be interpreted as a consequence of the attenuation of interference strength at the output of MMSE filter and the subsequent assumption of Gaussianity for its behavior. For the low complexity max log MAP detector, a significant improvement is observed in the performance as the rate of interference stream decreases which is in conformity with the earlier results of mutual information analysis (fig. 11). It is observed that for a given interference level, the performance is generally degraded as the rate (constellation size) of the interfering stream increases. The performance gap with respect to MMSE decreases as the desired and the interference streams grow in constellation size which can be attributed to the proximity to the Gaussianity of these larger constellations due to their high peak to average power ratio and to the optimality of MMSE for Gaussian alphabets.

4. References

- 802.11n (2006). *Enhancements for Higher Throughput*, IEEE 802.11 WG. IEEE 802.11n/D1.0 Draft Amendment.
- 802.16m (2007). *Draft IEEE 802.16m Evaluation Methodology*, IEEE 802.16m-07/037r1.

- Alamouti, S. (1998). A simple transmit diversity technique for wireless communications, *IEEE Journal on Selected Areas in Communications* **Vol. 16**(No. 8): 1451–1458.
- Bladsjö, D., Furuskär, A., Jäverbring, S. & Larsson, E. (1999). Interference cancellation using antenna diversity for EDGE - Enhanced data rates in GSM and TDMA/136, *IEEE Vehicular Technology Conference Proceedings, 1999. VTC 1999-Fall*, Vol. 4, pp. 1956–1960.
- Caire, G., Taricco, G. & Biglieri, E. (1998). Bit-interleaved coded modulation, **Vol. 44**(No. 3): 927–946.
- Carleial, A. (1975). A case where Interference does not reduce Capacity, *IEEE Transactions on Information Theory* **Vol. 21**(No. 5): 569–570.
- Dahlman, E., Ekstrom, H., Furuskär, A., Jading, Y., Karlsson, J., Lundevall, M. & Parkvall, S. (2006). The 3G long-term evolution - radio interface concepts and performance evaluation, *IEEE 63rd Vehicular Technology Conference VTC-Spring*, Vol. 1, pp. 137–141.
- Debbah, M., Muquet, B., de Courville, M., Muck, M., Simoens, S. & Loubaton, P. (2000). A MMSE successive interference cancellation scheme for a new adjustable hybrid spread OFDM system, *IEEE 51st Vehicular Technology Conference Proceedings, 2000. VTC 2000-Spring Tokyo*, Vol. 2, pp. 745–749 vol.2.
- Foschini, G. J. & Gans, M. J. (1998). On limits of wireless communication in a fading environment when using multiple antennas, *Wireless Personal Communications*. **Vol. 6**(No. 3): 311–335.
- Gesbert, D., Kiani, S., Gjendemsj, A. & Oien, G. (2007). Adaptation, coordination, and distributed resource allocation in interference-limited wireless networks, *Proceedings of the IEEE* **Vol. 95**(No. 12): 2393–2409.
- Ghaffar, R. & Knopp, R. (2008a). Dual-antenna BICM reception with applications to MIMO broadcast and single frequency cellular system, *IEEE 19-th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2008)*, Cannes.
- Ghaffar, R. & Knopp, R. (2008b). Low complexity BICM demodulation for MIMO transmission, *9th IEEE Workshop on Signal Processing Advances for Wireless Communications, SPAWC 2008*, Recife.
- Ghaffar, R. & Knopp, R. (2009a). Interference Suppression for Next Generation Wireless Systems, *IEEE 69-th Vehicular Technology Conference VTC-Spring 2009*, Barcelona.
- Ghaffar, R. & Knopp, R. (2009b). Spatial Interference Cancellation Algorithm, *Proc. IEEE Wireless Communications and Networking Conference WCNC 2009, Budapest*, Budapest, Hungary.
- Ghaffar, R. & Knopp, R. (2009c). Spatial Interference Cancellation and Pairwise Error Probability Analysis, *IEEE International Conference on Communications, ICC 2009*.
- Golden, G., Foschini, C., Valenzuela, R. & Wolniansky, P. (1999). Detection algorithm and initial laboratory results using v-blast space-time communication architecture, *Electronics Letters* **Vol. 35**(No. 1): 14–16.
- Gonzalez-Lopez, M., Vazquez-Araujo, F., Castedo, L. & Garcia-Frias, J. (2006). Optimized serially-concatenated LDGM and Alamouti codes for approaching MIMO Capacity, *IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2006*, pp. 1–5.
- Han, T. & Kobayashi, K. (1981). A new achievable rate region for the interference channel, *IEEE Transactions on Information Theory* **Vol. 27**(No. 1): 49–60.
- Hochwald, B. & Brink, S. T. (2003). Achieving near-capacity on a multiple-antenna channel, *IEEE Transactions on Communications*, **Vol. 51**(No. 3): 389–399.

- Larsson, E. G. & Stoica, P. (2003). *Space-Time Block Coding for Wireless Communications*, Cambridge University Press, Cambridge, U.K.
- Larsson, E. & Jalden, J. (2008). Fixed-complexity soft MIMO detection via partial marginalization, *IEEE Transactions on Signal Processing*, **Vol. 56**(No. 8): 3397–3407.
- Liu, Y., Lau, K., Takeshita, O. & Fitz, M. (2002). Optimal rate allocation for superposition coding in quasi-static fading channels, *IEEE International Symposium on Information Theory, ISIT 2002*, pp. 111–.
- LTE (2006). *Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)*, 3GPP TR 25.913 v7.3.0,.
- LTE-A (2008). *Requirements for Further Advancements for EUTRA*, 3GPP TR 36.913,.
- Medvedev, I., Bjerke, B., Walton, R., Ketchum, J., Wallace, M. & Howard, S. (2006). A comparison of MIMO receiver structures for 802.11N WLAN - Performance and complexity, *IEEE 17th International Symposium on Personal Indoor and Mobile Radio Communications*, pp. 1–5.
- Poor, H. & Verdu, S. (1997). Probability of error in MMSE multiuser detection, *IEEE Transactions on Information Theory* **Vol. 43**(No. 3): 858–871.
- Russell, M. & Stuber, G. (1995). Interchannel interference analysis of OFDM in a mobile environment, *IEEE 45th Vehicular Technology Conference*, Vol. 2, pp. 820–824.
- Shamai, S. & Steinier, A. (2003). A broadcast approach for a single-user slowly fading MIMO channel, *IEEE Transactions on Information Theory* **Vol. 49**(No. 10): 2617–2635.
- Telatar, I. E. (1999). Capacity of multi-antenna Gaussian channels, *European Transactions on Telecommunications* **Vol. 10**(No. 6): 585–595.
- Varanasi, M. K. & Guess, T. (1997). Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian MAC channel, *Asilomar Conference on Signals, Systems and Computers*.
- Verdu, S. (1989). Computational complexity of multiuser detection, *Algorithmica*, Vol. 4, pp. 303–312.
- Verdu, S. (1998). *Multiuser Detection*, Cambridge University Press, Cambridge, U.K.
- Winters, J. (1984). Optimum combining in digital mobile radio with cochannel interference, *IEEE Journal on Selected Areas in Communications* **Vol. 2**(No. 4): 528–539.
- Wolniansky, P., Foschini, G., Golden, G. & Valenzuela, R. (1998). V-blast: an architecture for realizing very high data rates over the rich-scattering wireless channel, *International Symposium on Signals, Systems, and Electronics, 1998. ISSSE 98*, pp. 295–300.

Advanced Hybrid–ARQ Receivers for Broadband MIMO Communications

Tarik Ait-Idir^{1,2}, Houda Chafnaji^{1,2}, Samir Saoudi²
and Athanasios Vasilakos³

¹ *Communications Systems Department, INPT, Madinat Al-Irfane, Rabat,
Morocco*

² *Signal and Communications Department, Institut Telecom/Telecom Bretagne, Brest, and
Université Européenne de Bretagne (UEB),
France*

³ *University of Western Macedonia,
Greece*

1. Introduction

Multiple input multiple output (MIMO) and hybrid–automatic repeat request (ARQ) mechanisms play a key role in the evolution of current wireless communications systems towards high data rate wireless packet access. In MIMO techniques, the spatial dimension of the MIMO channel is exploited through the use of multiple antennas at both the transmitter and receiver sides. This translates into an improvement in the spectrum efficiency and/or the link quality Wolniansky et al. (1998). Hybrid–ARQ protocols provide an important source of time diversity through the combination of channel coding and ARQ. This is performed with the aid of *packet combining* techniques where erroneous data packets are kept in the receiver to help detect/decode retransmitted frames.

In broadband MIMO communications, the MIMO wireless link suffers from intersymbol interference (ISI) caused by multipath propagation. This effect can be mitigated using channel equalization and/or Hybrid–ARQ. In this chapter, we focus on the joint design of the packet combiner and the channel equalizer for MIMO ARQ transmission over the broadband wireless channel. We start the chapter by reviewing the various approaches for joint packet combining and equalization. We introduce the considered broadband MIMO ARQ transmission scheme. Then, we derive the structure of the optimal maximum *a posteriori* (MAP) turbo packet combiner and study its outage performance. Finally, we introduce a new class of low-complexity minimum mean square error (MMSE)-based turbo packet combiners and analyze their implementation requirements and block error rate (BLER) performance.

2. Advanced Receivers for MIMO ARQ

In the last few years, a special interest has been paid to the design of advanced MIMO ARQ receivers where packet combining and signal processing, i.e., detection/equalization, are jointly performed. The concept of integrated equalization (IEQ) has been proposed in the framework

of MIMO systems with flat fading for joint multiple antenna interference (MAI) suppression and packet combining (see for instance Onggosanusi et al. (2003) and Samra & Ding (2006)). Turbo coded ARQ schemes with iterative MMSE frequency domain equalization (FDE) for MIMO code division multiple access (CDMA) has been proposed in Garg & Adachi (2006). Recently, we have introduced a new family of packet combining techniques for broadband MIMO ARQ systems, where the decoding of a data packet is performed with the aid of an iterative (turbo) processing between the soft combiner, i.e., joint packet combining and equalization unit, and the soft input soft output (SISO) decoder (see Ait-Idir & Saoudi (2009)). The following sections of this chapter focus on this new class of broadband MIMO ARQ techniques. Our notation is introduced in the next section followed by the communication model of the considered MIMO ARQ system.

3. Notation

In this chapter, scalars are denoted by small-case letters, vectors by small-case boldface letters, and matrices by upper-case boldface letters. Superscripts T and H denote the transpose and transpose conjugate, respectively. \mathbf{I}_N is the $N \times N$ identity matrix, and $\mathbf{0}_{N \times M}$ is the $N \times M$ zeros matrix. \otimes is the Kronecker product, and $\Pr\{\cdot\}$ denotes the probability of a given event.

4. Broadband MIMO ARQ Transceiver Scheme

4.1 MIMO ARQ Transmission

Let us consider a broadband multi-antenna, i.e., multiple input multiple output (MIMO), system operating over a frequency selective fading channel and using an ARQ protocol at the upper layer. The transmitter and the receiver are equipped with N_T transmit (index $t = 1, \dots, N_T$) and N_R receive (index $r = 1, \dots, N_R$) antennas, respectively. The MIMO channel suffers from intersymbol interference (ISI) and is composed of L symbol-spaced taps (index $l = 0, \dots, L - 1$). Each data stream is first encoded with the aid of a ρ -rate channel encoder, interleaved using a semi-random interleaver Π , then modulated and space-time multiplexed over the N_T transmit antennas. This transmission scheme corresponds to the so-called space-time bit interleaved coded modulation (STBICM). Let \mathcal{S} denote the constellation set, and $M = \log_2 \{|\mathcal{S}|\}$ its cardinality. A sequence of M coded and interleaved bits $b_{1,t,i}, \dots, b_{M,t,i}$ available on antenna t is transmitted at discrete-time instant $i = 0, \dots, T - 1$ over symbol $s_{t,i} \in \mathcal{S}$ according to the mapping function $\Psi : \{0, 1\}^M \rightarrow \mathcal{S}$. The symbol vector to be transmitted at time i is denoted

$$\mathbf{s}_i \triangleq [s_{1,i}, \dots, s_{N_T,i}]^T \in \mathcal{S}^{N_T}. \quad (1)$$

The rate of this transmission scheme is therefore $R = \rho M N_T$. The transmit symbol energy is normalized to one. Assuming infinitely deep space-time interleaving we get, $\mathbb{E} [\mathbf{s}_i \mathbf{s}_i^H] = \mathbf{I}_{N_T}$. At the receiver side, a positive/negative acknowledgment ACK/NACK message is sent back to the transmitter upon the decoding of the information block. When the transmitter receives a NACK message due to an erroneously decoded packet, subsequent transmission rounds occur until the packet is correctly received or a preset maximum number of rounds K (index $k = 1, \dots, K$) is reached. Parameter K is called the *ARQ delay*. Reception of a ACK message indicates a successful decoding and the transmitter moves on to the next packet. We assume an error-free feedback channel (the signaling channel carrying the ACK/NACK message), and perfect packet error detection (using a cyclic redundancy check –CRC code). We focus on Chase-type ARQ, i.e., the entire information block is retransmitted using the same STBICM

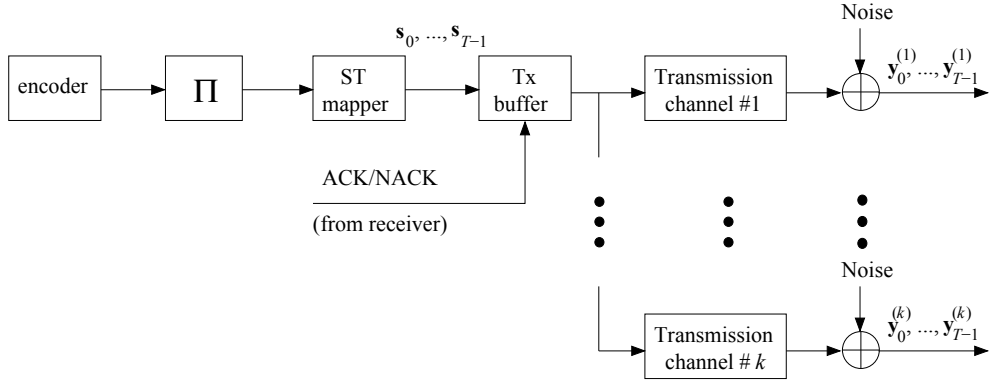


Fig. 1. STBICM ARQ Transmission over a short-term quasi-static MIMO channel

code. To prevent inter-block interference (IBI), we use either zero padding (ZP) or cyclic prefix (CP)-aided transmission.

The MIMO-ISI channel is assumed to be short-term quasi-static block fading, i.e., constant during one ARQ round and independently changes from round to round. The long-term dynamic corresponds to the case when the channel is constant during all ARQ rounds corresponding to the transmission of the same information packet (see El Gamal et al. (2006)). The short-term assumption is justified by the fact ARQ protocols are mainly used to improve the link quality in the case of delay-tolerant applications where the processing delay is not a major constraint. Let $\mathbf{H}_l^{(k)}$ denote the $N_R \times N_T$ complex matrix of the l th tap connecting the transmitter and the receiver at ARQ round k . The elements of $\mathbf{H}_l^{(k)}$ are zero-mean circularly symmetric Gaussian random variables, i.e., $h_{r,t,l}^{(k)} \sim \mathcal{CN}(0, \sigma_l^2)$, where $h_{r,t,l}^{(k)}$ denotes the (r, t) th element of matrix $\mathbf{H}_l^{(k)}$ and σ_l^2 is the energy of tap l . The total channel energy is normalized as $\sum_{l=0}^{L-1} \sigma_l^2 = 1$. The discrete baseband signal received by antenna r at ARQ round k and time instant i is given by

$$y_{r,i}^{(k)} = \sum_{l=0}^{L-1} \sum_{t=1}^{N_T} h_{r,t,l}^{(k)} s_{t,i-l} + n_{r,i}^{(k)}, \quad (2)$$

where $\mathbf{n}_i^{(k)} \triangleq [n_{1,i}^{(k)}, \dots, n_{N_R,i}^{(k)}]^\top \sim \mathcal{CN}(\mathbf{0}_{N_R \times 1}, \sigma^2 \mathbf{I}_{N_R})$ is the thermal noise at the N_R receive antennas. The STBICM ARQ transmission scheme over a short-term MIMO channel, i.e., k distinct MIMO channels, is depicted in Fig. 1.

4.2 MIMO ARQ Turbo Receiver

In Ait-Idir & Saoudi (2009), *turbo packet combining* has been introduced as an efficient technique for combining multiple transmissions in the case of broadband MIMO ARQ systems. In turbo packet combining, the decoding of a data packet is performed in an iterative (turbo) fashion through the exchange of soft (extrinsic) information between the SISO packet combiner and the SISO decoder. The soft combiner exploits signals received at multiple ARQ rounds to compute extrinsic log-likelihood ratios (LLR)s. Note that in conventional LLR-level

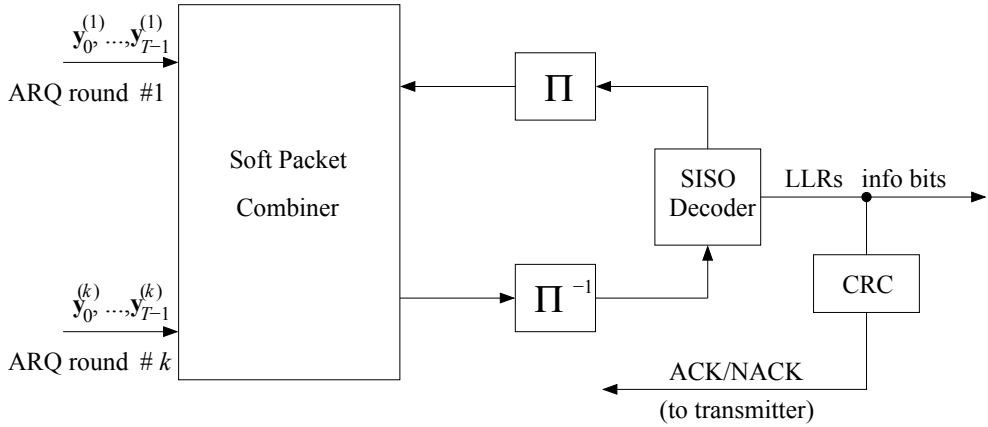


Fig. 2. Block diagram of the turbo packet combining-aided MIMO ARQ receiver

combining techniques, the soft outputs obtained at different transmissions are simply added together before channel decoding.

The general block diagram of the turbo packet combining-aided MIMO ARQ receiver is depicted in Fig. 2. Let

$$\boldsymbol{\phi}_{t,i}^a \triangleq [\phi_{1,t,i}^a, \dots, \phi_{M,t,i}^a]^\top \quad (3)$$

denote the $M \times 1$ real vector of *a priori* LLRs corresponding to coded and interleaved bits $b_{1,t,i}, \dots, b_{M,t,i}$, and available at the input of the soft combiner at a certain iteration of ARQ round k . Using *a priori* information $\boldsymbol{\phi}_{1,0}^a, \dots, \boldsymbol{\phi}_{N_T,T-1}^a$ and signals received at rounds $1, \dots, k$, the soft packet combiner computes extrinsic LLR vectors

$$\boldsymbol{\phi}_{t,i}^e \triangleq [\phi_{1,t,i}^e, \dots, \phi_{M,t,i}^e]^\top, \quad (4)$$

which are de-interleaved and sent to the SISO decoder to obtain *a posteriori* LLRs about useful bits and extrinsic information about coded bits. The generated extrinsic LLR values are then interleaved and fed back to the soft combiner to help compute soft information during the next turbo iteration of the same ARQ round. After a preset number of iterations, the decision about the data packet is performed, and the ACK/NACK message is sent back to the transmitter accordingly. Note that during the first iteration, *a priori* information corresponds to the soft information available from the last iteration of previous ARQ round $k - 1$.

5. Information-Theoretic Issues

In this section, we derive the optimal maximum *a posteriori* (MAP) turbo packet combining receiver for broadband MIMO ARQ transmission, and investigate its outage performance. We first show that optimal turbo packet combining can be formulated as a MIMO-ISI turbo equalization problem. We then obtain the outage probability of the broadband MIMO ARQ channel, and analyze the impact of the ARQ delay, i.e., maximum number of ARQ rounds K , on the outage performance.

5.1 Optimal Turbo Packet Combining

To derive the optimal maximum *a posteriori* (MAP) turbo packet combiner at ARQ round k , let us consider the following signal vector that groups signals corresponding to all ARQ rounds $1, \dots, k$,

$$\mathbf{y}^{(k)} \triangleq \left[\mathbf{y}_{T-1}^{(1)\top}, \dots, \mathbf{y}_{T-1}^{(k)\top}, \dots, \mathbf{y}_0^{(1)\top}, \dots, \mathbf{y}_0^{(k)\top} \right]^\top \in \mathbb{C}^{kN_R T}, \quad (5)$$

where

$$\mathbf{y}_i^{(u)} \triangleq \left[y_{1,i}^{(u)}, \dots, y_{N_R,i}^{(u)} \right]^\top \in \mathbb{C}^{N_R} \quad (6)$$

is the vector of received signals at ARQ round $u = 1, \dots, k$ and time instant i .

The formulation in (5) is of a great importance because it allows us to view each ARQ round as a set of virtual N_R receive antennas. Note that in this section we assume that all signals and channel matrices corresponding to previous ARQ rounds are available at the receiver side at round k . In Section 6, we will present an optimized turbo packet combining technique that makes use of all signals and channel matrices without being required to be explicitly stored in the receiver. With respect to (2) and (5) and assuming a ZP-aided transmission strategy, the signal vector $\mathbf{y}^{(k)}$ corresponding to the transmission of the entire symbol frame over k MIMO-ISI channels can be expressed as,

$$\mathbf{y}^{(k)} = \mathbf{H}^{(k)} \mathbf{s} + \mathbf{n}^{(k)}, \quad (7)$$

where $\mathbf{H}^{(k)}$ is a block Toeplitz matrix given as

$$\mathbf{H}^{(k)} \triangleq \begin{bmatrix} \boxed{\begin{matrix} \mathbf{H}_0^{(1)} \\ \vdots \\ \mathbf{H}_0^{(k)} \end{matrix}} & \cdots & \boxed{\begin{matrix} \mathbf{H}_{L-1}^{(1)} \\ \vdots \\ \mathbf{H}_{L-1}^{(k)} \end{matrix}} & & \\ & \ddots & & \ddots & \\ & & \boxed{\begin{matrix} \mathbf{H}_0^{(1)} \\ \vdots \\ \mathbf{H}_0^{(k)} \end{matrix}} & \cdots & \boxed{\begin{matrix} \mathbf{H}_{L-1}^{(1)} \\ \vdots \\ \mathbf{H}_{L-1}^{(k)} \end{matrix}} \end{bmatrix}_{kN_R T \times N_T T}, \quad (8)$$

and vectors \mathbf{s} and $\mathbf{n}^{(k)}$ are defined as,

$$\mathbf{s} \triangleq \left[\mathbf{s}_{T-1}^\top, \dots, \mathbf{s}_0^\top \right]^\top \in \mathcal{S}^{N_T T}, \quad (9)$$

$$\mathbf{n}^{(k)} \triangleq \left[\mathbf{n}_{T-1}^{(1)\top}, \dots, \mathbf{n}_{T-1}^{(k)\top}, \dots, \mathbf{n}_0^{(1)\top}, \dots, \mathbf{n}_0^{(k)\top} \right]^\top \in \mathbb{C}^{kN_R T}. \quad (10)$$

The communication model (7) corresponds to a MIMO-ISI equalization problem with N_T transmit and kN_R virtual receive antennas. It allows for jointly (over all ARQ rounds) canceling both multiple antenna interference (MAI) and ISI, while exploiting all the diversities available in the MIMO-ISI ARQ channel. Using the MAP criterion and given *a priori* LLRs, the extrinsic information about coded and interleaved bit $b_{m,t,i}$ can be expressed as,

$$\phi_{m,t,i}^e = \log \frac{\Pr \left\{ \mathbf{y}^{(k)} \mid b_{m,t,i} = 1; \mathbf{H}_0^{(1)}, \dots, \mathbf{H}_{L-1}^{(k)}, \text{ a priori LLRs} \right\}}{\Pr \left\{ \mathbf{y}^{(k)} \mid b_{m,t,i} = 0; \mathbf{H}_0^{(1)}, \dots, \mathbf{H}_{L-1}^{(k)}, \text{ a priori LLRs} \right\}}. \quad (11)$$

Then, by invoking the multi-round communication model (7), $\phi_{m,t,i}^e$ can be obtained as,

$$\phi_{m,t,i}^e = \log \frac{\sum_{\mathbf{s} \in \mathcal{S}_{m,t,i}^1} \exp \left\{ -\frac{1}{2\sigma^2} \left\| \mathbf{y}^{(k)} - \mathbf{H}^{(k)} \mathbf{s} \right\|^2 + \sum_{(m',t',i') \neq (m,t,i)} \Psi_{m'}^{-1}(s_{t',i'}) \phi_{m',t',i'}^a \right\}}{\sum_{\mathbf{s} \in \mathcal{S}_{m,t,i}^0} \exp \left\{ -\frac{1}{2\sigma^2} \left\| \mathbf{y}^{(k)} - \mathbf{H}^{(k)} \mathbf{s} \right\|^2 + \sum_{(m',t',i') \neq (m,t,i)} \Psi_{m'}^{-1}(s_{t',i'}) \phi_{m',t',i'}^a \right\}}, \quad (12)$$

where the subset $\mathcal{S}_{m,t,i}^b$ is defined as $\mathcal{S}_{m,t,i}^b \triangleq \{ \mathbf{s} \in \mathcal{S}^{N_T T} \mid \Psi_m^{-1}(s_{t,i}) = b \}, b = 0, 1$.

5.2 Outage Performance Analysis

Outage probability is a useful tool that allows to analyze the performance of non-ergodic channels, i.e., quasi-static channels. It provides a lower bound on the BLER, and is generally defined as the probability that the mutual information, as a function of the channel realization and the average signal to noise ratio (SNR) γ per receive antenna, is below the transmission rate R Tse & Viswanath (2005).

The outage probability of an ARQ protocol can be derived using the *renewal theory* (see Wolff (1989)) as in Caire & Tuninetti (2001) and El Gamal et al. (2006). In the case of a broadband MIMO system with Chase-type ARQ, perfect packet error detection, and error-free ACK/NACK feedback, the outage probability can be derived using the multi-round communication model (7) and the renewal theory (see Ait-Idir & Saoudi (2009)) as

$$P_{out}^R(\gamma) = \Pr \left\{ \frac{1}{K} I(\mathbf{s}; \mathbf{y}^{(K)} \mid \mathbf{H}^{(K)}, \gamma) < R, \bar{\mathcal{A}}_1, \dots, \bar{\mathcal{A}}_{K-1} \right\}, \quad (13)$$

where K is the ARQ delay, and $\bar{\mathcal{A}}_k$ denotes the event that an ACK message is fed back at round k , i.e., the data packet is positively acknowledged at ARQ round k . The quantity $\frac{1}{K} I(\mathbf{s}; \mathbf{y}^{(K)} \mid \mathbf{H}^{(K)}, \gamma)$ denotes the mutual information rate at the last ARQ round K , and is expressed in the case of a MIMO broadband channel with Gaussian inputs as

$$\frac{1}{K} I(\mathbf{s}; \mathbf{y}^{(K)} \mid \mathbf{H}^{(K)}, \gamma) = \frac{1}{KT} \sum_{i=0}^{T-1} \log_2 \left(\det \left(\mathbf{I}_{KN_R} + \frac{\gamma}{N_T} \mathbf{\Lambda}_i^{(K)} \mathbf{\Lambda}_i^{(K)H} \right) \right), \quad (14)$$

where $\mathbf{\Lambda}_i^{(K)}$ is the discrete Fourier transform (DFT) of the $KN_R \times N_T$ virtual MIMO-ISI channel at round K and frequency bin i , i.e.,

$$\mathbf{\Lambda}_i^{(K)} \triangleq \sum_{l=0}^{L-1} \begin{bmatrix} \mathbf{H}_l^{(1)} \\ \vdots \\ \mathbf{H}_l^{(K)} \end{bmatrix} \exp \left\{ -j \frac{2\pi}{T} il \right\}. \quad (15)$$

Note that the factor $\frac{1}{K}$ appearing in the outage probability (13) is due to the fact that a transmission scheme with Chase-type ARQ, and an ARQ delay K is equivalent to a repetition coding scheme where K parallel sub-channels are used to transmit the same symbol frame.

In the following, we investigate the impact of the ARQ delay K on the outage performance. We consider a MIMO-ISI channel with $L = 2$ taps having equal powers, i.e., $\sigma_0^2 = \sigma_1^2 = \frac{1}{2}$. At each ARQ round k , the mutual information rate $\frac{1}{k} I(\mathbf{s}; \mathbf{y}^{(k)} | \mathbf{H}_0^{(k)}, \mathbf{H}_1^{(k)}, \gamma)$ of the MIMO ARQ system after k rounds is evaluated similarly to (14). If the target rate R is not reached and $k < K$, the system moves on to the next round $k + 1$. An ARQ process is stopped and another is started, either because of system outage, i.e., the mutual rate after K rounds is below R , or the rate R is achieved at a certain round $k \leq K$. In all scenarios, we take $T = 256$ discrete time instants.

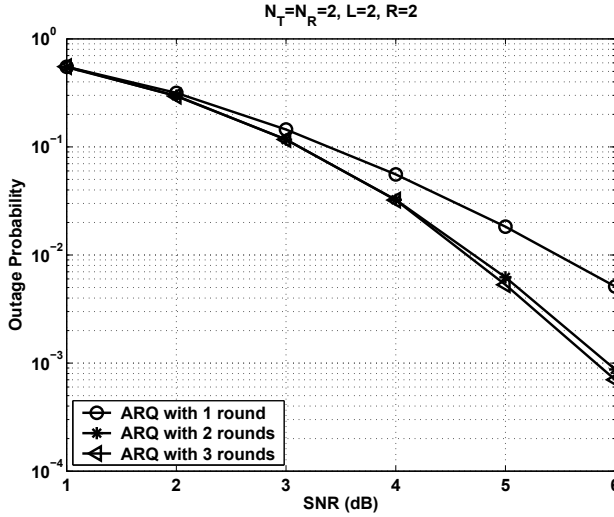


Fig. 3. Outage probability performance for two transmit and two receive antennas

In Fig. 3, we plot the outage probability performance for the $N_T = N_R = 2$ antenna configuration, with a target rate $R = 2$. We observe that the ARQ diversity gain, due to the short-term static channel dynamic, clearly appears when the ARQ delay is set to $K = 2$. It offers a significant SNR gain compared with the non-ARQ case ($K = 1$). When, the ARQ delay is increased to $K = 3$, the outage performance is similar to that of $K = 2$. This means that if the system is in outage in the second ARQ round, then it will almost be in outage in the third round. In Fig. 4, we investigate the outage performance when the number of transmit antennas is increased to $N_T = 4$, while $N_R = 2$. The target rate is set to $R = 4$. As in the previous configuration, both ARQ delays $K = 2$ and $K = 3$ provide almost the same outage performance, while the overall diversity gain is more important than that corresponding to $N_T = N_R = 2$. This can be seen from the steeper slopes of outage curves. Note that the ARQ diversity due to multiple transmissions does not completely translate into a receive diversity gain (related to the N_R virtual receive antennas at each ARQ round). This is due to the fact that the target rate R has to be maintained, as it can be seen from the expression of the outage probability (13). This means

that the diversity gain does not linearly increase with increase of the ARQ delay K . This issue has been addressed by El Gamal et al. (2006) in the case of flat fading MIMO ARQ systems.

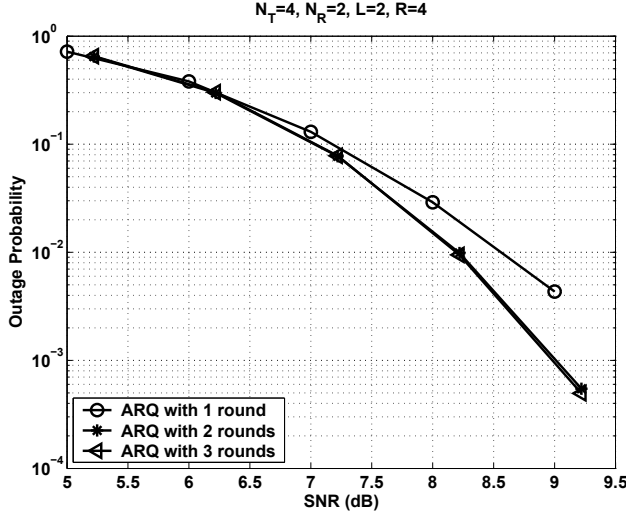


Fig. 4. Outage probability performance for four transmit and two receive antennas

6. MMSE-Based Turbo Packet Combining

The optimal MAP turbo packet combining algorithm we have presented so far in Subsection 5.1 has a computational complexity that exponentially increases with the increase in the number of ARQ rounds. In addition, all signals and channel matrices have to be stored in the receiver to perform combining at each ARQ round. In this section, we present alternative MMSE-based sub-optimal techniques (see Ait-Idir & Saoudi (2009)) that allow for performing turbo packet combining with reduced computational complexity and memory requirements. We first introduce the so-called *signal-level turbo combining* technique where packet combining is performed at the signal level similarly to MAP combining but based on MMSE processing. Then, we present the *symbol-level turbo combining* algorithm where packets are combined at the input of the soft demapper after performing MMSE channel equalization separately for each transmission. We also provide a brief presentation of the conventional LLR-level combining technique. For all combining schemes, we assume a ZP-aided block transmission. The multi-round block communication model is therefore described by a block Toeplitz channel matrix as in (7).

6.1 Signal-Level Turbo Combining

The signal-level turbo packet combining technique is a low-complexity MMSE-based combining algorithm that performs packet combining jointly with MAI and ISI cancellation using a block length equal to $\varepsilon = \varepsilon_1 + \varepsilon_2 + 1 \ll T$, where ε_1 and ε_2 are the lengths of the forward and backward filters, respectively. Packet combining at ARQ round k is therefore performed with

respect to the following ε -length block communication model

$$\underline{\mathbf{y}}_i^{(k)} = \underline{\mathbf{H}}^{(k)} \underline{\mathbf{s}}_i + \underline{\mathbf{n}}_i^{(k)}, \quad (16)$$

where

$$\underline{\mathbf{y}}_i^{(k)} \triangleq \left[\mathbf{y}_{i+\varepsilon_1}^{(1)\top}, \dots, \mathbf{y}_{i+\varepsilon_1}^{(k)\top}, \dots, \mathbf{y}_{i-\varepsilon_2}^{(1)\top}, \dots, \mathbf{y}_{i-\varepsilon_2}^{(k)\top} \right]^\top \in \mathbb{C}^{kN_R\varepsilon}, \quad (17)$$

$$\underline{\mathbf{n}}_i^{(k)} \triangleq \left[\mathbf{n}_{i+\varepsilon_1}^{(1)\top}, \dots, \mathbf{n}_{i+\varepsilon_1}^{(k)\top}, \dots, \mathbf{n}_{i-\varepsilon_2}^{(1)\top}, \dots, \mathbf{n}_{i-\varepsilon_2}^{(k)\top} \right]^\top \in \mathbb{C}^{kN_R\varepsilon}, \quad (18)$$

$$\underline{\mathbf{s}}_i \triangleq \left[\mathbf{s}_{i+\varepsilon_1}^\top, \dots, \mathbf{s}_{i-\varepsilon_2-L+1}^\top \right]^\top \in \mathcal{S}^{N_T(\varepsilon+L-1)}, \quad (19)$$

and $\underline{\mathbf{H}}^{(k)}$ is a block Toeplitz matrix which is defined similarly to (8) but using ε block rows and $(\varepsilon + L - 1)$ block columns.

First of all, the soft packet combiner computes conditional means and variances of transmitted symbols using *a priori* information vectors $\boldsymbol{\phi}_{t,i}^a$, where $t \in \{1, \dots, N_T\}$, and $i \in \{0, \dots, T - 1\}$, available from the previous iteration. Note that, as it has been mentioned before, the soft information generated by the SISO decoder during the last iteration of ARQ round $k - 1$ serves as *a priori* information at the first iteration of round k . Conditional symbol expectations, also called *soft symbols*, serve for regenerating soft interference caused by multiple antenna transmission and multipath propagation, i.e., MAI plus ISI. Conditional symbol variances and channel matrices corresponding to ARQ rounds $1, \dots, k$ are used for computing the multi-round MMSE filter that serves for performing signal combining. At each iteration of ARQ round k , the signal-level turbo packet combiner exploits the communication model (16) to cancel soft MAI plus ISI (estimated at the current iteration of round k) from all signals received at rounds $1, \dots, k$. Then, it combines the resulting soft interference-free signals to generate soft decisions about transmitted symbols.

Let $\tilde{s}_{t,i} \triangleq \mathbb{E} [s_{t,i} | \boldsymbol{\phi}_{t,i}^a]$ and $\tilde{\sigma}_{t,i}^2 \triangleq \mathbb{E} [|s_{t,i} - \tilde{s}_{t,i}|^2 | \boldsymbol{\phi}_{t,i}^a]$ denote the conditional mean and variance of symbol $s_{t,i}$. By invoking the special block Toeplitz structure of matrix $\underline{\mathbf{H}}^{(k)}$, and the structure of the multi-round signal vector (17), the signal-level turbo packet combiner computes, at a particular iteration of ARQ round k , soft decision $\tilde{\zeta}_{t,i}^{(k)}$ about symbol $s_{t,i}$ using the following forward-backward filtering structure,

$$\tilde{\zeta}_{t,i}^{(k)} = \mathbf{F}_t^{(k)} \underline{\mathbf{z}}_i^{(k)} | \mathbf{B}_t^{(k)} \tilde{\mathbf{s}}_{i|t}, \quad (20)$$

where $\tilde{\mathbf{s}}_{i|t}$ is the $N_T(\varepsilon + L - 1)$ -length soft symbol vector corresponding to (19) with zero at the $(\varepsilon_1 N_T + t)$ th position corresponding to soft symbol $\tilde{s}_{t,i}$. $\mathbf{F}_t^{(k)}$ and $\mathbf{B}_t^{(k)}$ are the forward and backward filters related to antenna t . $\underline{\mathbf{z}}_i^{(k)}$ is a vector that contains properly weighed and combined copies of signals received at rounds $1, \dots, k$. It does not change in the course of turbo iterations corresponding to the same ARQ round, and is produced at round k according to the following recursion

$$\begin{cases} \underline{\mathbf{z}}_i^{(k)} &= \underline{\mathbf{z}}_i^{(k-1)} + \underline{\mathbf{H}}^{(k)H} \underline{\mathbf{y}}_i^{(k)}, \\ \underline{\mathbf{z}}_i^{(0)} &= \mathbf{0}_{N_T(\varepsilon+L-1) \times 1}. \end{cases} \quad (21)$$

The signal vector $\mathbf{y}_i^{(k)}$ and the channel matrix $\mathbf{H}^{(k)}$ correspond to the ε -length block communication model of ARQ round k , and are given as,

$$\mathbf{H}^{(k)} \triangleq \begin{bmatrix} \mathbf{H}_0^{(k)} & \cdots & \mathbf{H}_{L-1}^{(k)} \\ & \ddots & \\ & & \mathbf{H}_0^{(k)} & \cdots & \mathbf{H}_{L-1}^{(k)} \end{bmatrix}_{N_R \varepsilon \times N_T(\varepsilon+L-1)}, \quad (22)$$

$$\mathbf{y}_i^{(k)} \triangleq \left[\mathbf{y}_{i+\varepsilon_1}^{(k)\top}, \dots, \mathbf{y}_{i-\varepsilon_2}^{(k)\top} \right]^\top \in \mathbb{C}^{N_R \varepsilon}. \quad (23)$$

Filters $\mathbf{F}_t^{(k)}$ and $\mathbf{B}_t^{(k)}$ are iteration-dependent, and are computed as

$$\mathbf{F}_t^{(k)} = \left(\sigma^2 + \left(1 - \tilde{\sigma}_t^2 \right) \mathbf{e}_t^\top \mathbf{\Lambda}^{(k)} \mathbf{Y}^{(k)} \mathbf{e}_t \right)^{-1} \mathbf{e}_t^\top \mathbf{\Lambda}^{(k)}, \quad (24)$$

$$\mathbf{B}_t^{(k)} = \mathbf{F}_t^{(k)} \mathbf{Y}^{(k)}. \quad (25)$$

where \mathbf{e}_t is a $N_T(\varepsilon + L - 1)$ -length vector of zeros with one at the $(\varepsilon_1 N_T + t)$ th position, i.e.,

$$\mathbf{e}_t \triangleq \left[\underbrace{0, \dots, 0}_{\varepsilon_1 N_T + t - 1}, 1, \underbrace{0, \dots, 0}_{(\varepsilon_2 + L) N_T - t} \right]^\top. \quad (26)$$

$\tilde{\sigma}_t^2$ is the unconditional variance of symbols transmitted over antenna t , and is computed as the time average of conditional variances $\tilde{\sigma}_{t,0}^2, \dots, \tilde{\sigma}_{t,T-1}^2$, i.e., $\tilde{\sigma}_t^2 = \frac{1}{T} \sum_{i=0}^{T-1} \tilde{\sigma}_{t,i}^2$. The square matrix $\mathbf{\Lambda}^{(k)}$ is updated at each turbo iteration according to,

$$\mathbf{\Lambda}^{(k)} = \mathbf{I}_{N_T(\varepsilon+L-1)} - \mathbf{Y}^{(k)} \left(\mathbf{Y}^{(k)} + \sigma^2 \mathbf{\Xi}^{-1} \right)^{-1}, \quad (27)$$

where

$$\mathbf{\Xi} = \mathbf{I}_{\varepsilon+L-1} \otimes \begin{bmatrix} \tilde{\sigma}_1^2 & & \\ & \ddots & \\ & & \tilde{\sigma}_{N_T}^2 \end{bmatrix} \in \mathbb{C}^{N_T(\varepsilon+L-1) \times N_T(\varepsilon+L-1)}, \quad (28)$$

and $\mathbf{Y}^{(k)}$ is recursively computed as,

$$\begin{cases} \mathbf{Y}^{(k)} &= \mathbf{Y}^{(k-1)} + \mathbf{H}^{(k)H} \mathbf{H}^{(k)}, \\ \mathbf{Y}^{(0)} &= \mathbf{0}_{N_T(\varepsilon+L-1) \times N_T(\varepsilon+L-1)}. \end{cases} \quad (29)$$

Details regarding the derivation of the forward-backward filtering structure (20) can be found in Ait-Idir & Saoudi (2009)

Extrinsic information $\phi_{m,t,i}^e$ about coded and interleaved bit $b_{m,t,i}$, $m = 1, \dots, M$, can be obtained using the decision statistic $\zeta_{t,i}^{(k)}$ of (20) as,

$$\phi_{m,t,i}^e = \log \frac{\sum_{s \in \mathcal{S}_m^1} \exp \left\{ -\frac{1}{2\delta_t^{(k)^2}} \left| \zeta_{t,i}^{(k)} - a_t^{(k)} s \right|^2 + \sum_{m' \neq m} \Psi_{m'}^{-1}(s) \phi_{m',t,i}^a \right\}}{\sum_{s \in \mathcal{S}_m^0} \exp \left\{ -\frac{1}{2\delta_t^{(k)^2}} \left| \zeta_{t,i}^{(k)} - a_t^{(k)} s \right|^2 + \sum_{m' \neq m} \Psi_{m'}^{-1}(s) \phi_{m',t,i}^a \right\}}, \quad (30)$$

where it is assumed that conditional soft demapper input $\tilde{\zeta}_{t,i}^{(k)} \mid s_{t,i}$ is Gaussian with mean $\alpha_t^{(k)}$ and variance $\delta_t^{(k)^2}$, and are given by,

$$\begin{cases} \alpha_t^{(k)} &= \mathbf{B}_t^{(k)} \mathbf{e}_t \\ \delta_t^{(k)^2} &= (1 - \alpha_t^{(k)}) \alpha_t^{(k)}, \end{cases} \quad (31)$$

and the set \mathcal{S}_m^b is defined as $\mathcal{S}_m^b \triangleq \{s \in \mathcal{S} \mid \Psi_m^{-1}(s) = b\}$ for $b = 0, 1$.

6.2 Symbol-Level Turbo Combining

In symbol-level turbo packet combining, MMSE turbo equalization is separately performed for each ARQ round k based on the ε -length block communication model of round k ,

$$\mathbf{y}_i^{(k)} = \mathbf{H}^{(k)} \mathbf{s}_i + \mathbf{n}_i^{(k)}, \quad (32)$$

where \mathbf{s}_i , $\mathbf{H}^{(k)}$, and $\mathbf{y}_i^{(k)}$ are given by (19), (22), and (33), respectively, and

$$\mathbf{n}_i^{(k)} \triangleq \left[\mathbf{n}_{i+\varepsilon_1}^{(k)\top}, \dots, \mathbf{n}_{i-\varepsilon_2}^{(k)\top} \right]^\top \in \mathbb{C}^{N_{R\varepsilon}} \quad (33)$$

is the spatially and temporally white Gaussian noise at the input of the equalizer at ARQ round k , i.e., $\mathbf{n}_i^{(k)} \sim \mathcal{CN}(\mathbf{0}_{N_{R\varepsilon} \times 1}, \sigma^2 \mathbf{I}_{N_{R\varepsilon}})$. Soft combining is iteratively performed at the level of unconditional MMSE filter outputs by combining the output at each iteration of ARQ round k with those obtained at the last iteration of previous rounds $1, \dots, k-1$.

The forward and backward filters $\mathbf{F}_t^{(k)}$ and $\mathbf{B}_t^{(k)}$ in the case of symbol-level turbo packet combining can easily be derived using the equations provided in the previous subsection with $k = 1$. Now, let $\tilde{\zeta}_{t,i}^{(k)}$ denote the MMSE filter output corresponding to symbol $s_{t,i}$ at a specific turbo iteration of round k . The conditional decision statistic $\tilde{\zeta}_{t,i}^{(k)} \mid s_{t,i}$ is Gaussian, with mean $\check{\alpha}_t^{(k)}$ and variance $\check{\delta}_t^{(k)^2}$ given similarly to (31). For ARQ rounds $u = 1, \dots, k-1$, $\tilde{\zeta}_{t,i}^{(u)}$ denotes the decision statistic obtained at the last iteration of round u . Therefore, the soft combiner provides extrinsic information $\phi_{m,t,i}^e$ about coded and interleaved bit $b_{m,t,i}$ as

$$\phi_{m,t,i}^e = \log \frac{\sum_{s \in \mathcal{S}_m^1} \exp \left\{ -\frac{1}{2} \left(\tilde{\zeta}_{t,i}^{(k)} - s \check{\alpha}_t^{(k)} \right)^H \Delta_t^{(k)-1} \left(\tilde{\zeta}_{t,i}^{(k)} - s \check{\alpha}_t^{(k)} \right) + \sum_{m' \neq m} \Psi_{m'}^{-1}(s) \phi_{m',t,i}^a \right\}}{\sum_{s \in \mathcal{S}_m^0} \exp \left\{ -\frac{1}{2} \left(\tilde{\zeta}_{t,i}^{(k)} - s \check{\alpha}_t^{(k)} \right)^H \Delta_t^{(k)-1} \left(\tilde{\zeta}_{t,i}^{(k)} - s \check{\alpha}_t^{(k)} \right) + \sum_{m' \neq m} \Psi_{m'}^{-1}(s) \phi_{m',t,i}^a \right\}}, \quad (34)$$

where vector

$$\check{\zeta}_{t,i}^{(k)} \triangleq \left[\check{\zeta}_{t,i}^{(1)}, \dots, \check{\zeta}_{t,i}^{(k)} \right]^\top \in \mathbb{C}^k \quad (35)$$

gathers MMSE filter soft outputs corresponding to all rounds $1, \dots, k$. $\check{\alpha}_t^{(k)}$ and $\Delta_t^{(k)}$ are the mean and covariance of the conditional Gaussian vector $\check{\zeta}_{t,i}^{(k)} \mid s_{t,i}$, and are given by,

$$\check{\alpha}_t^{(k)} \triangleq \left[\check{\alpha}_t^{(1)}, \dots, \check{\alpha}_t^{(k)} \right]^\top \in \mathbb{C}^k, \quad (36)$$

$$\Delta_t^{(k)} = \begin{bmatrix} \delta_t^{(1)^2} & & & \\ & \ddots & & \\ & & \delta_t^{(k)^2} & \\ & & & \ddots \end{bmatrix}_{k \times k} \quad (37)$$

6.3 LLR-Level Turbo Combining

In conventional LLR-level combining, LLR values of transmitted bits obtained at multiple ARQ rounds are stored in the receiver and simply added together to update the LLRs at each ARQ round. In our framework, LLR-level turbo combining is carried out by separately performing MMSE turbo equalization for multiple transmissions using the ε -length block communication model (32). Extrinsic LLRs $\phi_{m,t,i}^e$ corresponding to coded and interleaved bits $b_{m,t,i} \forall m, t, i$ are then computed at each iteration of ARQ round k using only decision statistics $\zeta_{t,i}^{(k)}$ introduced in the previous subsection. LLR values obtained at multiple ARQ rounds are then added together to produce the soft LLR outputs

$$\text{LLR}_{m,t,i}^{(k)} = \sum_{u=1}^k \phi_{m,t,i}^{e(u)}, \quad \forall m, t, i, \quad (38)$$

which are de-interleaved and fed back to the SISO decoder. Note that in (38), $\phi_{m,t,i}^{e(u)}$ denotes the extrinsic LLR at the last iteration of ARQ round $u = 1, \dots, k$.

6.4 Implementation Issues

In signal-level turbo combining, the computation of the forward and backward filters $\mathbf{F}_t^{(k)}$ and $\mathbf{B}_t^{(k)}$ involves, at each turbo iteration of ARQ round k , one inversion of the $N_T(\varepsilon + L - 1) \times N_T(\varepsilon + L - 1)$ matrix $\mathbf{Y}^{(k)} + \sigma^2 \mathbf{\Xi}^{-1}$. The cost of computing $\mathbf{F}_t^{(k)}$ and $\mathbf{B}_t^{(k)}$ is therefore in the order of $\mathcal{O}(N_T^3 \varepsilon^3)$ complex operations. This indicates that the computational complexity of the signal-level combining scheme is less sensitive to k . The number of ARQ rounds only influences the number of additions required for performing recursions (21) and (29), which is in the order of $N_T^2(\varepsilon + L - 1)^2 + N_R \varepsilon T$ complex additions. Note that the operations of computing $\mathbf{H}^{(k)H} \mathbf{H}^{(k)}$ and $\mathbf{H}^{(k)H} \mathbf{y}_i^{(k)}$ are also required in the case of symbol-level combining. Therefore, the computational cost of forward and backward filters is almost the same for both signal and symbol-level turbo combining schemes. The LLR-level turbo combining algorithm approximately involves the same amount of operations as symbol-level combining.

In the case of signal-level combining, memory requirements are determined by (21) and (29), where two $N_T(\varepsilon + L - 1) \times N_T(\varepsilon + L - 1)$ and $N_T(\varepsilon + L - 1) \times T$ complex matrices are required to accumulate channel matrices $\mathbf{H}^{(k)H} \mathbf{H}^{(k)}$, and signal vectors $\mathbf{z}_0^{(k)}, \dots, \mathbf{z}_{T-1}^{(k)}$, respectively. Note that the two recursions (21) and (29) play a key role in signal-level turbo combining since they avoid the storage of all signals and channel matrices as in MAP turbo combining. In symbol-level combining, only N_T complex matrices of size $K \times T$ and two $K \times N_T$ complex matrices are required to store filter outputs and their corresponding parameters given by (35), (36), and (37). Therefore, signal-level combining requires slightly more memory than its symbol-level counterpart. Finally, in LLR-level turbo combining, a real vector of size $N_T M T$ is required to combine extrinsic values. Therefore, the three combining strategies have similar implementation requirements. They slightly differ in the number of additions and storage memory.

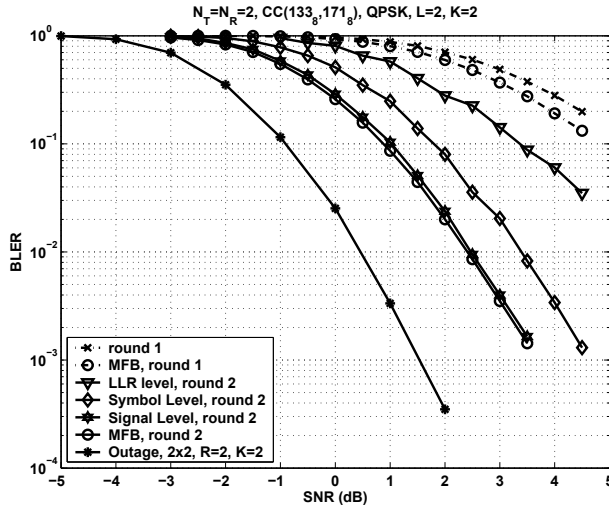


Fig. 5. BLER performance comparison for $N_T = N_R = 2$ and QPSK.

6.5 BLER Performance

In this subsection, we provide simulated BLER performance for the packet combining strategies studied in the previous subsections. The main focus of the analysis we provide is to show that signal-level turbo combining has better ISI cancellation capability and diversity gain compared with the other combining schemes.

We consider an STBICM transmitter with a 64-state $\frac{1}{2}$ -rate convolutional code whose polynomial generators are $(133_8, 171_8)$. The length of the code frame is 1800 bits. The modulation scheme is quadrature phase shift keying (QPSK). The MIMO-ISI channel has the same profile as in Subsection 5.2, i.e., two equal power taps. The ARQ delay is chosen $K = 2$ according to the theoretic analysis in Subsection 5.2. In all figures, the BLER is per ARQ round, and the SNR is per symbol per receive antenna.

We compare the resulting BLER performance with the outage probability and the matched filter bound (MFB). Note that for the purpose of fair comparison, the computation of the outage probability does not take into account the rate distortion as in (14). The MFB curves are obtained for each transmission assuming perfect ISI cancellation and maximum ratio combining (MRC) of all time, space, multipath, and delay diversity branches.

In Fig. 5, we consider an STBICM code with $N_T = N_R = 2$ transmit and receive antennas. This corresponds to a rate $R = 2$. The filter length is chosen equal to $\varepsilon = 9$ ($\varepsilon_1 = \varepsilon_2 = 4$) for all combining schemes. A quick inspection of Fig. 5, shows that both signal and symbol-level turbo combining offer a significant performance improvement after the second ARQ round compared with LLR-level combining. The signal-level scheme has better ISI cancellation capability compared with symbol-level combining. It almost achieves the MFB, while the symbol-level scheme presents a gap of approximately 1dB compared with the MFB. Also, note that both signal and symbol-level combining achieve the asymptotic slope of the outage probability.

In Fig. 6, we evaluate the BLER performance of a ST-BICM code with $N_T = 4$. This corresponds to a rate $R = 4$. The number of receive antennas is $N_R = 2$. Note that this

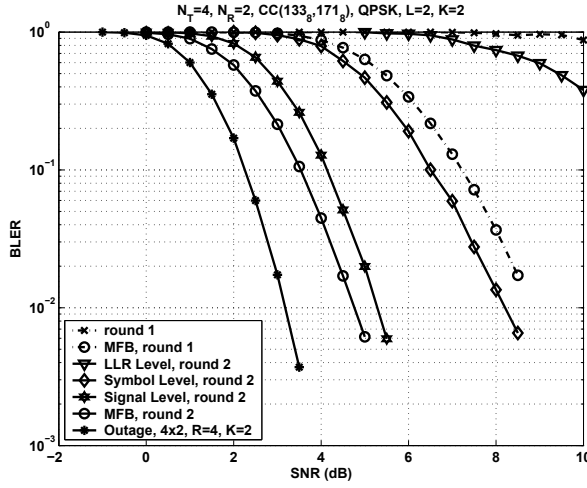


Fig. 6. BLER performance comparison for $N_T = 4$, $N_R = 2$ and QPSK.

type of unbalanced MIMO configurations where the transmitter is equipped with more antennas than the receiver is suitable for the forward link. The filter length is increased to $\varepsilon = 13$ ($\varepsilon_1 = \varepsilon_2 = 6$) for all combining schemes. The signal-level combining technique is shown to achieve BLER performance close to the MFB with a gap less than 0.5dB, while both the LLR-level and the symbol-level techniques have degraded BLER performance. The signal-level combining manifests itself in almost achieving the diversity gain of the MIMO-ISI ARQ channel, while it is shown that symbol-level combining fails to do so. This is because in the second ARQ round, the signal-level scheme constructs a 4×4 virtual MIMO-ISI channel for ISI cancellation and symbol detection, while the MIMO configuration remains unbalanced in the case of symbol and LLR-level combining.

7. Conclusion

In this chapter, we considered the design of efficient iterative turbo packet combining algorithms for broadband MIMO systems with Chase-type ARQ. We derived the structure of the optimal MAP turbo packet combining technique that exploits all the diversities available in the MIMO-ISI ARQ channel to perform packet combining, and analyzed its outage probability. As optimal MAP turbo packet combining has a huge computational cost and memory requirements, we introduced a new class of low-complexity MMSE-based turbo packet combining schemes. In MMSE-based signal-level combining, each ARQ round is viewed as a set of virtual receive antennas, and packet combining is jointly performed with ISI cancellation at the signal level. In MMSE-based symbol-level combining, multiple transmissions are separately turbo equalized, and combining is performed at the level of filter outputs. The simulation results provided in this chapter indicate that signal-level combining provides better BLER performance than that of symbol-level and conventional LLR-level combining.

8. References

- Ait-Idir, T. & Saoudi, S. (2009) Turbo packet combining strategies for the MIMO-ISI ARQ channel, *IEEE Transactions on Communications* vol. 57, no. 12, December 2009, 3782-3793
- Caire, G. & Tuninetti, D. (2001). ARQ protocols for the Gaussian collision channel, *IEEE Transactions on Information Theory*, Vol.47, No.4, July -2001, 1971-1988
- El Gamal, H.; Caire, G. & Damen, M. O. (2006) The MIMO ARQ channel diversity-multiplexing-delay tradeoff, *IEEE Transactions on Information Theory*, Vol.52, No.8, August -2006, 3601-3621
- Garg, D. & Adachi, F. (2006) Packet access using DS-CDMA with frequency-domain equalization *IEEE Journal of Selected Areas in Communications*, vol. 24, no. 1, Jan. 2006
- Onggosanusi, E. N.; Dabak, A. G.; Hui, Y. & Jeong, G. (2003) Hybrid ARQ transmission and combining for MIMO systems, *Proceedings of IEEE International Conference on Communications*, pp. 3205-3209, ISBN 0-7803-7802-4, Anchorage, May 2003
- Samra, H. & Ding, Z. (2006) New MIMO ARQ protocols and joint detection via sphere decoding, *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 473-482, Feb. 2006
- Tse, D. & Viswanath, P. (2005). *Fundamentals of Wireless Communication*, Cambridge University Press, ISBN 978-0-521-84527-4
- Wolff, R. (1989). *Stochastic Modeling and the Theory of Queues*, Upper Saddle River, NJ: Prentice-Hall, 1989
- Wolniansky, P. W.; Foschini, G. J. & Valenzuela, R. A. (1998). V-BLAST: an architecture for realizing very high data rates over the rich scattering wireless channel, in *Proc. Int. Symp. Signals, Systems, Electron.*, Pisa, Italy, Sep. 1998

Cooperative ARQ: A Medium Access Control (MAC) Layer Perspective

Jesús Alonso-Zárate*, Elli Kartsakli**, Luis Alonso**
and Christos Verikoukis*

**Centre Tecnològic de Telecomunicacions de Catalunya (CTTC)*

***Universitat Politècnica de Catalunya (UPC-EPSC)*

1. Introduction

Cooperative Automatic Retransmission reQuest (C-ARQ) schemes have become a very active research topic over the last years. C-ARQ schemes constitute a practical way of executing cooperation in wireless networks with already existing equipment. C-ARQ schemes exploit feedback from the receiver, i.e. cooperation is only executed when needed, and thus are sometimes referred to as cooperation on-demand cooperative schemes.

In short, the idea of C-ARQ is to exploit the fact that, due to the broadcast nature of the wireless channel, any transmission can be received by any of the stations in the transmission range of the transmitter. What has been traditionally considered as interference, is exploited in C-ARQ schemes to attain spatial diversity. Upon a transmission error, a retransmission can be requested from any (or some) of the stations which overheard the original transmission, which can act as spontaneous helpers (or relays). The result is that the destination of a packet can receive different copies of the same information arriving via statistically independent transmission paths, i.e., space diversity.

C-ARQ schemes have been already studied in the literature from a theoretical point of view and there is no doubt that, under some conditions, they can dramatically boost the performance of wireless communications compared to traditional ARQ, where retransmissions are performed only from the source. However, involving a number of users in a communication link requires coordination. To this end, efficient Medium Access Control (MAC) protocols are necessary to get the maximum efficiency of the communications. In this chapter we emphasize the important role of the MAC layer in this context of C-ARQ.

Along the chapter, we first review in Section 2 the motivation and operation of C-ARQ schemes into detail. We go through the parameters that affect the performance of these schemes and we point out the role of the MAC layer. Taking into account the specific requirements of the MAC layer in this kind of schemes, we present in Section 3 a novel high-performance MAC protocol specifically tailored for this purpose. Computer-based simulations are presented to evaluate the performance of the protocol. Finally, Section 4 concludes the chapter.

2. Cooperative ARQ (C-ARQ)

2.1 Background and Motivation

Traditionally, ARQ schemes have been used in communication networks to guarantee the reliable delivery of data packets. Upon the reception of a packet with errors, retransmissions are requested from the source (and along the same channel) until either the packet can be properly decoded or it is discarded for the benefit of the backlogged data.

Several variations of ARQ schemes have been proposed in the past to improve the performance of communications. These schemes perform well in wired networks where there is no correlation between consecutive packet error probabilities, i.e., packet errors are random and sparse. However, their performance in wireless networks is compromised by phenomena such as the shadowing and fading of the radio channel. In wireless channels, packet errors might come into bursts, and thus if a packet is received with errors, the immediate retransmissions will be also received with errors with high probability if they are performed through the same channel (Zorzi et al., 1997).

C-ARQ schemes constitute a practical solution to combat this fading nature of the wireless channel. Their operation is described in the following section.

2.2 Description of C-ARQ

Consider a wireless network formed by an arbitrary number of stations equipped with half-duplex radio frequency transceivers. In order to be able to execute a C-ARQ scheme, all the stations must listen to (overhear) every ongoing transmission in order to be able to cooperate if required. In addition, they should keep a copy of any received data packet (regardless of its destination address) until it is acknowledged (positively or negatively) by the destination. This packet is discarded whenever the destination successfully decodes the original packet.

It is assumed that, although both error detection and Forward Error Correction (FEC) bits are attached to all the transmitted data packets, errors can still occur due to the severe wireless channel impairments. Whenever a destination receives a data packet with unrecoverable errors, it broadcasts a retransmission request in the form of a control packet. This packet is referred to as the Call for Cooperation (CFC) packet. A *cooperation phase* is then initiated.

A subset of the stations which overheard both the original transmission from the source and the CFC from the destination, become *active relays* or *helpers*. As it will be further discussed later, some relay selection criteria can be attached to the CFC in order to activate the most appropriate subset of stations to act as helpers. Orthogonally in time (TDMA), frequency (FDMA or OFDMA), or code (CDMA), these active relays attempt to retransmit a copy of the original packet to assist in the failed transmission. For the sake of clarity in the explanation and without loss of generality, the data packets retransmitted by the relays will be referred to as *cooperative packets*.

Eventually, the destination might either receive a correct copy of the original packet from a relay or may be able to properly combine the different retransmissions from the relays to successfully decode the original packet. Otherwise, if the destination is not able to recover the data packet after some predefined time (cooperation time-out), it discards it. In any of the two cases, the cooperation phase is finished.

Although slight different variations to this general operation can be found in the literature, most of the proposed C-ARQ schemes follow this description. It is worth mentioning that the CFC has sometimes received the name of Negative ACK (NACK) in the literature (Dianati et al. 2006). However, this name falls short in describing the real function of the CFC. Besides informing the Negative ACK, it also calls for cooperation and, indeed, it could attach some relay selection criteria, among other control information required for the execution of a cooperative technique.

An example of operation of a C-ARQ mechanism is illustrated in Fig. 1. Therein, the communication between a source and a destination stations is assisted by an arbitrary number (N) of relays. In this particular example, the relays retransmit data orthogonally in time until the destination station can send the ACK.

The performance of a C-ARQ scheme might be mainly influenced by the following *four* parameters:

1) *The relay selection criteria*; as it could be expected, the number of potential helpers and the “quality” of those helpers will have a direct impact on the efficiency of the C-ARQ scheme. For this reason, there are several works focused on the design of efficient techniques to select either the best or a subset of the best potential helpers to act as relays (Gómez et al., 2007; Biswas & Morris, 2005).

2) *The PHY forwarding technique executed by the relays* (Nosratinia et al., 2004):

a. *Amplify and forward techniques*, when the relays transmit an amplified version of the original received signal, without demodulating or decoding it.

b. *Compress and forward techniques*, when the relays transmit a compressed version of the original transmitted signal, without decoding it.

c. *Decode and forward techniques*, when the relays transmit recoded copies of the original message. Note that using decode and forward, the recoding process can be done on the basis of repeating the original codification, recoding the original data (or only a relevant part of it), or using more sophisticated Space-Time Codes (STC) (Fitzek & Katz, 2006).

3) *The number of required retransmissions* necessary to decode a packet which can mainly depend on:

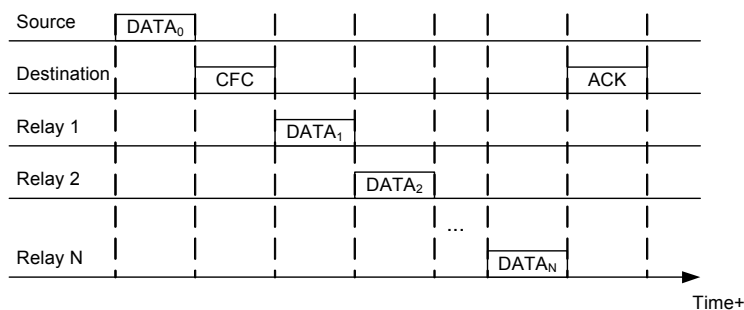


Fig. 1. C-ARQ Scheme with Time-Orthogonal Relays

a. The channel conditions between the source and the destination, the source and the relays, and the relays and the destination (Gómez & Pérez-Neira, 2006; Pfletschinger & Navarro, 2008).

b. The transmission scheme, which includes the forwarding technique executed by the relays and the combination technique executed by the destination station to combine the different retransmissions received from independent paths. The approach of combining different erroneous copies of a same packet to decode the original packet has been tackled in the past (Charabarty et al., 2005; Morillo-Pozo & García-Vidal, 2007).

4) *The MAC protocol* which is necessary to tackle with the contention among the relays. Just as an example, the ideal scheduling among the relays represented in Fig. 1 is impossible to attain in fully distributed networks without a central coordinator. Therefore, the set of active relays should contend for the channel in order to retransmit the packets. Efficient MAC protocols are necessary to execute a C-ARQ scheme in order to exploit the benefits of cooperation in wireless networks.

2.3 Motivation and Contributions of the Chapter

C-ARQ schemes have been so far analyzed from a fundamental point of view and mainly with emphasis on the PHY layer (Dianati et al. 2006; Zimmermann et al., 2004; Zimmermann et al., 2005; Gupta et al., 2004; Cerruti et al., 2008; Morillo-Pozo et al., 2005). These previous works put in evidence that C-ARQ schemes can yield an improvement in performance, lower energy consumption, and interference, as well as an extended coverage area by allowing communication at low Signal to Noise Ratios (SNRs). However, all of these contributions assume simplified topologies (with one or very few relays) and perfect scheduling among the relays at the MAC level. This scheduling might be difficult to attain in the fully decentralized scenario represented by the *cloud* of relays without infrastructure. Therefore, both the design of efficient MAC protocols and the evaluation of the actual performance of C-ARQ techniques considering the MAC overhead are mandatory if C-ARQ schemes are to find real application. Indeed, *this* is the main motivation for this chapter.

The focus in this chapter is on *time-orthogonal C-ARQ schemes*, which might be the easiest approach to implement with already existing off-the-shelf equipment. By slightly modifying the wireless controller (or driver), existing wireless cards could implement a C-ARQ scheme. The emphasis is on the design and analysis of a novel MAC protocol to deal with the unique characteristics of the contention process that takes place among the active relays within a cooperation phase. Note that in the considered C-ARQ schemes, upon the initialization of the cooperation phase, the network has the three following *unique* characteristics:

- 1) The spontaneous "sub-network" formed by the active relays is ad hoc and thus there is no infrastructure responsible for managing the access to the channel.
- 2) This sub-network formed by the active relays surrounding the node calling for cooperation is *suddenly* (sharply) set into saturation conditions whenever the cooperation phase is initiated. Upon the transmission of a CFC packet, all the active relays have a data packet ready to transmit in order to assist the failed transmission. Therefore, heavy contention takes place in a previously idle network.
- 3) Opposite to general communications systems, now fairness is not a major issue to achieve. Indeed, the main goal is to attempt to assist the failed transmission as fast and reliable as possible, minimizing the use of the radio resources.

These three characteristics determine the way MAC protocols should be designed within the context of C-ARQ schemes in wireless networks. Considering the aforementioned

characteristics, we present in this chapter the design and performance evaluation of a novel high-performance MAC protocol for C-ARQ schemes, named DQCOOP.

It is worth mentioning at this point that, in the literature, there exists a family of *cooperative MAC* protocols which have not been designed for the execution of C-ARQ schemes in wireless networks, but they are aimed at solving other kind of interesting cooperative issues. For completeness, they are overviewed in the following section.

2.4 Related Work: Cooperative MAC Protocols

Some MAC protocols for cooperative communications have been proposed in the literature. Most of them have been designed to achieve a throughput enhancement, but actually none of them takes into account all of the unique characteristics of the on-demand C-ARQ schemes. It has to be mentioned that all these MAC protocols have been designed more as routing protocols with a cross-layer design that takes into account the transmission rates to decide the shortest route to a destination than as MAC protocols themselves. In what follows, a summary of the most relevant contributions is summarized.

In (Liu et al. 2007) two versions of the CoopMAC protocol are designed in the context of 802.11b WLANs in order to solve the performance anomaly induced by the multi-rate capability of the Distributed Coordination Function (DCF) of the standard (IEEE 802.11 Standard, 2007). Users with low transmission rate occupy the channel for long periods of time, reducing the overall throughput of the system and reducing the throughput seen by stations with higher transmission rates. The main idea of CoopMAC protocols is that stations transmit first to intermediate stations at a higher rate and then those intermediate stations transmit to the access point, reducing the total transmission delay. In the first version of CoopMAC, referred to as CoopMAC I, any station keeps updated a table with those stations that could potentially help in a transmission. Before transmitting any packet, a station calculates the shorter transmission path, either using direct communication with the intended destination or through any of the potential helpers with an entry in the table. In the case of using a relay, a previous handshake is done between the source station and the selected relay in order to ensure the validity of the route. The main drawback of CoopMAC I is that it requires the addition of three new fields in the Request To Send (RTS) frame and the addition of a new control frame named Helper ready To Send (HTS). As an alternative, CoopMAC II is proposed to overcome this problem. This second version of CoopMAC uses available empty fields in regular IEEE 802.11 control frames and eliminates the handshake between destination and helpers. Although the implementation is simpler, version II is more vulnerable to a change in the availability of a helping station caused by mobility. Computer simulations in (Liu et al., 2007) demonstrate the improved performance achieved with either CoopMAC I or II. Moreover, Korakis et al. implemented the protocol in actual WLAN cards, as reported in (Korakis et al., 2006). The main contribution of their work is the description of the overall implementation process and the limitations found when attempting to actually implement the protocol. These limitations were mainly due to the constraints imposed by the time sensitive tasks performed by the firmware of the wireless cards. In addition, the CoopMAC has been also adapted to wireless networks using directional antennas in (Tau et al., 2007).

On the other hand, both the Cooperative-MAC (CMAC) and FEC CMAC (FCMAC) protocols were presented in (Shankar et al., 2005) within the context of 802.11e networks to improve the overall performance and to ensure a certain QoS. In CMAC, a station detecting

an erroneous packet transmission between any other pair of source-destination stations decides to cooperate by retransmitting a copy of the overheard transmission as long as the received packet has no errors. A random backoff mechanism with a constant backoff window is applied to avoid collisions among different helpers. The size of the contention window of the helpers has to be very small in comparison to the contention window of the source in order to ensure that helpers retransmit their copy before the original source retransmits on its own the failed packet. Each helper transmits the copy of the packet at most once, to ensure that all available helpers cooperate and thus the benefits of diversity are obtained. On the other hand, FCMAC extends the operation of CMAC by fragmenting data packets into smaller blocks. Each block contains its own inner FEC field and the whole packet contains an outer FEC. Upon error detection of a whole packet, only a predefined number of randomly selected blocks among those received without errors are retransmitted. If the retransmitted blocks are those that were received with errors at destination, then the performance is improved. Otherwise, the increased overhead becomes useless. A possible solution consists in adding a negative acknowledgement (NACK) sent out by the destination upon error detection, indicating which are the blocks received with errors. However, the use of NACK in CMAC would imply higher overhead and again, it would require hardware modifications, thus breaking with the claimed backwards compatibility. The main limitation of CMAC and FCMAC is that they rely on the fact that helpers can learn whether other transmissions between any pair of source and destination are successful or not only by overhearing the radio channel.

In (Wang & Yang, 2005), the Cooperative Diversity Medium Access with Collision Avoidance (CD-MACA) protocol is proposed within the context of wireless ad hoc networks operating over the CSMA/CA protocol. Whenever a source terminal fails to receive the CTS packet, all those stations that had properly received it, take the place of the source terminal and retransmit the data packet. An analytical model based on Markov chain theory is proposed to obtain the achievable throughput of the system considering cooperation. Although the general idea of CD-MACA is rather interesting, the definition in (Wang & Yang, 2005) is quite general and several implementation details are not considered.

From an energy-efficient perspective, another cooperative MAC protocol is also presented within the context of ad hoc networks in (Azing et al., 2005). This proposal integrates cooperative diversity into two different wireless routing protocols by embedding a distributed cooperative MAC. The initial path establishment performed by the routing protocol can be done either considering cooperation or not. Cooperation is then achieved by forcing all the stations to act as a distributed virtual antenna, through which simultaneous transmissions are separated with CDMA.

In (Sadek et al., 2006) a cooperative MAC protocol was presented within the context of a mesh network formed by an access point, a number of regular stations, and one fixed wireless router (relay). A fixed TDMA scheme is applied and empty slots are used for cooperative relaying. The relay station keeps a copy of all those packets that are not properly received by the Access Point (AP). At the beginning of each time slot, the relay listens to the channel. If the channel is idle, it retransmits the packet at the head of its queue. Based on this main idea, two specific algorithms are proposed to exploit the benefits of cross-layer design between the PHY and MAC layers.

All these MAC protocols have been designed to achieve an improvement in the network performance by transmitting through faster multi-hop routes. However, none of them takes

into account the unique characteristics of the C-ARQ schemes for their implementation in on-demand cooperative schemes. DQCOOP is presented in the next section as a novel MAC protocol that has been tailored to meet the requirements of the C-ARQ scenario. It constitutes the adaptation of the high-performance DQMAN protocol (Alonso-Zárate et al., 2008a) to this kind of scenarios.

3. DQMAN for C-ARQ: DQCOOP

The aim of this section is to present DQCOOP as an extension and adaptation of the high-performance DQMAN (Alonso-Zárate et al., 2008a) to match the unique requirements posed by the C-ARQ schemes. DQMAN, in its turn, is the extension of the infrastructure-based DQCA protocol (Alonso-Zárate et al., 2008b) for wireless ad hoc networks. The new resultant protocol is called DQCOOP. The rules of DQMAN and DQCA will not be described into detail in this chapter as they can be found in (Alonso-Zárate et al., 2008a) and (Alonso-Zárate et al., 2008b), respectively.

In short, the basic idea of DQMAN is that any idle station with data to transmit listens to the channel for a randomized period of time before establishing its cluster. This Clear Channel Assessment (CCA) period gets the name of Master Selection Phase (MSP). If the channel is idle for the whole MSP, then a cluster is established. The station becomes master and starts broadcasting a periodical clustering beacon (CB) that allows neighbor stations to get synchronized and become slaves. The master operates as such for as long as there is data activity within its cluster. Therefore, the cluster structure changes along time as a function of the aggregate traffic load of the network. Once the cluster is established, the master station transmits its own data and it acts as the AP of a WLAN wherein DQCA can be executed. For completeness, we review the basic protocol rules of DQCA in the next section.

3.1 DQCA Overview

The purpose of this section is to highlight the basic features of DQCA. As demonstrated in (Alonso-Zárate et al., 2008b), DQCA outperforms the widely commercially spread Distributed Coordination Function (DCF) of the IEEE 802.11 Standard and remains stable even when the traffic load occasionally exceeds the channel capacity.

DQCA is a MAC protocol designed to manage the access to the channel in the uplink of an infrastructure WLAN. Time is divided into MAC frames, and each frame is divided in three parts separated by a Short Inter Frame Space (SIFS) necessary to tolerate propagation delays, turnaround times, and processing delays. The three parts, depicted in Fig. 2, are:

- i) A Contention Window (CW) further divided into m access minislots wherein the nodes can send a short chip sequence named Access Request Sequence (ARS) to request access to the channel. An ARS is a short chip sequence that contains no explicit information but has a specific and predefined pattern that allows the AP to distinguish between an idle minislot, the presence of just one ARS, or the occurrence of a collision between two or more simultaneous ARS.
- ii) A data slot reserved for the transmission of data packets.
- iii) A feedback part wherein the AP broadcasts a Feedback Packet (FBP) that contains the data acknowledgment, the state of the each of the minislots of the CW for contention resolution algorithm, and a 'final message bit' that is enabled (set to one) by the AP to identify the last data packet (fragment) of a message. Of course, nodes must also include a

‘final message bit’ in their data packet transmissions in order to advertise the transmission of the final fragment of each message.

All the nodes execute three sets of simple rules at the end of each MAC frame. By simply using the feedback information attached to the FBP, they can update the state of two distributed queues (explained below) to execute the access algorithm. According to the protocol rules, DQCA operates as a random access protocol when the traffic load is low (an immediate access rule of the protocol allows a station to get access to the channel immediately if the distributed queues are empty), and it switches smoothly and automatically to a reservation protocol as the traffic load increases. Therefore, it attains the better of the access methods.

The protocol operation is based on two concatenated distributed queues, the Collision Resolution Queue (CRQ) and the Data Transmission Queue (DTQ). The CRQ is responsible for the resolution of collisions among ARS and the DTQ handles the transmission of data. The number of occupied positions (or elements) in each queue is represented by an integer counter (RQ and TQ for the CRQ and the DTQ, respectively). Both counters have the same value for all the nodes in the system and are updated according to a set of rules at the end of each frame. Each node must also maintain and update another set of counters that reveal its position in the queue (pRQ and pTQ for the CRQ and the DTQ, respectively). By the term “position” it is meant the relative order of arrival (or age) of the node in the respective queue. In the CRQ, each position (or element) is occupied by a set of nodes that suffered an ARS collision (i.e. attempted an ARS transmission in the same access minislot of the same CW). The DTQ contains the nodes that successfully reserved the channel through an ARS and therefore each queue element corresponds to exactly one node.

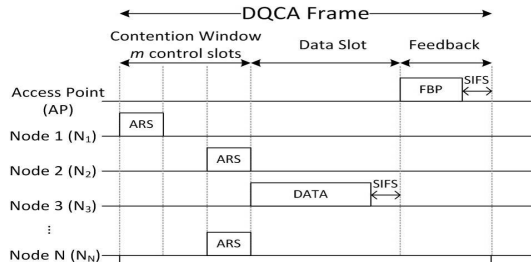


Fig. 2. DQCA Frame Structure

3.2 Motivation and Problem Statement

The intuitive idea behind DQCOOP is that the *destination* asking for cooperation gets the role of *master* and coordinates the retransmissions from the relays, which become *slaves*, as in DQMAN. Then, a temporary cluster is established around the destination and a variation of DQCA can be executed. This is represented in Fig. 3.

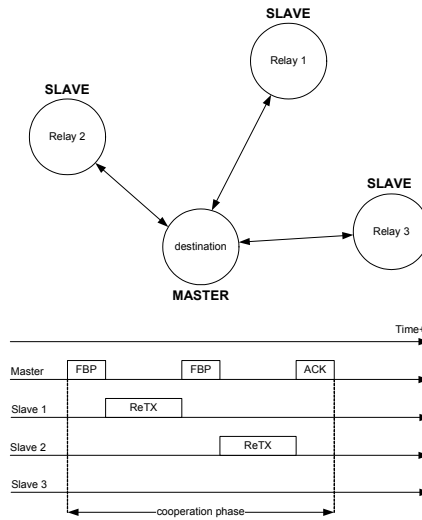


Fig. 3. Master-Slave Architecture of DQCOOP (simplified example)

The master, i.e., the destination, initiates the periodic broadcast of the FBP and creates a temporary cluster. A cooperation phase is initiated. The slaves, i.e., the relays, request access to the channel to retransmit their cooperative packet (retransmissions of the original source transmissions) by executing a variation of the DQCA rules. It is assumed that the relays attempt to retransmit persistently until the cooperation phase is finished. Whenever the cooperation phase is finished either an ACK or a NACK packet is transmitted, indicating either the successful or unsuccessful recovery of the data packet originally received with errors, respectively.

However, DQMAN, as defined in (Alonso-Zárate et al., 2008a), would be inefficient in managing the access to the channel in a C-ARQ scheme. This is mainly due to the fact that upon cooperation request (broadcast by the destination), the group of active relays forms an ad hoc network wherein all the active stations suddenly have a data packet ready to be transmitted. This turns temporarily the network from idle to saturation conditions. This idle-to-saturation sharp transition would cause DQMAN to spend a non-negligible start-up time before attaining its high performance, mainly due to:

- 1) The simultaneous channel access requests from the active relays in the first frame immediately after the transmission of the CFC would have a high probability of collision. Therefore, some empty frames would be needed until the first collision could be solved and data retransmissions could actually start.
- 2) Upon the transmission of the first FBP, all the active relays (slaves) would retransmit in the following frame by executing the *immediate access rule* of DQCA (Slotted ALOHA access for low traffic loads). All these transmissions would collide, causing a waste of resources for the duration of a complete MAC frame.
- 3) Even with the immediate access rule disabled, an empty frame would be present when the collision resolution process starts due to the MAC frame structure with the feedback broadcast at the end of the frame.

Therefore, it is necessary to expand and adapt the DQMAN operation to take into consideration the aforementioned issues that may potentially degrade its performance in C-ARQ schemes. DQCOOP is presented in the next section with the goal of attaining the near-optimum performance of DQMAN within the context of the considered C-ARQ scheme.

3.3 Protocol Description

The core operation of DQCOOP is highly based on DQMAN. However, the *clustering algorithm* and the *MAC protocol* (frame structure and protocol rules) are modified to meet the requirements of the C-ARQ scheme. Their descriptions are presented in the next two sections.

3.3.1 Clustering Algorithm

In DQCOOP, the clustering algorithm of DQMAN is modified as it follows:

- 1) The destination, and not the transmitter as in DQMAN, takes the *master* role when a cooperation phase is initiated with the transmission of a CFC packet. Some of the relays which received the original data packet (received with errors by the destination to trigger a cooperation phase) and also receive the CFC transmitted by the destination become active relays. These active relays get the role of *slaves*. A cluster is then established. The master periodically broadcasts a FBP, in the same way as in DQMAN, to provide the slaves with the minimum feedback information necessary to execute the protocol rules at the end of each frame.
- 2) There is no CCA prior to the establishment of the cluster. This means that the destination station does not have to contend with other users to get access to the channel. Therefore the contention within the MSP associated with DQMAN is avoided with DQCOOP. This can be actually performed as the CFC is transmitted instead of the ACK when receiving a packet with errors. ACK packets are usually given priority over all kind of traffic (in wireless networks), and thus there is no need for contention in this case.
- 3) The cluster is broken up whenever the master either manages to decode the original packet or discards the packet. The cooperation phase is ended with the transmission of the ACK packet. Otherwise, if a maximum time-out expires and the original packet cannot be decoded, a NACK packet is transmitted and the cluster is broken up as well. That is, in fact all the stations become idle upon the transmission of either the ACK or the NACK by the master.

3.3.2 The MAC Protocol: Frame Structure and Protocol Rules

When a cooperation phase is initiated, time is divided into five parts as represented in Fig. 4. Upon the transmission/reception of each FBP, all the stations execute the protocol rules of DQCA. The five parts of a cooperation phase within the context of DQCOOP are:

- 1) A **CFC transmission**. The cooperation phase is initiated when a CFC is broadcast by the destination station upon the reception of a data packet with errors. This CFC takes the form of a special FBP and indicates that **immediate access is forbidden**.
- 2) An **initial contention window composed of m_0 minislots** follows the CFC transmission wherein every active relay station randomly selects (with equal probability) one out of the m_0 minislots where to send an Access Request Sequence (ARS).
- 3) A **FBP transmission**. A FBP is broadcast by the master station with the **feedback**

information regarding the state of each of the m_0 previous minislots. As in DQMAN, for each minislot, this information can have one out of three values. It can be empty (E), i.e., no ARS transmitted, success (S), i.e., exactly one ARS transmitted, or collision (C), i.e., more than one ARS transmitted in the same minislot (no matter how many).

4) **A number of regular DQMAN consecutive MAC frames** follow this first FBP until the cooperation phase is ended. The rules of DQMAN, with the exception of the immediate access rule, are executed to manage the data retransmissions and the resolution of the collisions. **The contention window of these frames has m minislots**, where in general $m < m_0$, although this is not a mandatory condition.

5) **An ACK or NACK transmission.** Whenever the destination is able to **successfully decode** the original packet, it broadcasts an **ACK** packet indicating the end of the cooperation phase. A **NACK** is transmitted if the packet cannot be decoded at some point in time.

Short Inter Frame Spaces (SIFS) are left between each of the parts of the cooperation phase to compensate for non-negligible propagation and data processing delays and turnaround times to switch the radio transceiver from receiving to transmitting mode.

It is worth mentioning that the value of m_0 must be tuned according to the expected number of active relays. The higher the number of active relays, the higher the value of m_0 in order to reduce the probability that all the access requests collide in the first frame. However, a high value for m_0 has a cost in terms of control overhead. On the other hand, as long as at least one access request is successful, the data transmission process can be initiated from the first MAC frame, avoiding thus the loss of resources.

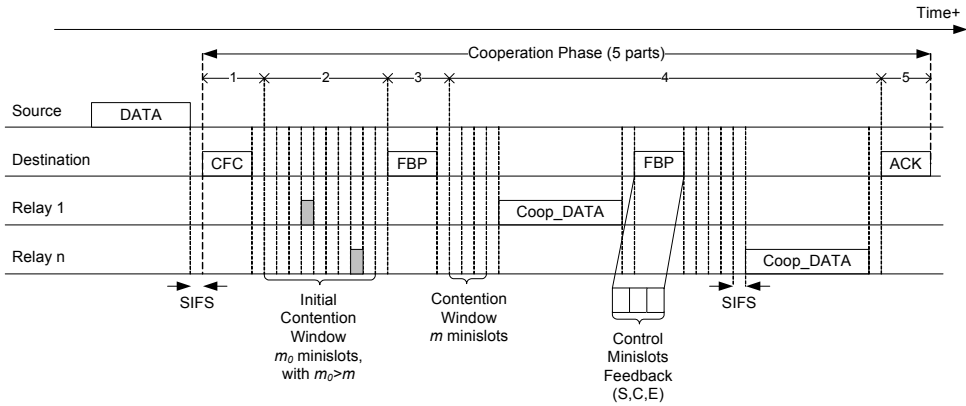


Fig. 4. DQCOOP MAC Frame Structure

3.3.3 Operational Example

A simple network layout with six stations is considered, all of them in the transmission range of each other. A source station (S) transmits to a destination station (D) with the support of relays $R1$, $R2$, $R3$, and $R4$. TQ and RQ represent the size of the DTQ and the CRQ, respectively, and pTQ_i and $pCRQ_i$ represent the position of the i^{th} user in the DTQ and CRQ, respectively.

The cooperation phase is represented in Fig. 5 and explained as follows:

- 1) Upon the reception of the data packet with errors, D initiates a cooperation phase by broadcasting a CFC. This packet sets the start of frame 0.
- 2) Frame 0 contains 5 access minislots ($m_0=5$). The set of relays $\{R1, R2, R3, R4\}$ select the set of minislots $\{3, 1, 5, 5\}$.
- 3) At the end of frame 0, D broadcasts the FBP with the following feedback information regarding the state of the minislots, i.e., $\{Success, Empty, Success, Empty, Collision\}$.
- 4) Upon the execution of the protocol rules, $R2$ gets the first position of DTQ, $R1$ gets the second position of DTQ, and both $R3$ and $R4$ get the first position of CRQ. In terms of the four integer number representing the queues, this can be written as $\{pTQ1, pTQ2, pTQ3, pTQ4\}=\{2,1,0,0\}$ and $\{pRQ1, pRQ2, pRQ3, pRQ4\}=\{0, 0, 1, 1\}$. On the other hand, $TQ=2$ and $RQ=1$.
- 5) During frame 1, both the data transmission and the collision resolution work in parallel. At the beginning of the frame, containing 2 access minislots ($m=2$), $R3$ and $R4$ attempt to solve their collision. They reselect an access minislot where to send an ARS. In this case, they select minislots 1 and 2 respectively, and thus they successfully solve their collision.
- 6) On the other hand, $R2$, which is at the first position of DTQ, transmits data (a retransmission of the original packet).
- 7) At the end of frame 1, the FBP broadcast by D indicates that a transmission has been successful and the next station in DTQ should transmit in the following frame. In addition, the feedback information on the state of the minislots allows $R3$ and $R4$ to queue, orderly in time, in DTQ.

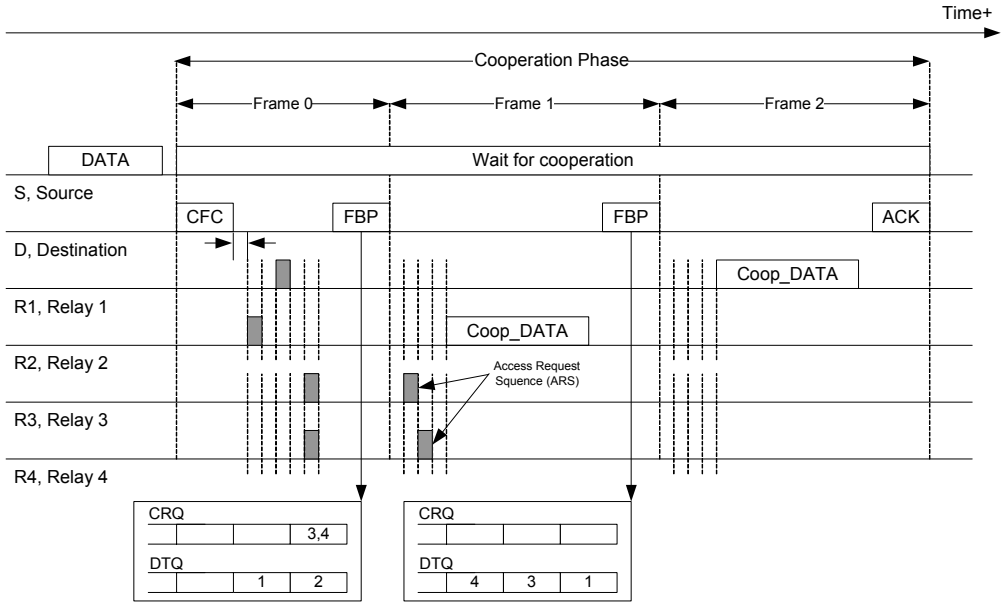


Fig. 5. DQCOOP Example of Operation

- 8) In frame 2, there are no collisions to be solved and thus the minislots are empty. $R1$ transmits data.
- 9) Upon the reception of the retransmission from $R1$, D is able to successfully decode the original packet. Therefore, it transmits an ACK packet indicating the end of the cooperation phase. All the relays discard the buffered cooperative packet.

3.4 Performance Evaluation

The performance of DQCOOP is evaluated in this section through a C++ custom-made simulator. In order to focus on the evaluation of the cooperation phases, a single-hop network wherein all the data transmissions from a fixed source to a fixed destination are received with errors is considered. That is, the destination always broadcasts a CFC packet upon the reception of every original data packet received from the source station. Moreover, the source has always a packet ready to be transmitted to the destination.

In this performance evaluation, we will measure the average packet transmission delay defined as the period of time elapsed from the moment that a packet is first transmitted from the source until it can be decoded at destination after receiving K retransmissions. For all the experiments, we assume that a constant number of relays are activated within each cooperation phase. Furthermore, and without loss of generality, the destination is considered to require a constant number K of retransmissions from the relay set to decode the original packet.

The simulation parameters are summarized in Table 1.

Parameter	Value	Parameter	Value
Control Rate	6 Mbps	MAC header	34 bytes
Data Rate (Source)	24 Mbps	PHY preamble	96 μ s
Data Rate (Relays)	54 Mbps	ACK, CFC, FBP length	14 bytes
Packet Length	1500 bytes	<i>SlotTime</i>	10 μ s
ARS	10 μ s	SIFS	10 μ s

Table 1. Simulation Parameters

3.4.1 Number of Minislots in the Cooperation Phase

The average packet transmission delay as a function of the number of active relays in a cooperative phase is represented in Fig. 6 when $K=3$ and for different values of m_0 and m , which in addition accomplish that $m_0=m$. Each curve represents the results obtained with different number of access minislots (m).

For low values of m , the average packet transmission delay gets lower as the value of m increases thanks to the faster collision resolution process. However, increasing the number of access minislots also increases the MAC overhead. The addition of an extra minislot entails an extension of the frame duration (devoted to overhead) and also enlarges the size of the FBP that contains the state of each one of the minislots. Therefore, as it can be seen in the figure, for high values of m , e.g., $m=10$, the fact that the collision resolution becomes shorter in time does not pay off the increase in the protocol overhead when the number of active relays is low and thus the average packet transmission delay gets higher. This can be better appreciated in Fig. 7 where the average packet transmission delay for the scenario with 5 relays is plotted as a function of the number of access minislots $m=m_0$. In this curve it is easier to see that, for low number of access minislots, an increase in the number of minislots leads to lower average packet transmission delays. However, over a given threshold, the faster resolution of collisions due to the longer contention window does not compensate for the MAC overhead and the average packet transmission delay increases

with the number of access minislots. For this reason, it is necessary to find a good compromise between the faster collision resolution and the protocol overhead. This tradeoff will be further discussed later in the next section.

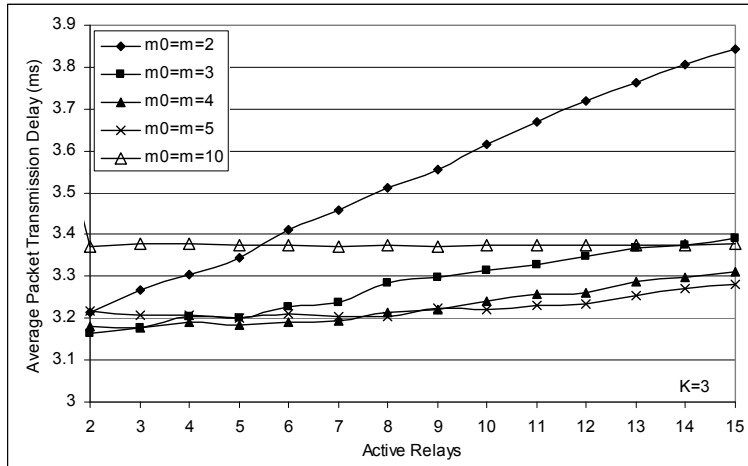


Fig. 6. Average Packet Transmission Delays for Different Values of $m_0=m$

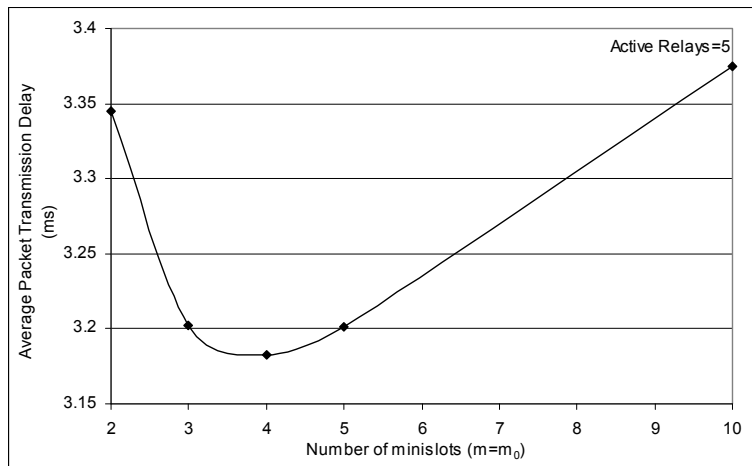


Fig. 7. Average Packet Transmission Delay for Different Values of $m_0=m$

Getting back to the results in Fig. 6, they show that the average packet transmission delay drops remarkably when the number of access minislots is at least equal to 3. Higher values of m do not result in any substantial reduction of this time. Therefore, as it happens with the DQCA protocol (Alonso-Zárate et al., 2008b), a good operational point for DQCOOP is to set $m=3$.

It is interesting to evaluate whether this discussion is still valid for any arbitrary number of required retransmissions (K). The average packet transmission delay is plotted in Fig. 8 as a function of the value of K and for different values of $m_0=m$ when the number of active relays is 15. In all cases, there is a considerable reduction of the average packet transmission delay when shifting from 2 to 3 minislots. However, there is no much interest in increasing the number of access minislots to higher values than 3, at least in terms of packet transmission delay. Therefore, it is important to reinforce the already known argument that the number of access minislots should be set to 3 in any DQCA-like protocol.

However, it seems reasonable to think that the value of m_0 (the number of access minislots within the very first frame after the transmission of the CFC) could be set to a higher value than m in order to *absorb* the first multiple access request arrival from all the active relays. Note that the first frame is the one that receives the maximum number of simultaneous access requests. In subsequent frames, the requests are split into smaller groups according to the m -ary tree-splitting collision resolution operation of DQMAN.

In the next section, m is set to 3 and the performance of the protocol is evaluated for different values of m_0 . The aim is to evaluate the reduction of the average packet transmission delay for $m_0 > m$.

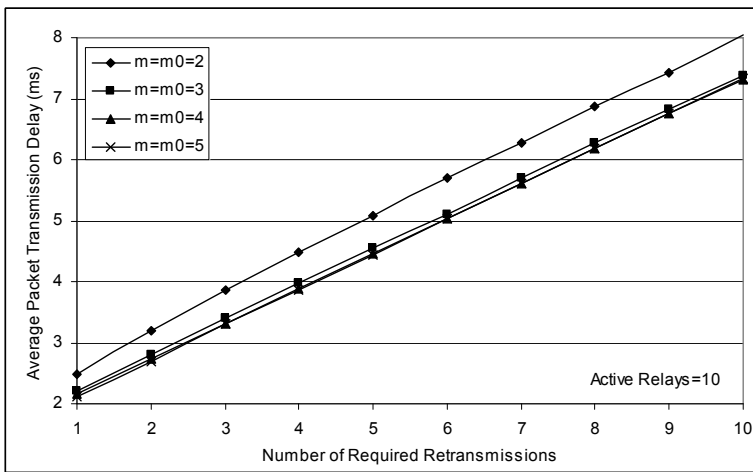


Fig. 8. Average Packet Transmission Times for Different Values of K

3.4.2 Number of Minislots in the Start-up Phase (m_0)

The performance of DQCOOP for different values of m_0 is evaluated in this section. As discussed before, an increase in the number of minislots of the first frame reduces the probability of collision in the first access requests upon initialization of a cooperation phase and, therefore, it should yield a lower average packet transmission delay. However, it also entails an increase of the protocol overhead (frame length and amount of required feedback information).

In order to quantify this tradeoff, first note that the duration of a cooperation phase can be decomposed as the sum of time devoted to the transmission of data and the overhead due to

the necessary MAC protocol. This overhead time includes silent intervals as well as the time devoted to the transmission of control packets. Considering this, the *relative overhead* is defined as the ratio between the overhead time in the cases that $m_0 > 1$ and the overhead time when $m_0 = 1$ (this latter case is the worst case in terms of overhead since all the relays collide in the first access request with probability one). This definition allows plotting the curves with different values of K in the same vertical axis and also makes the results independent of the absolute values of the transmission rates used for the simulation and the numerical evaluation.

The relative overhead is plotted in Fig. 9 as a function of the value of m_0 , for different number of required retransmissions (K), and considering a total number of 5 active relays. The first observation is that there is a close relationship between the overhead of the protocol and the value of m_0 . The value of the relative overhead is very sensitive to the value of m_0 if the number of required retransmissions is low. This means that if the value of K is low, the accurate tuning of the value of m_0 has a remarkable effect on the performance of the C-ARQ scheme. All the curves show a local minimum of the relative overhead for any pair of values of m_0 and K . However, on the other hand, the higher the values of K , the more flat the curves become. This means that if the number of required retransmission is high, the value of m_0 becomes a non-critical parameter on the performance of DQCOOP.

The main reason for this behavior is that when the number of required retransmissions is high and thus the duration of the cooperation phase is long, the impact of the overhead of the first frame on the performance of DQCOOP is low. Note that if K retransmissions are needed, at least K frames are necessary.

On the other hand, it seems reasonable to believe that the selection of the value of m_0 should depend on the number of active relays (which request access simultaneously in the first frame). In order to evaluate this relationship, the average packet transmission delay is plotted in Fig. 10 for $K=3$. Different curves are plotted for different number of active relays and as a function of the value of m_0 .

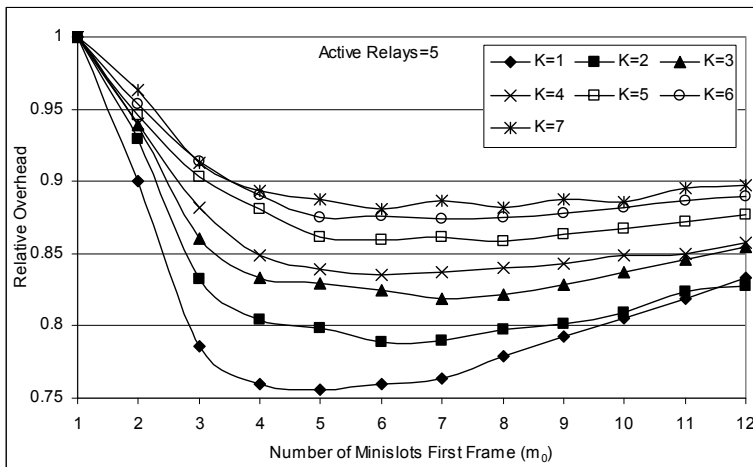


Fig. 9. Protocol Relative Overhead (DQCOOP)

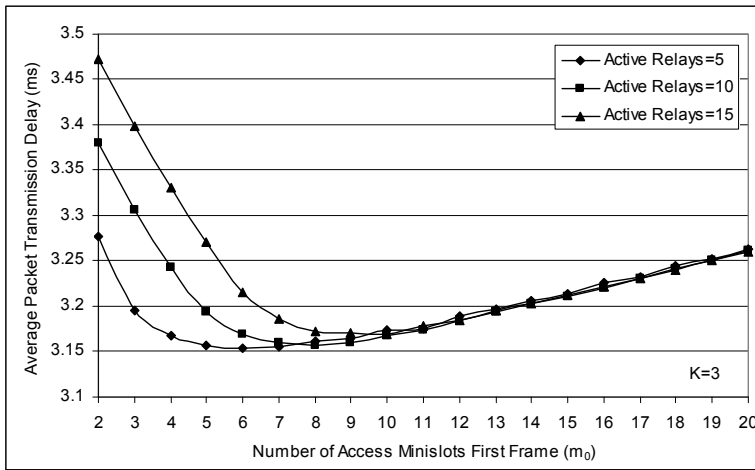


Fig. 10. Average Packet Transmission Delay as a Function of m_0

It is worth noting that when $m_0 \geq 10$ the three curves almost overlap. This means that, if this condition is fulfilled, the average packet transmission delay is almost equal and independent of the number of active relays. In addition, the value of the average packet transmission delay at $m_0=10$ is not substantially bigger than the one at the respective minimum values that can be found for $m_0=6$ (for 5 active relays), $m_0=7$ (for 10 active relays), and $m_0=10$ (for 15 active relays). This constitutes a worthwhile design guideline since by setting $m_0=10$ the average packet transmission delay for any value of K can be predicted with reliable accuracy regardless of the number of active relays in each cooperation phase (considering a practical situation with no more than 15 active relays). In addition, this fact relaxes the configuration requirements of the network, which is of remarkable interest when operating in fully decentralized and spontaneous networks.

4. Conclusions

In this chapter we have highlighted the important role of the MAC layer in the performance of C-ARQ schemes. Typically, these kinds of schemes have been evaluated from fundamental points of view and assuming perfect scheduling among the relays. However, we have shown that efficient MAC protocols are necessary to fulfill the specific requirements posed by C-ARQ schemes and to get the most of their potential to increase the efficiency of wireless communications.

In addition, we have presented the DQCOOP protocol as an extension and adaptation of DQMAN to efficiently coordinate the contention among the relays in a C-ARQ scheme. It has been necessary to redesign the initialization phase of a DQMAN cluster so as to manage the idle-to-sharp traffic transition that takes place upon the transmission of a CFC. Since the active relays attempt to help simultaneously, the first contention window of DQMAN has to be resized. In addition, the protocol frame structure and the protocol rules have been also modified to optimize the performance of DQMAN in the context of C-ARQ schemes.

The performance of the protocol has been evaluated with computer simulations. Results show that the performance of DQCOOP can be independent of number of active relays and the number of access minislots. This is a desirable characteristic in fully decentralized networks, as is the case of ad hoc networks, where there might be no previous knowledge of the network topology and configuration. Results also show that this independency can be simply accomplished by setting the number of access minislots to 3 (attaining a faster resolution of collisions compared to the transmission of data) and properly dimensioning the number of access minislots in the very first frame, which is also modified to avoid an otherwise certain empty data field. This last modification aims at absorbing the first simultaneous access request by all the active relays. In fact, results show that the number of access minislots in the very first frame can be overdimensioned at almost no cost, and thus the performance of DQCOOP can be independent of the number of relays. The cost of increasing by one unit the number of access minislots in terms of overhead pays off the reduced probability of collision in the first access request.

5. References

- Alonso-Zárate J., Gómez J., Verikoukis C., Alonso L., & Pérez-Neira A. (2006). Performance Evaluation of a Cooperative Scheme for Wireless Networks, *Proceedings of the IEEE PIMRC*, pp. 1-5, ISBN 1-4244-0329-4, Helsinki, Finland, September 2006
- Alonso-Zárate J., Kartsakli E., Skianis C., Verikoukis C., & Alonso L. (2008a), Saturation Throughput Analysis of a Cluster-based Medium Access Control Protocol for Single-hop Ad Hoc Wireless Networks, *Simulation: Transactions of the Society for Modeling and Simulation International*, Vol. 84, No. 12, 619-633, ISSN 0037-5497
- Alonso-Zárate J., Kartsakli E., Cateura A., Verikoukis C., & Alonso L. (2008b). A Near-Optimum Cross-Layered Distributed Queueing Protocol for Wireless LAN, *IEEE Wireless Communications Magazine, Special Issue on MAC protocols for WLAN*, Vol. 15, No. 2, (February 2008) 48-55, ISSN 1536-1284
- Azgin A., Altunbasak Y., & AlRebig G. (2005). Cooperative MAC and Routing Protocols for Wireless Ad Hoc Networks, *Proceedings of the IEEE GLOBECOM 2005*, ISBN: 0-7803-9414-3, December 2005
- Biswas S. & Morris R. (2005). ExOR: Opportunistic Multi-hop Routing for Wireless Networks, *Proceedings of the SIGCOMM '05*, pp. 133-144, ISBN 1-59593-009-4, New York, NY, USA, 2005.
- Campbell G. et al. (2002), Method and apparatus for detecting collisions and controlling access to a communications channel, *US Patent no. US6408009 B1*, June 2002.
- Cerruti I., Fumagalli A., & Gupta P. (2008). Delay Model of Single-Relay Cooperative ARQ Protocols in Slotted Radio Networks Poisson Frames Arrivals, *IEEE/ACM Trans. on Networking*, Vol. 16, No. 2, (April 2008) 371-382, ISSN 1063-6692
- Chakraborty S. S., Liinajarja M., & Ruttik K. (2005). Diversity and packet combining in Rayleigh fading channels, *IEE Proceedings-Communications*, Vol. 152, No. 3, (June 2005) 353 - 356, ISSN 1350-2425
- Dianati, M., Ling X., Naik K., & Shen X. (2006). A Node-Cooperative ARQ Scheme for Wireless Ad Hoc Networks, *IEEE Trans. on Vehicular Technology*, Vol. 55, No. 3, (May 2006) 1032-1044, ISSN 00189545

- Fitzek F. H. P. & Katz M. D. (2006). *Cooperation in Wireless Networks: Principles and Applications*, Ed. Springer, ISBN 978-1-4020-4710-7, The Netherlands
- García-Vidal J., Guerrero-Zapata M., Morillo J., & Fusté D. (2007). A Protocol Stack for Cooperative Wireless Networks, In: *Wireless Systems and Mobility in Next Generation Internet, in Lecture Notes in Computer Science (LNCS)*, 62-73, Springer, ISBN: 978-3-540-70968-8, Springer, (2007)
- Goldsmith A. (2005). *Wireless Communications*, Cambridge University Press, ISBN 0521837162
- Gómez J. & Pérez-Neira A. (2006). Average Rate Behavior for Cooperative Diversity in Wireless Networks, *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 5309-5402, ISBN 0-7803-9389-9, Kos, Greece, May 2006
- Gómez J., Alonso-Zárate J., Verikoukis C., Pérez-Neira A., & Alonso L. (2007). Cooperation On Demand Protocols for Wireless Networks, *Proceedings of the IEEE PIMRC*, pp. 5, ISBN: 978-1-4244-1144-3, Athens, Greece, September 2007
- Gupta P., Cerruti I., & Fumagalli A. (2004). Three Transmission Scheduling Policies for a Cooperative ARQ Protocol in Radio Networks, *Proceedings of the WNCG Conference*, Austin, October 2004
- IEEE 802.11 Standard, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE Std. 802.11-99, Revision 2007
- Korakis T., Natayanan S., Bagri A., & Panwar S. (2006). Implementing a Cooperative MAC Protocol for Wireless LAN, *Proceedings of the IEEE International Conference on Communications (ICC'06)*, pp. 4805-4810, ISBN: 1-4244-0355-3, June 2006
- Liu P., Tao Z., Lin Z., Erkip E., & Panwar S. (2006). Cooperative Wireless Communications: A Cross-Layer Approach, *IEEE Wireless Communications Magazine, Special Issue on Advances in Smart Antennas*, Vol. 13, No. 4, (August 2006) 84-92, , ISSN 1536-1284
- Liu P., Tao Z., & Panwar S. (2007). CoopMAC: A Cooperative MAC for Wireless LANs, *IEEE Journal on Selected Areas on Communications*, Vol. 25, No.2, (February 2007) 340-354, ISSN 0733-8716
- Morillo-Pozo J., García-Vidal J., & Pérez-Neira A. I. (2005). Collaborative ARQ in Wireless Energy-Constrained Networks, *Proceedings of the 2005 Joint Workshop on Foundations of Mobile Computing 2005 (DIAL-POM'05)*, ISBN 1-59593-092-2
- Morillo-Pozo J. D., & García-Vidal J. (2007). A Low Coordination Overhead C-ARQ Protocol with Frame Combining, *Proceedings of the IEEE PIMRC*, pp. 1-5, ISBN 978-1-4244-1144-3, Athens, Greece, September 2007
- Nosratinia, A., Hunter, T. E., & Hedayat A. (2004). Cooperative Communications in Wireless Networks, *IEEE Communications Magazine*, Vol. 42, No. 10, (October 2004) 74-80, ISSN 0163-6804
- Pfletschinger S. & Navarro M. (2008). Link Adaptation with Retransmissions for Partial Channel State Information, *Proceedings of the IEEE GLOBECOM 2008*, pp. 1-6, ISBN 978-1-4244-2324-8, New Orleans, Louisiana, USA, November 2008
- Tao Z., Korakis T., Slutskiy Y., Panwar S., & Tassiulas L. (2007). Cooperation and Directionality: A Co-opdirectional MAC for Wireless Ad Hoc Networks, *Proceedings of the WiOpt 2007*, pp. 1-8, ISBN: 978-1-4244-0960-0, April 2007

- Shankar S., Chou C., & Ghosh M. (2005). Cooperative Communication MAC (CMAC) – A new MAC protocol for Next Generation Wireless LANs, *Proceedings of the IEEE International Conference on Wireless Networks, Communications, and Mobile Computing 2005*, pp. 1-6, ISBN: 0-7803-9305-8, June 2005
- Sadek B., Ray Liu K. J., & Ephremides A. (2006). Collaborative Multiple-Access Protocols for Wireless Networks, *Proceedings of the IEEE International Conference on Communications (ICC'06)*, pp. 4495-4500, ISBN: 1-4244-0355-3, June 2006
- Wang X. & Yang C. (2005). A MAC Protocol Supporting Cooperative Diversity for Distributed Wireless Ad Hoc Networks, *Proceedings of the IEEE PIMRC*, pp. 1396-1400, ISBN: 9783800729098, Berlin, Germany, September 2005
- Zimmermann M. D., Herhold P., & Fettweis G. (2004). The Impact of Cooperation on Diversity-Exploiting Protocols, *Proceedings of the 59th IEEE Vehicular Technology Conference*, pp. 410-414, ISBN 0-7803-8255-2, Milan (Italy), May 2004
- Zimmermann E., Herhold P., & Fettweis G. (2005). On the Performance of Cooperative Relaying Protocols in Wireless Networks, *European Trans. on Telecommunications*, Vol. 16, No. 1, (January 2005) 5-16, ISSN 1541-8251
- Zorzi M., Rao R. R., & Milstein L.B. (1997). ARQ error control for fading mobile radio channels, *IEEE Trans. on Vehicular Technology*, vol. 46, no. 2, (May 1997) 445-455, ISSN 00189545

A Hybrid Feedback Mechanism to Exploit Multiuser Diversity in Wireless Networks

Yahya S. Al-Harthi

*Electrical Engineering Department
King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia*

Abstract

One of the most promising approaches to boost the communication efficiency in wireless systems is the use of multiuser diversity (MUDiv), where the fading of channels is exploited. The mechanism of scheduling the user with the best channel condition is called opportunistic scheduling (OS). In this paper we propose a joint polling and contention based feedback (JPCF) algorithm that exploits MUDiv while reducing the feedback load. The guard time, which is between bursts, is divided into minislots that alternate between polling-based feedback minislot (p -minislot) and contention-based feedback minislot (c -minislot). During the minislot, users feedback their channel qualities if above a predetermined threshold. We analyze the scheduling algorithm under slow Rayleigh fading assumption and derive the closed-form expressions of the feedback load as well as the system capacity. We also consider the delay resulting from the time needed to schedule a user and derive the system throughput. The scheduling algorithm is compared with other scheduling algorithms.

1. Introduction

With the emerging of new multimedia applications and the huge demand and growth for such applications, the need to provide high-speed high-rate transmission techniques is demanding. One technique that has the capability of supporting such applications is multiuser diversity (MUDiv), where the fading of channels is exploited (1). Assume that a reasonably large number of users are actively transmitting/receiving packets in a given cell, and they experience independent time-varying fading conditions. By transmitting data to only the instantaneous "on-peak" user, *opportunistic scheduling* (OS) can efficiently utilize the wireless resources and thus dramatically improve the overall system throughput (2), (3). In order to schedule the best user, in terms of channel quality, each user measures his instantaneous signal-to-noise ratio (SNR) and feeds it back to the network scheduler. Currently Qualcomm's high data rate (HDR) system require similar scheme (4). As the number of active users becomes high, resources are wasted in carrying this amount of feedback, which lead to inefficient use of the spectrum. This issue motivated researchers to propose new techniques to reduce the feedback load while exploiting MUDiv.

Many investigations have been conducted to exploit multiuser diversity while keeping the overhead as minimum as possible. For instance, the impact of the degree of quantization of the SNR measurements on the throughput of constant rate transmission was investigated in

(5). It was shown that reducing the feedback rate (few quantization levels) yields a good performance compared to the unquantized feedback. In single carrier systems, (6) proposed a discrete rate switch-based multiuser diversity (DSMUDiv) algorithm that reduced the feedback load and feedback rate while preserving the performance of OS. The scheduling algorithm relied on a probing mechanism. One drawback of this algorithm is that as the average SNR decreases the need for full probing increases, therefore, the algorithm suffers from high feedback load at low average SNR values. The work was extended to multi-carrier systems in (7), where further reduction in the feedback load was shown. Similarly, in (8) users compare their sum-rate on all sub-carriers to a predefined threshold value. In (9), the threshold value was optimized to meet a specified outage probability. Users with channel quality above the threshold are allowed to feedback their SNR measurements, while all others remain silent. In case no feedback, a random user is selected, which leads to loss in capacity. Also, the feedback values are unquantized (analog values), which increases the feedback rate. In (10), the work was extended, where the scheduler requests full feedback if none of the users' channel qualities are above the threshold. Although the loss in capacity was compensated by full feedback as opposed to (9), the feedback load was increased. In (11), multiple feedback thresholds are used in order to reduce the feedback load and exploit multiuser diversity. The feedback load was reduced at an expense of scheduling delay. In addition, a set of switched-based multiuser access schemes were proposed in (12) in order to reduce the feedback load. In (13), (14), the feedback rate was reduced to a one bit feedback per user. The scheduler uses these feedback bits to partition all users into two sets and assigns the channel to one user belonging to the set experiencing favorable channel conditions. Although this reduces the feedback load and feedback rate but some loss is expected due to the low resolution of the quantized feedback (one bit). Other work have considered multi-carrier systems, for instants, in (15) a clustering scheme was proposed to reduce the feedback load while slightly degrading the performance. Each user feeds back a figure-of-merit listing its strongest clusters of subcarriers. Similarly, in (16) a fixed number of subcarriers for each user is fed back instead of a full feedback. These subcarriers are either the best K out of M subcarriers or K predetermined subcarriers. All previously mentioned work considered a polling threshold-based scheduling schemes to reduce the feedback load.

Other work considered random access algorithms to exploit multiuser diversity while reducing the overhead. In (17) a distributed multiaccess scheme was proposed. Based on a channel quality threshold, users report the best m channels with quality above the threshold value. Optimization was performed to find the best threshold and m values to maximize the system capacity. In (18), a random access threshold-based feedback scheme was proposed to exploit diversity gain while reducing the feedback load. Parameter optimization is performed to allow only users with high channel gains to feedback, which maximizes the system capacity. In (19), a medium access control protocol was designed based on a splitting algorithms to resolve collisions over a sequence of minislots, and determine the user with the best channel. Contention resolution algorithms while exploiting the diversity gain were proposed in (20). Finally, reduced feedback overhead algorithms were studied in (21). Static splitting was considered for the best effort traffic scenario. Whereas, for the traffic mixture scenario combine contention and polling based feedback was considered to maintain quality of service.

While Polling-based algorithms, like the DSMUDiv, guarantee best user selection, it suffers from scheduling delay. On the other hand, although contention-based algorithms reduces the overhead more then centralized scheduling it suffers from loss in capacity. In this chapter we introduce an opportunistic scheduling algorithm that is not only a polling-based feedback

algorithm like in (6), but it is a joint polling and contention based feedback (JPCF) algorithm. During the guard time, which is divided into minislots, users feedback their channel qualities, based on a predetermined channel threshold, either in a polling-based feedback minislot (p -minislot) or/and a contention-based feedback minislot (c -minislot). When a the best user is found data transmission begins. The scheduling algorithm is analyzed under slow Rayleigh fading assumption and closed-form expressions of both the feedback load and the system capacity are derived. We also look at the effect of the delay on the throughput, where we derive the system throughput. The scheduling algorithm is compared with the DSMUDiv (6) and the optimal (full feedback) selective diversity scheduling algorithms.

This chapter is organized as follows. Section 2 introduces the system model. Section 3 and 4 present the scheduling algorithm and the mathematical analysis of the JPCF algorithm, respectively. In Section 5 we present some numerical examples. Finally, Section 6 ends the paper with some concluding remarks.

2. System Model

We consider a single free interference cell in a wireless network with K active users communicating with a base station (BS). We assume downlink scheduling, where only one user is allowed to receive data transmission in each time slot. The communication is based on time division duplex (TDD), where downlink and uplink channels are reciprocal.

2.1 Downlink Transmission Model

Let,

$$y_i(T) = h_i(T) \cdot x(T) + n_i(T); i = 1, 2, 3, \dots, K \quad (1)$$

be the baseband channel model, where $x(T) \in \mathbb{C}$ is the transmitted signal in time slot T and $y_i(T) \in \mathbb{C}$ is the received signal of user i in time slot T . The noise processes $n_i(T)$ are independent and identical distributed (*i.i.d.*) sequences of zero mean complex Gaussian noise with variance σ_n^2 . The fading channel gain from the BS to the i th user in time slot T is $h_i(T)$. We adopt a quasi-static fading channel model where $h_i(T)$ is *i.i.d.* from burst to burst but remains constant over each burst. We consider a flat Rayleigh fading model, assuming the fading coefficients of all users are *i.i.d.* Therefore, $h_i(T)$ is a zero-mean complex Gaussian random variable. The amplitude of $h_i(T)$, $\alpha_i(T) = |h_i(T)|$ is Rayleigh distributed with the probability density function (PDF) given by,

$$f_{\alpha_i}(\alpha_i) = \frac{2\alpha_i}{\Omega_i} \exp\left(-\frac{\alpha_i^2}{\Omega_i}\right) \quad (2)$$

where $\Omega_i = E[\alpha_i^2]$ is the average fading power of the i th user.

2.2 Quantized Feedback

In this work we assume that users adapt their modulation level based on the instantaneous channel condition. Using this transmission strategy, which is called adaptive modulation (AM), if the channel is strong at a given time, the transmission can occur with a higher constellation size. Otherwise, a lower constellation size has to be used. Note that the switching thresholds of the adaptive transmission modes are functions of the modulation scheme and target error performance. We consider adaptive multilevel quadrature amplitude modulation (M-QAM) scheme (22). Specifically, we consider a transmission scheme employing uncoded adaptive discrete rate M-QAM schemes with constellation sizes $\mathbb{M} = \{M_n : M_n < M_{n+1}, 0 \leq$

$n \leq N\}$, where $M_0 = 1$ is the user outage (deep fade), and M_N is the highest modulation level. We assume perfect channel estimation and negligible time delay between channel estimation and signal set adaptation, and as such, the rate adaptation can happen instantaneously. We also assume error free feedback channels. If we denote the target average bit error probability (BEP) by BEP_o , thresholds or switching thresholds can be obtained according to (22, eq.(30)):

$$\begin{aligned}\gamma_{th}^{(1)} &= [\text{erfc}^{-1}(2 \cdot \text{BEP}_o)]^2, \\ \gamma_{th}^{(n)} &= -\frac{2}{3}(2^n - 1) \ln(5 \cdot \text{BEP}_o); \quad n = 2, 3, \dots, N, \\ \gamma_{th}^{(N+1)} &= +\infty,\end{aligned}\tag{3}$$

where $\text{erfc}^{-1}(\cdot)$ denotes the inverse complementary error function.

Similar to (6), with the assumption of discrete rates, a user estimates his SNR and instead of feeding back the analog value it is mapped to a quantized value that represent the modulation level, which can be supported, and then this quantized value is fed back.

Define the set of quantized values (binary bits) $Q = \{q^{(n)} : q^{(n)} < q^{(n+1)}, 0 \leq n \leq N\}$ that represents the modulation levels, ($q^{(n)} \rightarrow_{map} M_n$). If we assume n modulation levels, then each quantized value $q^{(m)}$, $0 \leq m \leq n$, will be $\log_2 n$ bits in length. To illustrate the idea of the SNR mapping, assume γ_i is the estimated SNR of the i th user, then the quantized value is,

$$q_i = \begin{cases} q^{(0)} & \text{if } \gamma_i < \gamma_{th}^{(1)} \text{ (outage),} \\ q^{(n)} & \text{if } \gamma_{th}^{(n)} \leq \gamma_i < \gamma_{th}^{(n+1)} \\ q^{(N)} & \text{if } \gamma_i \geq \gamma_{th}^{(N)}. \end{cases}\tag{4}$$

2.3 Uplink Feedback Structure

Fig. 1 shows the feedback channel structure, which consists of minislots. The number of minislots constructing the feedback channel can vary from 1 minislot, in the case the first user is granted the channel access, to $2K - 1$ minislots, where K is the total number of users, in the case all users feed back their channel qualities. This variation in length assumes that data transmission can begin at different time periods. Such assumption can be possible if dynamic spectrum access is implemented. The JPCF scheduling algorithm assumes that the feedback is polling-based or contention-based. The feedback channel is divided into polling-based feedback minislots (p -minislot) and contention-based feedback minislots (c -minislot), where each p -minislot is followed by a c -minislot, except the last p -minislot. The contention protocol is based on a modified slotted ALOHA protocol (23). As any random multiple access protocol one out of three possible outcomes will occur: no access, one access, and a collision. Multiple users feeding back the requested information will cause signals to interfere at the receiver which we consider as a collision. When a collision occurs no resolution is consider and the information is discarded.

3. Scheduling Algorithm

In this section, we will introduce our proposed algorithm, the JPCF scheduling algorithm. Similar to the optimal selective scheduling algorithm, the JPCF algorithm always guarantees that the best user is granted the channel access. The difference between the two is in the selection process. In the JPCF algorithm decisions on whom to schedule are based on threshold

test. In simple terms, the BS compares the channel quality of a user to a predetermined threshold value and selects the user if his channel quality exceeds the threshold. In case no user is found with channel quality exceeding the threshold, then the user with the highest channel quality among all is selected.

Let us define the following:

- $\mathbb{L} = \{(D(i), l) : 1 \leq l \leq K \text{ is the probing order, and } D(i) \text{ is the ID of user } i\}$. The set is generated randomly each time slot.
- The predetermined threshold value is set to $\gamma_{th}^{(N)}$. According to (3), $\gamma_{th}^{(N)}$ is the SNR threshold of the highest modulation level and $q^{(N)}$ is the quantized threshold value.
- The probability of contention in the c -minislot is ρ .

The feedback mechanism of the JPCF algorithm is based on both polling period and contention period. Each period is a minislot during. Fig. 1 shows the feedback channel structure, which consists of p -minislots and c -minislots. The following describes the scheduling algorithm (Fig. 2 shows the flowchart):

- (i) Through the broadcast channel all users know their polling order (\mathbb{L}).
- (ii) The feedback channel always begins with a p -minislot.
- (iii) During the p -minislot the BS polls the channel state information of the user (his SNR) who is in order.
- (iv) The feedback channel is terminated and the data transmission begins if the polled user has $q_l = q^{(N)}$. By knowing (γ_l) and according to (4) q_l is determined.
- (v) If $q_l \neq q^{(N)}$, then his information is stored and the feedback channel continues with a c -minislot following the previous p -minislot and step (vi) is performed.
- (vi) During the c -minislot, which is a contention based minislot, users feedback their SNRs, with probability ρ , if it satisfy $q = q^{(N)}$. Otherwise, they keep silent.
- (vii) If the contention is successful, then the feedback channel is terminated and the data transmission begins for that successful user. Otherwise, the feedback channel continues with a p -minislot following the previous c -minislot and step (iii) is performed.
- (viii) In case all users are polled, the BS picks the user with the highest channel quality among them and the data transmission begins for that user.
- (ix) In case a tie occurs, a random pick is performed.

4. Performance Analysis

The performance of the JPCF algorithm is evaluated in this section. Basically, we concentrate on the feedback load, the system capacity, and the scheduling delay. Closed form expressions for all three performance measures are also derived.

4.1 Feedback Load

We define the average feedback load (AFL) as the average number of responses until a user is scheduled. The feedback response includes the response on the p -minislot, and the response on the c -minislot. In the c -minislot, a contention resulting a success or a collision is counted as one response, and zero response is considered in case no contention occurs.

Consider two consecutive minislots (p -minislot and c -minislot) and denote the discrete random variable $\eta \in [1, 2]$ by the total number of feedback during them. Let \mathbf{U} be the set of events where a successful search occurs during the two minislots. Note that a *successful search* occurs when a user with $q = q^{(N)}$ is found. Therefore, the occurrence of a successful search is associated with one of the following events: (i) The polled user during the p -minislot has $q = q^{(N)}$, or (ii) The polled user during the p -minislot has $q < q^{(N)}$ and a successful contention occurs in the following c -minislot.

The probability of \mathbf{U} given l probes is:

$$\Gamma^{\mathbf{U}}(l, \rho) = \left[1 - F_{\gamma}(\gamma_{th}^{(N)}) \right] + \left[F_{\gamma}(\gamma_{th}^{(N)}) \left\{ \sum_{i=1}^{K-l} \binom{K-l}{i} \left(1 - F_{\gamma}(\gamma_{th}^{(N)}) \right)^i \left(F_{\gamma}(\gamma_{th}^{(N)}) \right)^{K-l-i} \times \left(i\rho(1-\rho)^{i-1} \right) \right\} \right], \quad (5)$$

where

$$F_{\gamma}(\zeta) = P[\gamma < \zeta] = (1 - e^{-\frac{\zeta}{\gamma}}), \quad (6)$$

is the cumulative distribution function (CDF) of the SNR (γ). The first part of the right hand side of (5) refers to the success in the p -minislot and the second part refers in the success in the c -minislot.

let \mathbf{X} represent the set of events where no feedback occurs on the c -minislot and \mathbf{Y} to be the set of events where a feedback occurs on the c -minislot. Conditioning on l , the probability of \mathbf{X} and the probability of \mathbf{Y} , respectively, are:

$$\Gamma^{\mathbf{X}}(l, \rho) = \left[1 - F_{\gamma}(\gamma_{th}^{(N)}) \right] + \left[F_{\gamma}(\gamma_{th}^{(N)}) \left\{ \sum_{i=0}^{K-l} \binom{K-l}{i} \left(1 - F_{\gamma}(\gamma_{th}^{(N)}) \right)^i \left(F_{\gamma}(\gamma_{th}^{(N)}) \right)^{K-l-i} \times \left((1-\rho)^i \right) \right\} \right], \quad (7)$$

and

$$\Gamma^{\mathbf{Y}}(l, \rho) = F_{\gamma}(\gamma_{th}^{(N)}) \left\{ \sum_{i=0}^{K-l} \binom{K-l}{i} \left(1 - F_{\gamma}(\gamma_{th}^{(N)}) \right)^i \left(F_{\gamma}(\gamma_{th}^{(N)}) \right)^{K-l-i} \left(1 - (1-\rho)^i \right) \right\}. \quad (8)$$

The conditioned expected value of η is:

$$\bar{\eta}(l, \rho) = 1 \cdot \Gamma^{\mathbf{X}}(l, \rho) + 2 \cdot \Gamma^{\mathbf{Y}}(l, \rho). \quad (9)$$

Therefore, the AFL is:

$$\begin{aligned} \text{AFL}(\rho) &= \sum_{l=1}^{K-1} \left(\sum_{i=1}^l \bar{\eta}(i, \rho) \right) \prod_{i=1}^{l-1} \left(1 - \Gamma^{\mathbf{U}}(i, \rho) \right) \Gamma^{\mathbf{U}}(l, \rho) \\ &\quad + \left[\left(\sum_{l=1}^{K-1} \bar{\eta}(l, \rho) \right) + 1 \right] \prod_{l=1}^{K-1} \left(1 - \Gamma^{\mathbf{U}}(l, \rho) \right). \end{aligned} \quad (10)$$

4.2 System Capacity

The average spectral efficiency (ASE) is defined as the average transmitted data rate per unit bandwidth in bits/sec/Hz for specified power and target error performance. In this work we are considering discrete rates and using quantized SNR values (q). For example: if two users with $\gamma_1 \in [\gamma_{th}^{(i)}, \gamma_{th}^{(i+1)}]$, $\gamma_2 \in [\gamma_{th}^{(i)}, \gamma_{th}^{(i+1)}]$, and $\gamma_1 \neq \gamma_2$, then according to (4) both users feedback $q^{(i)}$. Therefore, scheduling either one will result a transmission rate of $\log_2 M_i$ bps/Hz.

Based on the algorithm's description in Section 3, it is clearly seen that both the DSMUDiv scheduling algorithm proposed in (6) and the JPCF scheduling algorithm will always select the best user. Therefore, resulting a similar performance in terms of spectral efficiency.

It has been shown in (6) that with i users in the system, scheduling the best user yields the following average spectral efficiency:

$$\begin{aligned} R(i) = & b_o \left([F_\gamma(\gamma_{th}^{(1)})]^i \right) \\ & + \sum_{n=1}^{N-1} b_n \left([F_\gamma(\gamma_{th}^{(n+1)})]^i - [F_\gamma(\gamma_{th}^{(n)})]^i \right) \\ & + b_N \left(1 - [F_\gamma(\gamma_{th}^{(N)})]^i \right) \end{aligned} \quad (11)$$

where $b_n = \log_2 M_n$ is the number of bits per constellation.

Therefore, the ASE of the JPCF algorithm is:

$$ASE = R(K). \quad (12)$$

In the above expression (12) it is assumed that the guard time duration is negligible, meaning that the feedback rate will not degrade the total system spectral efficiency. Practically, this is not valid. Therefore, the amount of bits transmitted as feedback has to be counted and at the end it will influence the performance of the system. To look into this issue, which is spectral efficiency degradation caused by the feedback traffic, we define a the system capacity as [bits/channel use]. Assuming N modulation levels, then we need $\log_2 N$ bits to represent them. Therefore, the system capacity is:

$$C_{sys} = R(K) - \frac{AFL \cdot \log_2 N}{S}, \quad (13)$$

where S is the number of symbols transmitted in the data transmission time slot.

In (13), the last term takes into account the amount of bits transmitted as feedback. Also, in the same expression the time delay is not taken into account, which is the guard time duration. The only consideration is the feedback rate, or the amount of bits transmitted as feedback. In the next section we investigate the effect of delay on the system performance.

4.3 Scheduling Delay

In this section we investigate the scheduling delay and its effect on the system performance. This time delay is part of the system resources and it is important to identify the amount of resources consumed when performing the JPCF algorithm.

4.3.1 Guard time

The scheduling process will take place during the guard time (τ_g), which is between bursts. The delay resulted from this scheduling process is measured as the time needed to schedule a user, which is simply the time duration of the minislots used (idle minislots are counted) until data transmission is allowed. By looking at Fig. 1, we can see that $\tau_f \leq \tau_g \leq (2K-1)\tau_f$. For simplicity, we assume that both p -minislot and c -minislot have the same time length (τ_f). Assuming a successful search at the l th p -minislot, then the guard time is:

$$\tau_g(l) = \tau_f(2l-1), \quad (14)$$

with probability:

$$P^{(p)}(l) = \left[\prod_{i=1}^{l-1} \left(1 - \Gamma^U(i, \rho) \right) \right] \cdot \left[1 - F_\gamma(\gamma_{th}^{(N)}) \right]. \quad (15)$$

On the other hand, if the successful search occurs at the l th c -minislot, then the guard time is:

$$\tau_g(l) = 2l\tau_f, \quad (16)$$

with probability:

$$P^{(c)}(l) = \left[\prod_{i=1}^{l-1} \left(1 - \Gamma^U(i, \rho) \right) \right] \cdot \left[\Gamma^U(l, \rho) - \left(1 - F_\gamma(\gamma_{th}^{(N)}) \right) \right]. \quad (17)$$

Therefore, the average guard time is:

$$\begin{aligned} \bar{\tau} = & \sum_{l=1}^{K-1} \left[\tau_f(2l-1)P^{(p)}(l) + 2l\tau_fP^{(c)}(l) \right] \\ & + \left((2K-1)\tau_f \right) \cdot \left[\prod_{i=1}^{K-1} \left(1 - \Gamma^U(i, \rho) \right) \right]. \end{aligned} \quad (18)$$

4.3.2 System throughput

The system throughput (STH) is the amount of data bits transmitted per time, where this time includes the data transmission time (T_d) and the guard time (τ_g). In (12), we looked at the amount of bits per data transmission time, where the effect of the scheduling delay was not included. To have a better insight, we derive the average system throughput (ASTH) by taking into account the effect of the guard time duration.

The average system throughput is:

$$\begin{aligned} \text{ASTH} = & \sum_{l=1}^{K-1} \left[\left(\frac{T_d - \tau_f(2l-1)}{T_d} \right) P^{(p)}(l) + \left(\frac{T_d - 2l\tau_f}{T_d} \right) P^{(c)}(l) \right] \cdot R(K) \\ & + \left(\frac{T_d - (2K-1)\tau_f}{T_d} \right) \cdot R(K) \cdot \left[\prod_{i=1}^{K-1} \left(1 - \Gamma^U(i, \rho) \right) \right]. \end{aligned} \quad (19)$$

4.4 Parameters Optimizations

The algorithm's objective is to strictly schedule the best user, therefore the search process will last until a user with $q = q^{(N)}$ is found, which depends on $\gamma_{th}^{(N)}$ and ρ . In this section we investigate the optimization of ρ with the objective to minimize the feedback load:

$$\begin{aligned} \{\rho\} = \arg \min_{\{\rho\}} \text{AFL}(\gamma_{th}^{(N)}, \rho, K), \\ \text{subject to } 0 \leq \rho \leq 1. \end{aligned} \quad (20)$$

Similarly, the optimization solution is well defined by the equation:

$$\begin{cases} \frac{\partial \text{AFL}(\rho, K)}{\partial \rho} = 0, \\ 0 \leq \rho \leq 1. \end{cases} \quad (21)$$

Note that the optimization may not be convex. The derivation of (10) is involved so we apply an exhaustive search method to find an optimal value. Table 3 shows the optimal values with different K given the parameters in Section 5. It is clearly seen from Table 3 that for a given K , ρ has two optimal values at the two average SNR regions. This is due to the collision. At low average SNR values, the value of ρ is increased to encourage good users to compete, whereas, at higher average SNR values the value of ρ is decreased to lower the possibility of collision.

5. Numerical Results

In this section we elaborate on the the performance of the JPCF algorithm by presenting some numerical examples. We compare its performance with both the DSMUDiv and the optimal selective diversity scheduling schemes. Parameters values are found in Tables 1 and 2. Fig. 3 shows the normalized average feedback load (*i.e.*, the average feedback load divided by the number of users K) of the JPCF algorithm for different values of ρ , which is the probability of contention. As the value of the average SNR changes from low to high, the optimal value of ρ , at which the feedback is minimized, changes from $\rho = 1$ to $\rho = 0.13$, for a given value of K . Table 3 shows the optimal values of ρ for different values of K . The reason for the change of the optimal value is the threshold value. For a given threshold value, the percent of users with channel quality exceeding it increases as the average SNR increases, which leads to higher probability of collision. As more users contend the feedback load increases, which forces the value of ρ to change to maintain the minimization in (20). The point at which this change happens depend on the value of $\gamma_{th}^{(N)}$ and the BEP_0 .

The comparison of the JPCF algorithm with both the DSMUDiv and the optimal selective diversity scheduling algorithms is presented in Fig. 4. In the figure, although the DSMUDiv algorithm has decreased the feedback load compared to the optimal algorithm, the JPCF algorithm has decreased the feedback load even more. This extra reduction comes from the introduction of the contention-based feedback (*c-minislot*), where not all users need to be polled as opposed to the DSMUDiv algorithm, which is a pure polling-based feedback algorithm. In terms of spectral efficiency, as shown in Section 4.2, both the JPCF algorithm and the DSMUDiv algorithm have the same spectral efficiency (here the feedback rate is not included). In (6), it has been proven that the DSMUDiv algorithm maintain the performance of the full feedback algorithm in terms of spectral efficiency, which means that the JPCF algorithm also maintain the same performance. The difference will come when you include the feedback rate ($\log_2 N$), which is defined as [bits/channel use]. From (13), we can see that as the feedback

load increases the system capacity decreases, therefore, we deduce that the JPCF algorithm has the highest system capacity compared to the other two algorithms, which can be seen in Fig. 5 and 6. As a benchmark we include the average spectral efficiency (ASE) that does not include the feedback rate. The gap between the system capacity and the ASE shrinks with the increase of number of symbols transmitted in the data transmission time slot.

Although the number of minislots of the JPCF algorithm varies from 1 to $2K - 1$, which is much greater than the DSMUDiv algorithm, where the minislots varies from 1 to K , the average guard time of the JPCF algorithm is much smaller than the guard time of the DSMUDiv algorithm as depicted in Fig. 7. Such advantage occurs at medium to high average SNR. As the average SNR decreases this advantage diminishes as seen in Fig. 8. The reason is that the JPCF algorithm has contention minislots which creates an additional delay on top of the delay created by the polling minislots. This delay increase as the probability of finding a user with channel gain exceeding the threshold value decreases, which is the case for low average SNR values. Fig. 9 shows the effect of different data transmission time durations on the system throughput. Obviously, higher values give better performance.

6. Conclusions

In this chapter we proposed a scheduling algorithm that maximizes the spectral efficiency while reducing the feedback load. The algorithm, called joint polling and contention based feedback (JPCF) algorithm, collects channel quality information of the users either in a polling form or in a contention form. Compared to the optimal (full feedback) algorithm, the JPCF algorithm has a similar spectral efficiency and a higher system capacity, which takes into account the effect of the feedback rate. Also, the JPCF algorithm shows more reduction in feedback load compared to the DSMUDiv algorithm. One drawback of the JPCF algorithm is the high delay compared to the DSMUDiv algorithm as the average SNR decreases, which affects the performance. As the average SNR increases, the delay encountered when using the JPCF algorithm drops below the delay created by using the DSMUDiv algorithm, which improves the system performance. The work presented in this paper includes analysis of the JPCF algorithm under slow Rayleigh fading assumption. Closed-form expressions of the feedback load, system capacity and scheduling delay are also presented in this paper.

Modulation Level	Switching Threshold (dB)
BPSK	$\gamma_{th}^{(1)} = 4.8$
4-QAM	$\gamma_{th}^{(2)} = 7.8$
16-QAM	$\gamma_{th}^{(3)} = 15$
64-QAM	$\gamma_{th}^{(4)} = 20$

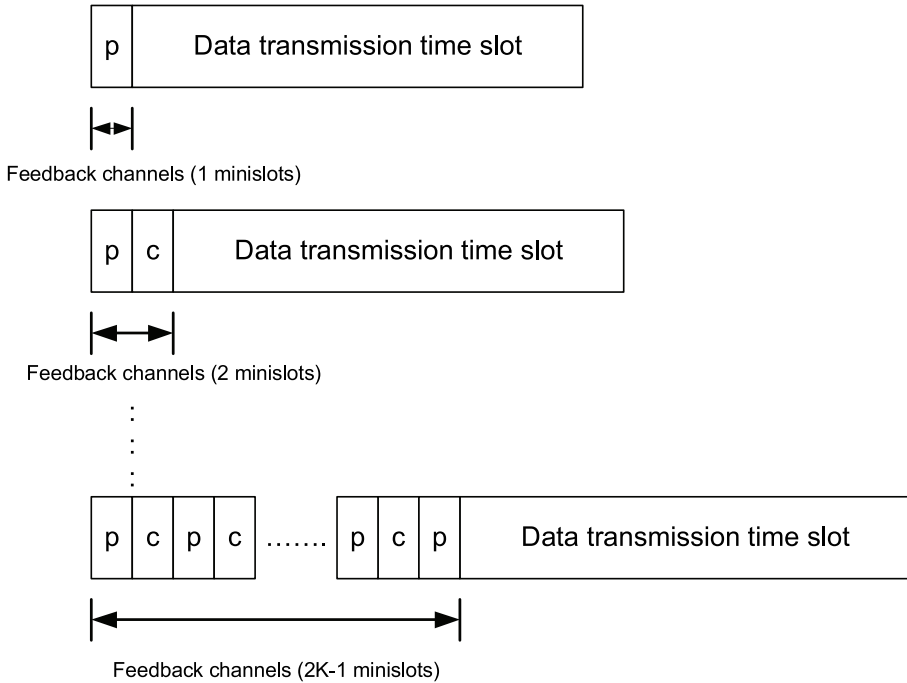
Table 1. A List of Selected Modulation Levels ($\text{BEP}_0 = 10^{-2}$)

Parameter	Value
N	4 modulations
K	30 users
τ_f	154 μ sec (based on reference (24))
T_d	5 msec (based on reference (24))

Table 2. A List of Parameters

$K=5$	$K=30$	$K=50$	$K=100$
$\rho = 0.2, \bar{\gamma} > 19$ dB	$\rho = 0.13, \bar{\gamma} > 15$ dB	$\rho = 0.12, \bar{\gamma} > 15$ dB	$\rho = 0.1, \bar{\gamma} > 14$ dB
$\rho = 1, \bar{\gamma} \leq 19$ dB	$\rho = 1, \bar{\gamma} \leq 15$ dB	$\rho = 1, \bar{\gamma} \leq 15$ dB	$\rho = 1, \bar{\gamma} \leq 14$ dB

Table 3. The Parameters Optimizations

Fig. 1. The framing structure of a TDD system. Each polling-based feedback minislot (p -minislot) is followed by a contention-based feedback minislot (c -minislot), where they carry the feedback information.

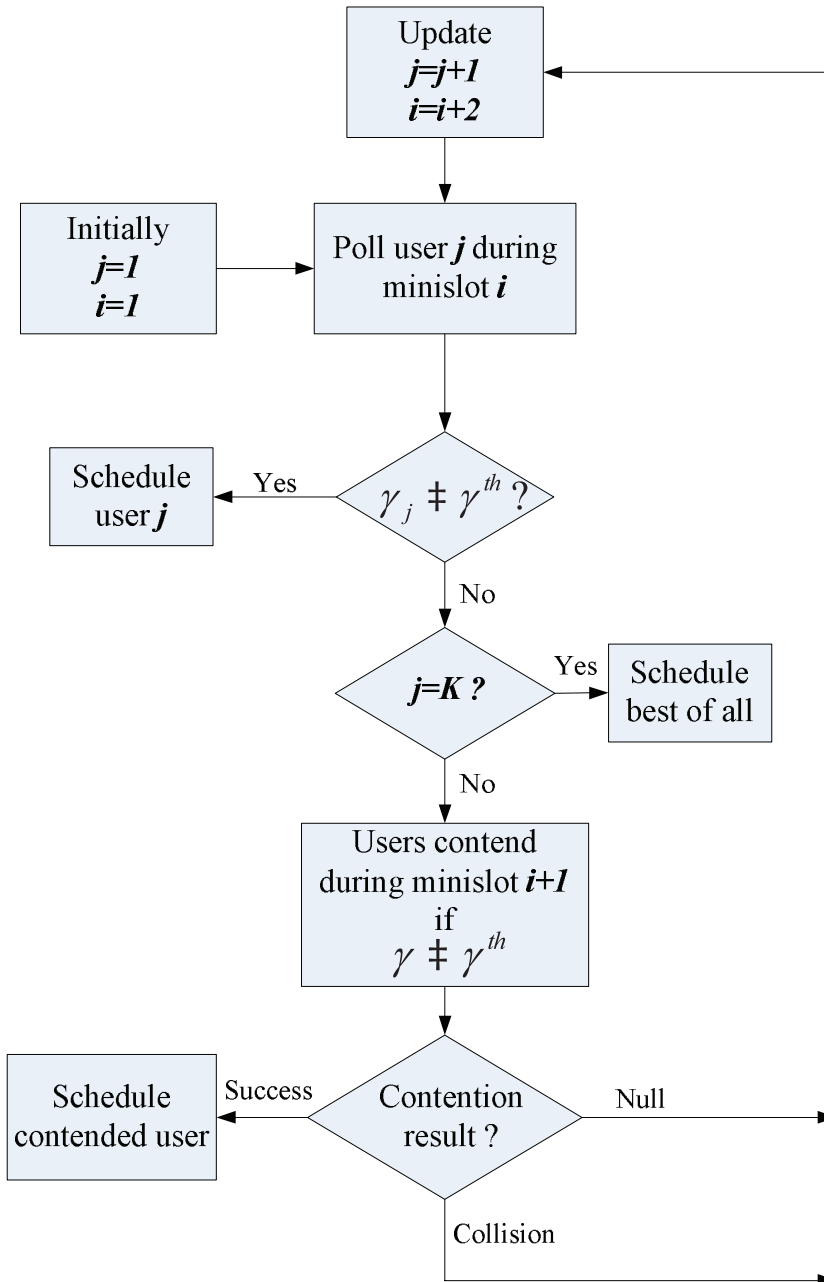


Fig. 2. A flowchart of the JPCF algorithm.

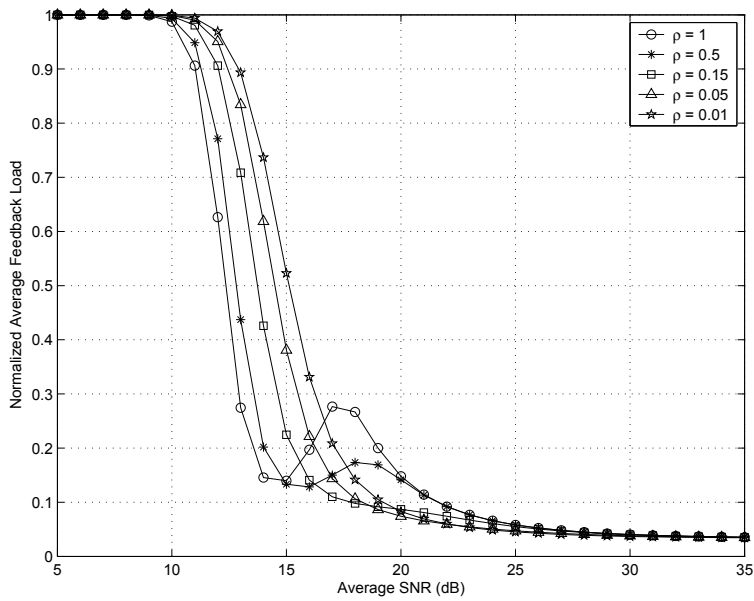


Fig. 3. Normalized average feedback load of the JPCF scheduling algorithm with different values of ρ .

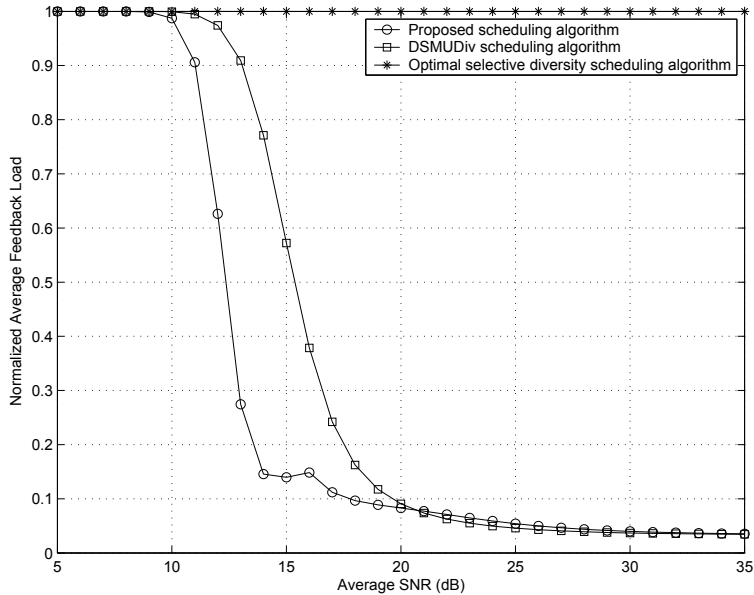


Fig. 4. Comparison of the normalized average feedback load of: (i) the proposed (JPCF) scheduling algorithm, (ii) the DSMUDiv scheduling algorithm, and (iii) the optimal (full feedback) scheduling algorithm.

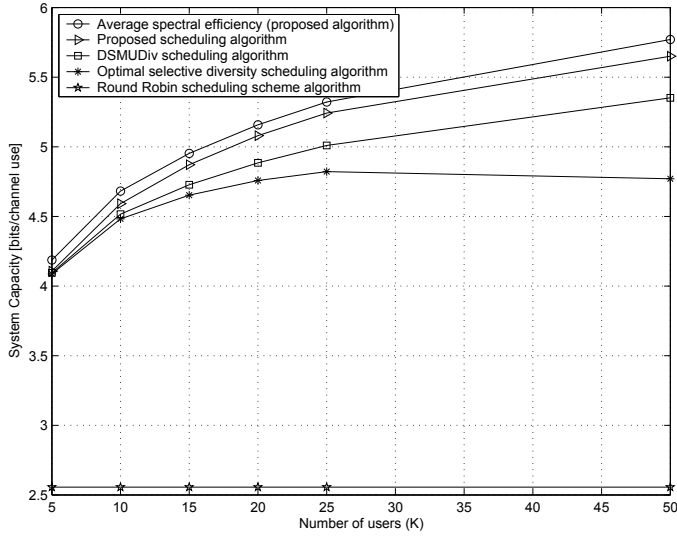


Fig. 5. System capacity of: (i) the proposed (JPCF) scheduling algorithm, (ii) the DSMUDiv scheduling algorithm, and (iii) the optimal (full feedback) scheduling algorithm. Setting $\rho = 1$, 100 symbols are transmitted, and $\bar{\gamma} = 15\text{dB}$.

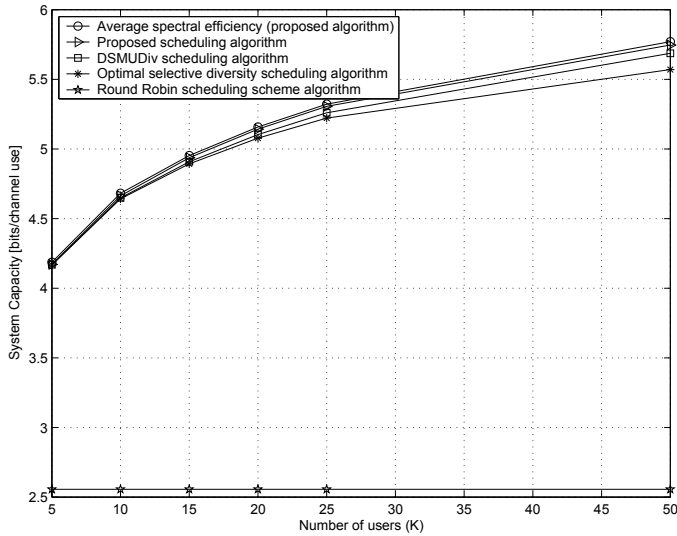


Fig. 6. System capacity of: (i) the proposed (JPCF) scheduling algorithm, (ii) the DSMUDiv scheduling algorithm, and (iii) the optimal (full feedback) scheduling algorithm. Setting $\rho = 1$, 500 symbols are transmitted, and $\bar{\gamma} = 15\text{dB}$.

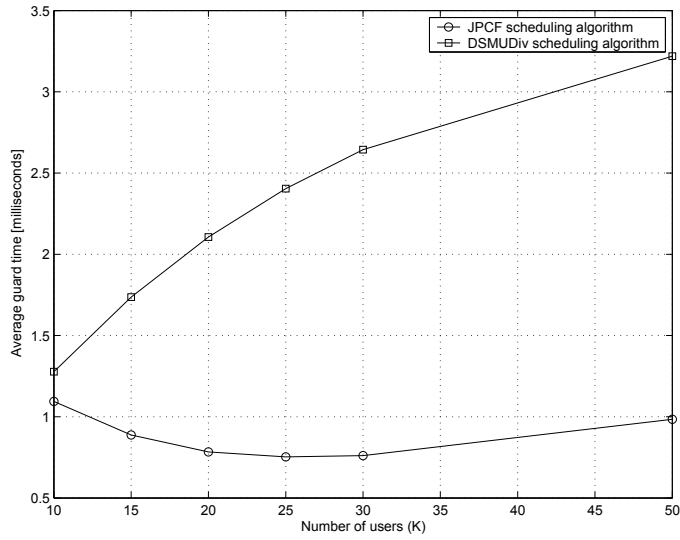


Fig. 7. Average guard time of: (i) the JPCF scheduling algorithm, and (ii) the DSMUDiv scheduling algorithm. Setting $\bar{\gamma} = 15\text{dB}$.

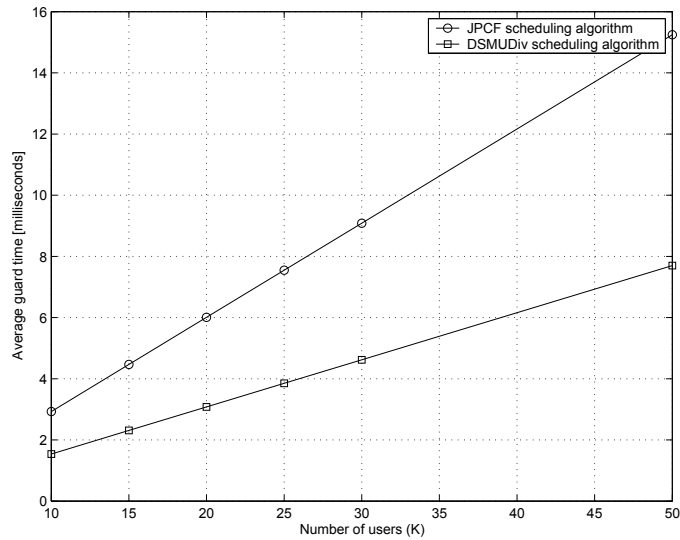


Fig. 8. Average guard time of: (i) the JPCF scheduling algorithm, and (ii) the DSMUDiv scheduling algorithm. Setting $\bar{\gamma} = 5\text{dB}$.

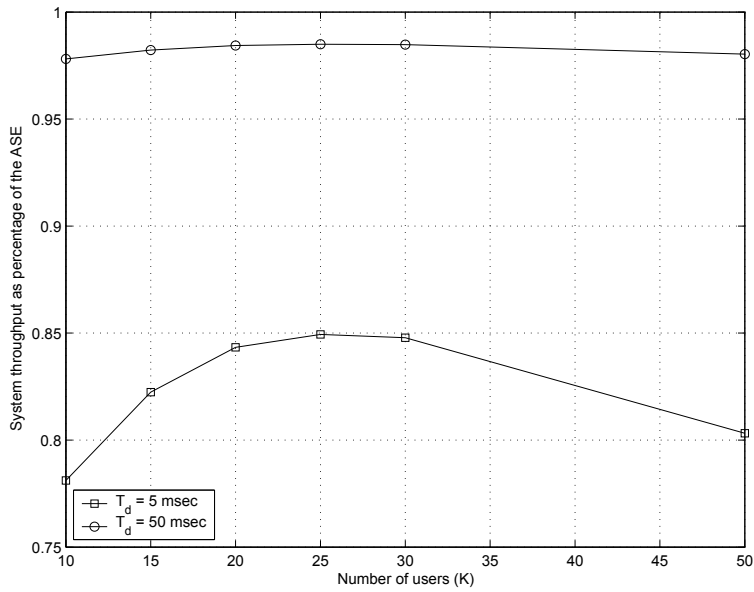


Fig. 9. System throughput as percentage of the average spectral efficiency of the JPCF scheduling algorithm. Setting $\bar{\gamma} = 15$ dB, and 1 symbol is transmitted.

7. References

- [1] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antenna," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.
- [2] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE International Communication Conference (ICC'95)*, Paris, France, June 1995, pp. 331–335.
- [3] D. N. C. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," in *Proc. IEEE International Symposium Information Theory (ISIT'97)*, Ulm, Germany, June 1997, p. 27.
- [4] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushyana, and S. Viterbi, "CDMA/HDR: A bandwidth efficient high speed wireless data service for nomadic users," *IEEE Communication Magazine*, vol. 38, no. 7, pp. 70–77, July 2000.
- [5] F. Floren, O. Edfors, and B. A. Molin, "The effect of feedback quantization on the throughput of a multiuser diversity scheme," in *Proc. IEEE Global Telecommunication Conference (GLOBECOM'03)*, San Francisco, California, December 2003, pp. 497–501.
- [6] Y. S. Al-Harathi, A. Tewfik, and M. S. Alouini, "Multiuser diversity with quantized feedback," *IEEE Transactions on Wireless Communication*, vol. 6, no. 1, pp. 330–337, January 2007.
- [7] Y. Al-Harathi, A. Tewfik, and M. S. Alouini, "Multiuser diversity-enhanced equal access with quantized feedback in multicarrier OFDM systems," in *Proc. IEEE Vehicular Technology Conference (VTC-Fall'05)*, Dallas, TX, September 2005, pp. 568–572.
- [8] M. A. Regaieg, N. Hamdi, and M. S. Alouini, "Switched-based reduced feedback OFDM multi-user opportunistic scheduling," in *Proc. IEEE International Symposium on Personal and Indoor and Mobile Radio Communication (PIMRC '05)*, Berlin, Germany, September 2005, pp. 2495–2499.
- [9] D. Gesbert and M. S. Alouini, "How much feedback is multi-user diversity really worth?," in *Proc. IEEE International Communication Conference (ICC'04)*, Paris, France, June 2004, pp. 234–238.
- [10] V. Hassel, M. S. Alouini, G. E. Øien, and D. Gesbert, "Rate-optimal multiuser scheduling with reduced feedback load and analysis of delay effects," *EURASIP Journal on Wireless Communications and Networking*, vol. 2006, no. 36424, pp. 7, 2006.
- [11] V. Hassel, M.-S. Alouini, D. Gesbert, and G. E. Øien, "Exploiting multiuser diversity using multiple feedback thresholds," in *Proc. IEEE Vehicular Technology Conference (VTC-Spring'05)*, Stockholm, Sweden, 30 May–1 June 2005, pp. 1302–1306.
- [12] B. Holter, M. S. Alouini, G. E. Øien, and H.-C. Yang, "Multiuser switched diversity transmission," in *Proc. IEEE Vehicular Technology Conference (VTC-Fall'04)*, Los Angeles, CA, September 2004, pp. 2038–2043.
- [13] Y. Xue and T. Kaiser, "Exploiting multiuser diversity with imperfect one-bit channel state feedback," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 1, pp. 183–193, January 2007.
- [14] S. Sanayei and A. Nosratinia, "Exploiting multiuser diversity with only 1-bit feedback," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC'05)*, New Orleans, LA, March 2005, pp. 978–983.
- [15] P. Svedman, S. K. Wilson, Jr. L. J. Cimini, and B. Ottersten, "A simplified opportunistic feedback and scheduling scheme for OFDM," in *Proc. IEEE Vehicular Technology Conference (VTC-Spring'04)*, Milan, Italy, May 2004, pp. 1878–1882.

- [16] Z. H. Han and Y. H. Lee, "Opportunistic scheduling with partial channel information in OFDMA/FDD systems," in *Proc. IEEE Vehicular Technology Conference (VTC-Fall'04)*, Los Angeles, CA, September 2004, pp. 511–514.
- [17] K. Bai and Junshan Zhang, "Opportunistic multichannel Aloha: distributed multiaccess control scheme for OFDMA wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 3, pp. 848–855, May 2006.
- [18] T. Tang and R. W. Heath, "Opportunistic feedback for downlink multiuser diversity," *IEEE Communications Letters*, vol. 9, no. 10, pp. 948–950, October 2005.
- [19] X. Qin and R. Berry, "Opportunistic splitting algorithms for wireless networks," in *Proc. IEEE Conference on Computer Communications (INFOCOM'04)*, Hong Kong, China, March 2004, pp. 1662–1672.
- [20] H. Koubaa, V. Hassel, and G. E. Øien, "Contention-less feedback for multiuser diversity scheduling," in *Proc. IEEE Vehicular Technology Conference (VTC-Fall'05)*, Dallas, TX, September 2005, vol. 3, pp. 1574–1578.
- [21] S. Patil and G. de Veciana, "Reducing feedback for opportunistic scheduling in wireless systems," *IEEE Transactions on Wireless Communications*, vol. 6, no. 12, pp. 4227–4232, December 2007.
- [22] M. S. Alouini and A. J. Goldsmith, "Adaptive modulation over Nakagami fading channels," *Kluwer Journal on Wireless Communications*, vol. 13, pp. 119–143, May 2000.
- [23] D. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, 1992.
- [24] IEEE Standards Department, "IEEE Std 802.11. part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," Technical Report, IEEE, NJ, September 1999.

Opportunistic Access Schemes for Multiuser OFDM Wireless Networks

Cédric Gueguen and Sébastien Baey
*Université Pierre et Marie Curie (UPMC) - Paris 6
 France*

1. Introduction

Bandwidth allocation in next generation broadband wireless networks (4G systems) is a challenging issue. The scheduling shall provide mobile multimedia transmission services with an adequate QoS. These new multimedia services with tight QoS constraints require increased system capacity together with high fairness. The past decades have witnessed intense research efforts on wireless communications. In contrast with wired communications, wireless transmissions are subject to many channel impairments such as path loss, shadowing and multipath fading. These phenomena severely affect the transmission capabilities and in turn the QoS experienced by applications, in terms of data integrity but also in terms of the supplementary delays or packet losses which appear when the effective bit rate at the physical layer is low.

Among all candidate transmission techniques for broadband transmission, Orthogonal Frequency Division Multiplexing (OFDM) has emerged as the most promising physical layer technique for its capacity to efficiently reduce the harmful effects of multipath fading. This technique is already widely implemented in most recent wireless systems like 802.11a/g or 802.16. The basic principle of OFDM for fighting the effects of multipath propagation is to subdivide the available channel bandwidth in sub-frequency bands of width inferior to the coherence bandwidth of the channel (inverse of the delay spread). The transmission of a high speed signal on a broadband frequency selective channel is then substituted with the transmission on multiple subcarriers of slow speed signals which are very resistant to intersymbol interference and subject to flat fading. This subdivision of the overall bandwidth in multiple channels provides frequency diversity which added to time and multiuser diversity may result in a very spectrally efficient system subject to an adequate scheduling.

2. Resource Allocation in a Multiuser OFDM Wireless Access Network

In this chapter, we focus on the proper allocation of radio resources among the set of mobiles situated in the coverage zone of a wireless access point in a centralized approach (Fig. 1). The packets originating from the backhaul network are buffered in the access point

which schedules the downlink transmissions. In the uplink, the mobiles signal their traffic backlog to the access point which builds the uplink resource mapping.

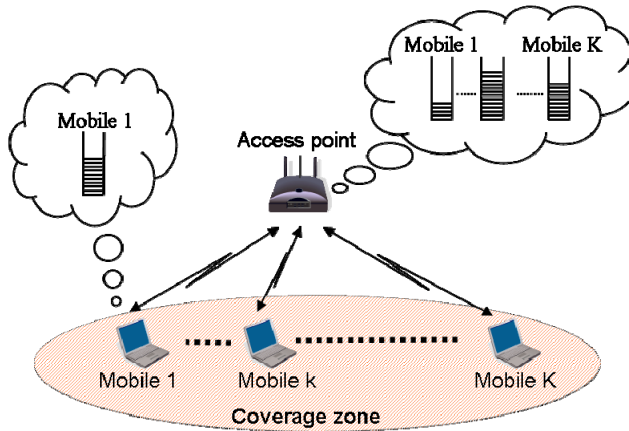


Fig. 1. Allocation of radio resources among the set of mobiles situated in the coverage zone of an access point.

The physical layer is operated using the frame structure described in Fig. 2. This structure is typical of OFDM wireless access networks like the OFDM mode of the IEEE 802.16-2004 (Hoymann, 2005). The total available bandwidth is divided in sub-frequency bands or subcarriers. The radio resource is further divided in the time domain in frames. Each frame is itself divided in time slots of constant duration. The time slot duration is an integer multiple of the OFDM symbol duration. The number of subcarriers is chosen so that the width of each sub-frequency band is inferior to the coherence bandwidth of the channel. Moreover, the frame duration is fixed to a value much smaller than the coherence time (inverse of the Doppler spread) of the channel. With these assumptions, the transmission on each subcarrier is subject to flat fading with a channel state that can be considered static during each frame.

The elementary resource unit (RU) is defined as any (subcarrier, time slot) pair. Each of these RUs may be allocated to any mobile with a specific modulation order. Transmissions performed on different RUs by different mobiles have independent channel state variations (Andrews et al., 2001). On each RU, the modulation scheme is QAM with a modulation order adapted to the channel state between the access point and the mobile to which it is allocated. This provides the flexible resource allocation framework required for opportunistic scheduling.

The system is operated using time division duplexing with four subframes: the *downlink feedback subframe*, the *downlink data subframe*, the *uplink contention subframe* and the *uplink data subframe*. The uplink and downlink data subframes are used for transmission of user data. In the downlink feedback subframe, the access point sends control information towards its mobiles. This control information is used for signalling to each mobile which RU(s) it has been allocated in the next uplink and downlink data subframes, the modulation order

selected for each of these RUs and the recommended emission power in the uplink. In the uplink contention subframe, the active mobiles send their current traffic backlog and information elements such as QoS measures and transmit power. The uplink contention subframe is also used by the mobiles for establishing their connections. This frame structure supposes a perfect time and frequency synchronization between the mobiles and the access point as described in (Van de Beek et al., 1999). Therefore, each frame starts with a long preamble used for synchronisation purposes. Additional preambles may also be used in the frame.

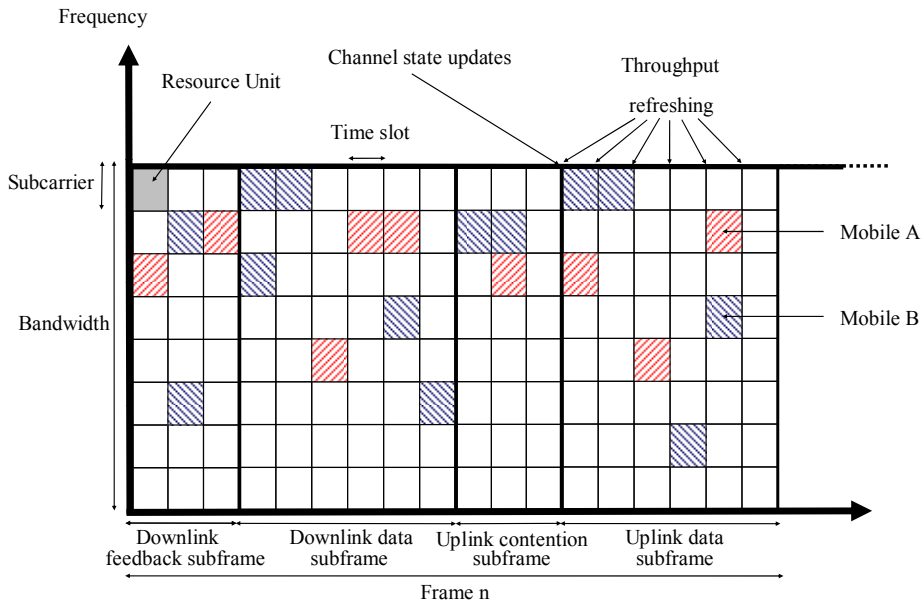


Fig. 2. Frame structure in TDD mode.

3. Scheduling Techniques in OFDM Wireless Networks

The MAC protocols currently used in wireless local area networks were originally and primarily designed in the wired local area network context. These conventional access methods like Round Robin (RR) and Random Access (RA) are not well adapted to the wireless environment and provide poor throughput. More recently intensive research efforts have been given in order to propose efficient schedulers for OFDM based networks and especially opportunistic schedulers which preferably allocate the resources to the active mobile(s) with the most favourable channel conditions at a given time. These schedulers take benefit of multiuser and frequency diversity in order to maximize the system throughput. All these schemes strongly rely on diversity for offering their good performances. Three major scheduling techniques have emerged: Maximum Signal-to-Noise Ratio (MaxSNR), Proportional Fair (PF) and recently, the Weighted Fair Opportunistic (WFO).

Note that for these schedulers, knowledge of the channel state is supposed to be available at the receiver (Li et al., 1999). The current channel attenuation on each subcarrier and for each mobile is estimated by the access node based on the SNR of the signal sent by each mobile during the uplink contention subframe. Assuming that the channel state is stable on a scale of 50 ms (Truman & Brodersen, 1997), and using a frame duration of 2 ms, the mobiles shall transmit their control information alternatively on each subcarrier so that the access node may refresh the channel state information once every 25 frames.

3.1 Maximum Signal-to-Noise Ratio

Many schemes are derived from the Maximum Signal-to-Noise Ratio (MaxSNR) technique (also known as Maximum Carrier to Interference ratio (MaxC/I)). In MaxSNR, priority is given at every scheduling event to the mobiles which have the greatest signal-to-noise ratio (SNR). It allocates the resource at a given time to the active mobile with the greatest SNR (Knopp & Humblet, 1995; Wong et al., 1999; Wang & Xiang, 2006). Denoting $m_{k,n}$ the maximum number of bits that can be transmitted on a time slot of Resource Unit n if this RU is allocated to the mobile k , MaxSNR scheduling consists in allocating the RU n to the mobile j which has the greatest $m_{k,n}$ such as:

$$j = \arg \max_k (m_{k,n}), k = 1, \dots, K, \quad (1)$$

Taking profit of multiuser and frequency diversity, MaxSNR scheduling continuously allocates the radio resource to the mobile with the best spectral efficiency. Consequently MaxSNR strongly increases the system throughput. Dynamically adapting the modulation and coding allows one to always make the most efficient use of the radio resource and come closer to the Shannon limit. However MaxSNR assumes that the user with the most favourable transmission conditions has information to transmit at the considered time instant. It does not take into account the variability of the traffic and the queuing aspects. Additionally, a negative side effect of this strategy is that the closest mobiles to the access point have disproportionate priorities over mobiles more distant since their path loss attenuation is much smaller. This results in a severe lack of fairness.

3.2 Proportional Fair

Proportional Fair (PF) algorithms have recently been proposed to incorporate a certain level of fairness while keeping the benefits of multiuser diversity (Viswanath et al., 2002; Kim et al., 2002; Anchun et al., 2003; Svedman et al., 2004; Kim et al., 2004). In PF based schemes, the basic principle is to allocate the bandwidth resources to a mobile j when its channel conditions are the most favourable with respect to its time average such as:

$$j = \arg \max_k \left(\frac{m_{k,n}}{M_{k,n}} \right), k = 1, \dots, K, \quad (2)$$

where $M_{k,n}$ is the time average of the $m_{k,n}$ values.

At a short time scale, path loss variations are negligible and channel state variations are mainly due to multipath fading, statistically similar for all mobiles. Thus, PF provides an equal sharing of the total available bandwidth among the mobiles as RR. Applying the opportunistic scheduling technique, system throughput maximization is also obtained as with MaxSNR. PF actually combines the advantages of the classical schemes and currently appears as the best bandwidth management scheme.

In PF-based schemes, fairness consists in guaranteeing an equal share of the total available bandwidth to each mobile, whatever its position or channel conditions. However, since the farther mobiles have a lower spectral efficiency than the closer ones due to pathloss, all mobiles do not all benefit of an equal average throughput despite they all obtain an equal share of bandwidth. This induces heterogeneous delays and unequal QoS. (Choi & Bahk, 2007; Gueguen & Baey (a), 2008; Holtzman, 2001) demonstrate that fairness issues persist in PF-based protocols when mobiles have unequal spatial positioning, different traffic types or different QoS targets. PF scheduling do not take into account the delay constraints and is not well adapted to multimedia services which introduce heterogeneous users, new traffic patterns with highly variable bit rates and stringent QoS requirements in terms of delay and packet loss.

3.3 Weighted Fair Opportunistic

More recently a new MAC scheduler, called "Weighted Fair Opportunistic (WFO)", has been proposed for efficient support of multimedia services in multiuser OFDM wireless networks (Gueguen & Baey (b), 2008). Built in a higher layers/MAC/PHY cross-layer approach, this scheme is designed for best profiting of the multiuser diversity taking advantage of the dynamics of the multiplexed traffics. It takes into account both the transmission conditions and the higher layer constraints (traffic patterns, QoS constraints). In order to provide an efficient support of multimedia services, WFO dynamically favours the mobiles that go through a critical period with respect to their QoS requirements using dynamic priorities.

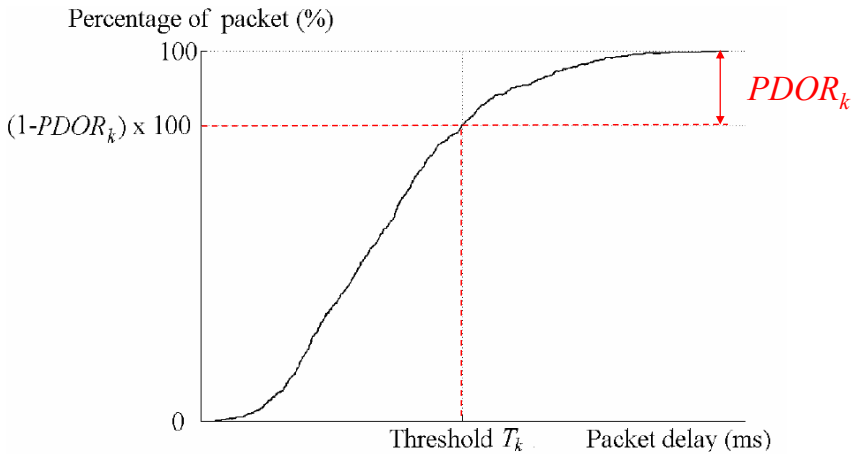


Fig. 3. Example packet delay CDF and experienced PDOR.

Evaluating if a mobile goes through a critical period should not only focus on the classical mean delay and jitter analysis. Indeed, a meaningful constraint regarding delay is the limitation of the occurrences of large values. Accordingly, (Gueguen & Baey (c), 2008) defines the concept of *delay outage* by analogy with the concept of outage used in system coverage planning. A mobile k is in delay outage (in critical period) when its packets experience a delay greater than a given threshold T_k defined by the mobile application requirements. The delay experienced by each mobile is tracked all along the lifetime of its connection. At each transmission of a packet of mobile k , the ratio of the total number of packets whose delay exceeded the threshold divided by the total number of packets transmitted since the beginning of the connection is computed. The result is called Packet Delay Outage Ratio (PDOR) of mobile k and is denoted $PDOR_k$. This measure is representative of the emergency for the mobile k to be served. Fig. 3 illustrates an example cumulative distribution of the packet delay of a mobile at a given time instant. A mobile can be considered as satisfied when, at the end of its connection, its delay constraint is met, i.e. its experienced PDOR is less than the application specific PDOR target.

In WFO scheduling, the required QoS, the experienced QoS and the transmission conditions are jointly considered in an extended cross-layer approach. The scheduling principle is to allocate a Resource Unit n to the mobile j which has the greatest WFO parameter value $WFO_{k,n}$ such as:

$$j = \arg \max_k (WFO_{k,n}), k = 1, \dots, K, \quad (3)$$

where $WFO_{k,n}$ is defined by:

$$WFO_{k,n} = m_{k,n} \times f(PDOR_k), \quad (4)$$

with f a strictly increasing polynomial function (Gueguen & Baey (d), 2008; Gueguen & Baey, 2009):

$$f(x) = 1 + \beta \times x^\alpha, \quad (5)$$

The exponent parameter α allows being sensitive and reactive to PDOR fluctuations which guarantees fairness at a short time scale. β is a normalization parameter that ensures that $f(PDOR_k)$ and $m_{k,n}$ are in the same order of magnitude.

With this scheduling, physical layer information (represented through the factor $m_{k,n}$) are used in order to take advantage of the time, frequency and multiuser diversity and maximize the system capacity. Higher layer information (represented through the factor $f(PDOR_k)$) are exploited in order to introduce dynamic priorities between flows for ensuring the same QoS level to all mobiles. With this original weighted system that introduces dynamic priorities between the flows, WFO keeps a maximum number of flows active across time but with relatively low traffic backlogs. This results in a well-balanced resource allocation. Preserving the multiuser diversity allows to continuously take a maximal benefit

of opportunistic scheduling and thus maximize the bandwidth usage efficiency. When the frequency diversity is sufficient, WFO better conceals the system capacity maximization, QoS support and fairness objectives than PF and MaxSNR schemes.

4. Performance Evaluation

In this section we compare the most acknowledged schedulers, the Round Robin (RR), the MaxSNR and the PF schemes with the most promising, the Weighted Fair Opportunistic (WFO) scheduling. Each is implemented with subcarrier by subcarrier allocation. Performance evaluation results are obtained using OPNET discrete event simulations. We focus on the main scheduling problem: maximize the system capacity while ensuring high fairness between mobiles localised at heterogeneous spatial positions in the cell.

In the simulations we assume 128 subcarriers and 5 time slots in a frame. The channel gain model on each subcarrier considers free space path loss and multipath Rayleigh fading (Parsons, 1992). We introduce a reference distance d_{ref} for which the free space attenuation equals a_{ref} . As a result the channel gain is given by:

$$a_{k,n} = a_{ref} \times \left(\frac{d_{ref}}{d_k} \right)^{3.5} \times \alpha_{k,n}^2, \quad (6)$$

where d_k is the distance to the access point of the mobile k and $a_{k,n}^2$ represents the flat fading experienced by this mobile k if it transmits or receives on subcarrier n . In the following, $a_{k,n}^2$ is Rayleigh distributed with an expectancy equal to unity.

The maximum transmit power satisfies:

$$10 \log_{10} \left(\frac{P_{max} \times T_s}{N_0} \times a_{ref} \right) = 31 \text{ dB}, \quad (7)$$

where T_s is the time duration of an OFDM symbol, P_{max} is the maximum achievable transmit power and N_0 is the single-sided power spectral density of noise. The BER target is taken equal to 10^{-3} . With this setting, the value of $m_{k,n}$ for the mobiles situated at the reference distance is 6 bits when $a_{k,n}^2$ equals unity.

We assume all mobiles run the same videoconference application. This demanding type of application generates a high volume of data with high sporadicity and requires tight delay constraints, which substantially complicate the task of the scheduler. Each mobile has only one service flow with traffic composed of an MPEG-4 video stream (Baey, 2004) and an AMR voice stream (Brady, 1969).

In these extended simulations, we analyzed the behaviour of the schedulers when mobiles occupy different geographical positions. The objective is to clearly exhibit the ability of the

opportunistic schedulers to provide fairness whatever the respective position of the mobiles. We first study a general context that includes mobility. We constitute two groups of 7 mobiles that both move straight across the cell, following the pattern described in Fig. 4 and Fig. 5. Each mobile has a speed of 3 km/h and the cell radius is taken equal to 5 km ($3 d_{ref}$). When a group of mobiles comes closer to the access point, the other group simultaneously goes farther away. Additionally, the threshold time T_k is fixed to the value 80 ms in order to consider real time constraints and the PDOR target is 5 %.

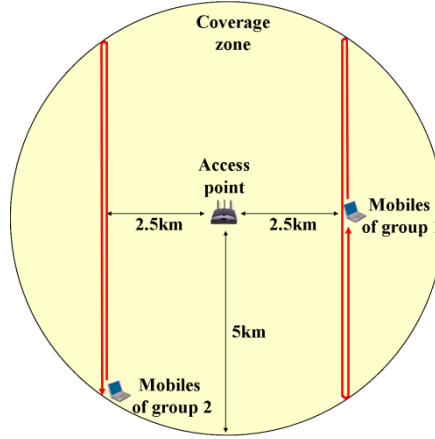


Fig. 4. Mobility pattern.

Considering the path loss, the Rayleigh fading and this mobility model, we have computed in Fig. 5 the evolution of the mean number of bits that may be transmitted per Resource Unit for each group of mobiles, averaging over all the Resource Units of a frame. This shows the impact of the mobile position on the mean $m_{k,n}$ values.

Regarding fairness, in wireless networks, it is well known that the closest mobiles to the access point generally obtain better QoS than mobiles more distant thanks to their higher spectral efficiency. Fig. 6 reports the mean PDOR experienced by each group of mobiles across the time. MaxSNR is highly unfair. Indeed, as soon as the mobiles move away from the access point, they experience high delays with a high number of packets in delay outage. PF offers better results. It brings more fairness and globally attenuates the delay peaks of the critical periods. However, we observe that WFO is the one that best smoothes these peaks. It adequately and continuously allocates the adequate priorities between the mobiles reacting to their relative movement across the cell. Providing a totally fair allocation of the bandwidth resources, the WFO scheduling smoothes the delay experienced by each mobile across time. Consequently, it further enhances the PF performances and the PDOR values are further decreased. WFO results in a very fair resource allocation that fully satisfies the delay constraints whatever the movement of the mobile.

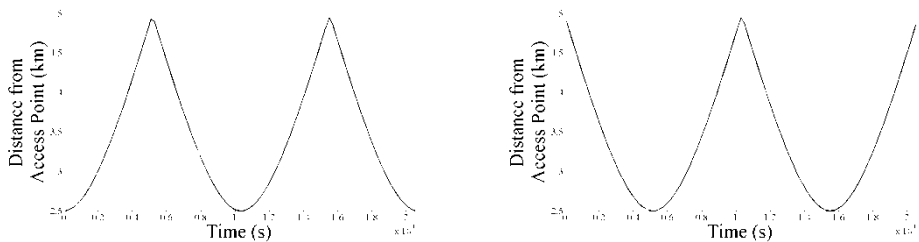


Fig. 5. Position of the mobiles across time (for mobiles of group 1 on the left and for mobiles of group 2 on the right).

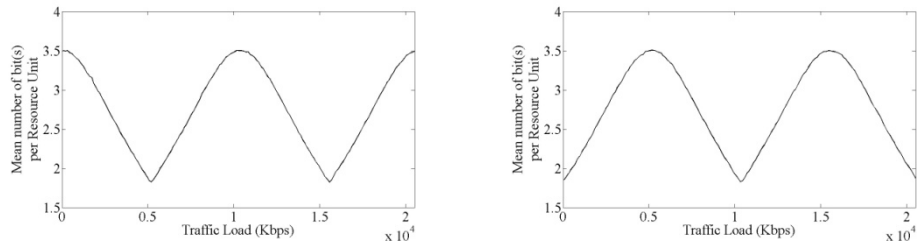


Fig. 6. Mean number of bit(s) per Resource Unit for each group of mobiles (for mobiles of group 1 on the left and for mobiles of group 2 on the right).

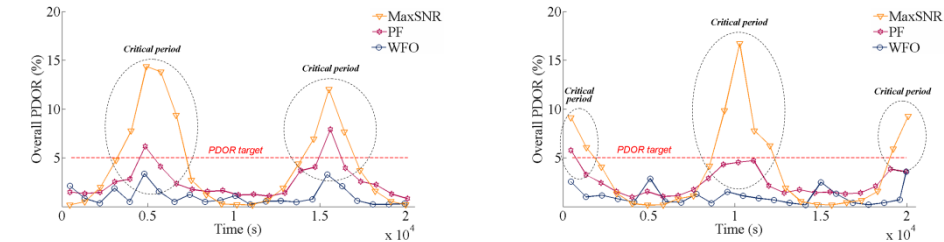


Fig. 7. PDOR fluctuation experienced by each group of mobiles (for mobiles of group 1 on the left and for mobiles of group 2 on the right).

In order to further underline the advantage of opportunistic schedulers compared to the classical Round Robin, we now study precisely the performance of the algorithms in a sub-scenario where all mobiles are static. A first half of mobiles are situated close to the access point and a second half 1.5 farther. The other parameters are identical for all the mobiles as described in Table 1. The total number of mobiles sets the traffic load.

Group	Distance d_k	Delay threshold T_k	Data rate
1	$2 d_{ref}$	80 ms	80 Kbps
2	$3 d_{ref}$	80 ms	80 Kbps

Table 1. Scenario setup with static mobiles.

First we focus on the fairness provided by each scheduler. Fig. 7a, 7b, 7c and 7d display the overall PDOR for different traffic loads considering the influence of the distance on the scheduling.

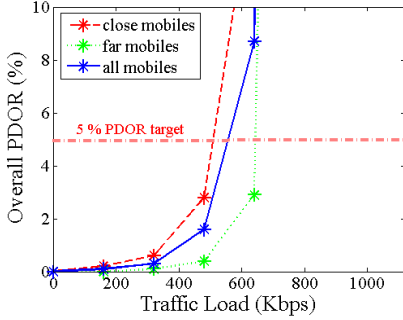


Fig. 8 a. With RR.

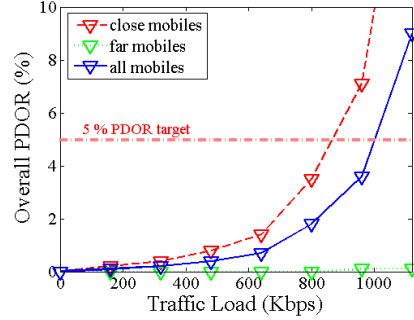


Fig. 8 b. With MaxSNR.

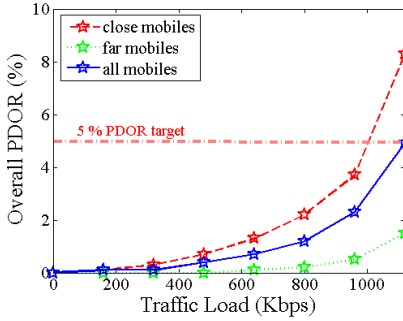


Fig. 8 c. With PF.

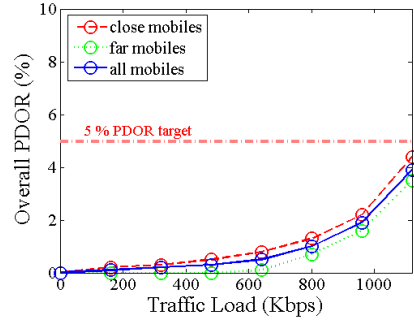


Fig. 8 d. With WFO.

Fig. 8. Measured QoS with respect to distance.

The classical RR fails to ensure the same PDOR to all mobiles. Actually, the RR fairly allocates the RUs to the mobiles without taking in consideration that far mobiles have a much lower spectral efficiency than closer ones. Moreover, the RR does not take benefit of multiuser diversity which results in a bad utilization of the bandwidth and in turn, poor system throughput. Consequently, an acceptable PDOR target of 5 % is exceeded even with relatively low traffic loads. Based on opportunistic scheduling, the three other schemes globally show better QoS performances supporting a higher traffic load. However, MaxSNR and PF still show severe fairness deficiencies. Close mobiles easily respect their delay requirement while far mobiles experience much higher delays and go past the 5 % PDOR target when the traffic load increases. In contrast, WFO provides the same QoS level to all mobiles whatever their respective position. WFO is the only one to guarantee a totally fair allocation. This allows reaching higher traffic loads with an acceptable PDOR for all mobiles. Additionally, looking at the overall PDOR for all mobiles at different traffic loads shows that, besides fairness, WFO provides a better overall QoS level as well.

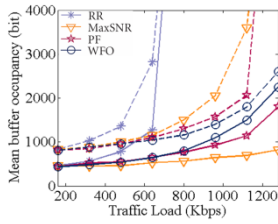


Fig. 9 a. Mean buffer occupancy for close mobiles (solid lines) and far mobiles (dashed lines).

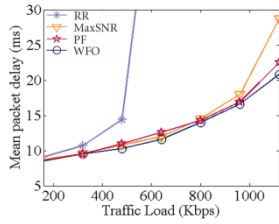


Fig. 9 b. Mean packet delay.

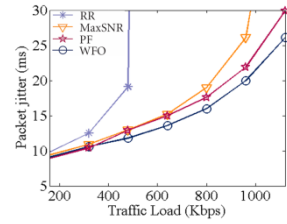


Fig. 9 c. Packet jitter.

Fig. 9. Buffer occupancy, delay and jitter.

Observing the mean buffer occupancy in Fig. 8a, WFO clearly limits the buffer occupancy to a same and reasonable value whatever the position of the mobile. This allows staying under the PDOR target for any traffic load. With its system of weights, WFO dynamically adjusts the relative priority of the flows according to their experienced delay. With this approach, sparingly delaying the closer mobiles, WFO builds on the breathing space offered by the easy respect of the delay constraints of the closer mobiles (with better spectral efficiency) for helping the farther ones. The WFO interesting performance results are corroborated in Fig. 8b and 8c where the overall values of the mean packet delay and jitter obtained using WFO are smaller.

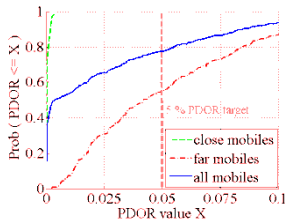


Fig. 10 a. CDF of end cycle PDOR with MaxSNR.

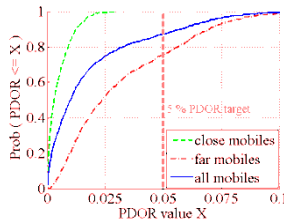


Fig. 10 b. CDF of end cycle PDOR with PF

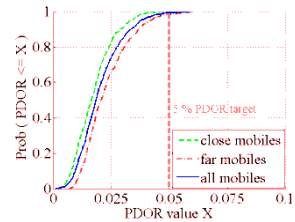


Fig. 10 c. CDF of end cycle PDOR with WFO

Fig. 10. Perceived QoS with different allocation schemes.

We then had a look at the QoS satisfaction level that each mobile perceives across the lifetime of a connection. We divided the connection of each mobile in cycles of five minutes and measured the PDOR at the end of each cycle. Fig. 9 shows the CDF of end cycle PDOR values for a traffic load of 960 Kbps, using respectively the MaxSNR, the PF and the WFO schemes (RR performances are not presented here since it is not able to support this high traffic load.). We also estimated the mobile dissatisfaction ratio. We checked if at the end of each cycle the delay constraint is met or not. We then computed the mobile dissatisfaction ratio defined as the number of times that the mobiles are not satisfied (experienced $PDOR \geq PDOR_{target}$) divided by the total number of cycles (cf. Fig. 10).

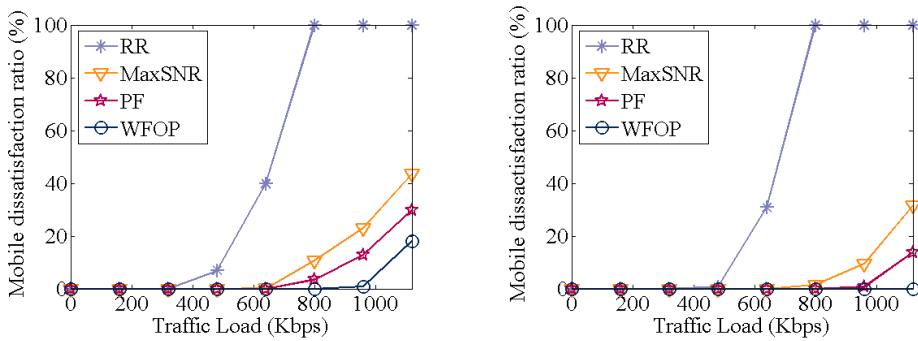


Fig. 11. Analysis of the respect of QoS constraints for different targeted QoS (Mobile dissatisfaction if $PDOR_{target} = 5\%$ on the left, if $PDOR_{target} = 10\%$ on the right).

Highly unfair, MaxSNR fully satisfies the required QoS of close mobiles at the expense of the satisfaction of far mobiles. Indeed, only 54.5 percents of these latter experience a final PDOR inferior to a PDOR target of 5 % (cf. Fig. 9a). Unnecessary priorities are given to close mobiles who easily respect their QoS constraints while more attention should be given to the farther. This inadequate priority management dramatically increases the global mobile dissatisfaction, which reaches 23 % as shown on Fig. 9a and Fig. 10 (on the left).

PF brings more fairness and allocates more priority to far mobiles. Compared to MaxSNR, PF offers a QoS support improvement with only 12.8 % of dissatisfied mobiles (cf. Fig. 9b and Fig. 10 (on the right)). Fairness is still not total since the farther mobiles have a lower spectral efficiency than the closer ones due to pathloss. All mobiles do not all benefit of an equal average throughput despite they all obtain an equal share of bandwidth. This induces heterogeneous delays and unequal QoS. This fairness improvement compared to MaxSNR indicates however that some flows can be slightly delayed to the benefit of others without significantly affecting their QoS.

WFO was built on this idea. The easy satisfaction of close mobiles (with better spectral efficiency) offers a degree of freedom which ideally should be exploited in order to help the farther ones. WFO allocates to each mobile the accurate share of bandwidth required for the satisfaction of its QoS constraints, whatever its position. With WFO, only 0.8 percents of the mobiles are dissatisfied (cf. Fig. 9c and Fig. 10 (on the left)). Additionally, compared to Fig. 9a and Fig. 9b, Fig. 9c exhibits superimposed curves, which prove the WFO high fairness, included at short term.

Fig. 10 shows that WFO brings the largest level of satisfaction. Indeed, for a tight PDOR target of 5 % (see on the left), the dissatisfaction ratio with a high traffic load of 1120 Kbps is equal to 18 % with WFO versus 29.7 % with PF, the best of the other scheduling schemes. If we set the PDOR target to 10 %, the dissatisfaction ratio with a high traffic load of 1120 Kbps is 0 % with WFO versus 13.8 % with the best of the other scheduling schemes (PF).

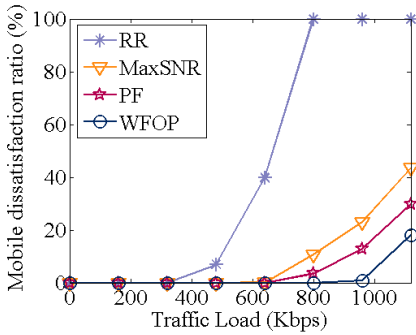


Fig. 12 a. Spectral efficiency.

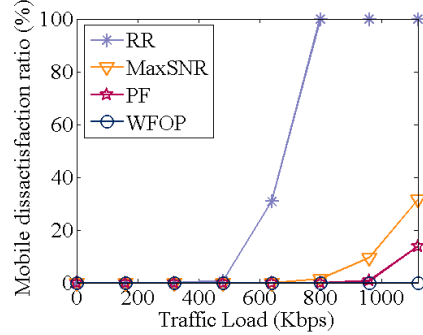


Fig. 12 b. Multiuser diversity

We finally studied the system capacity offered by the four scheduling algorithms. Fig. 11a shows the average number of bits carried on a used subcarrier by each tested scheduler under various traffic loads. As expected, the non opportunistic Round Robin scheduling provides a constant spectral efficiency, i.e. an equal bit rate per subcarrier whatever the traffic load since it does not take advantage of the multiuser diversity. The three other tested schedulers show better results. In contrast with RR, with the opportunistic schedulers (MaxSNR, PF, WFO), we observe an interesting inflection of the spectral efficiency curve when the traffic load increases. The joint analysis of Fig. 11a and Fig. 11b shows that the spectral efficiency of opportunistic scheduling is an increasing function of the number of active mobiles, thanks to the exploitation of this supplementary multiuser diversity. Consequently MaxSNR, PF and WFO increase their spectral efficiency with the traffic load and the system capacity is highly extended compared to networks which use classical scheduling algorithms. With these three schedulers, all mobiles are served even at the highest traffic load of 1280 Kbps.

The performance of the four schedulers can be further qualified by computing the theoretical maximal system throughput. Considering the Rayleigh distribution, it can be noticed that $a^2_{k,n}$ is greater or equal to 8 with a probability of only 0.002. In these ideal situations, close mobiles can transmit/receive 6 bits per RU while far mobiles may transmit/receive 4 bits per RU. If the scheduler always allocated the RUs to the mobiles in these ideal situations, an overall efficiency of 5 bits per RU would be obtained which yields a theoretical maximal system throughput of 1600 Kbps. Comparing this value to the highest traffic load in Fig. 11a (1280 Kbps) further demonstrates the good efficiency obtained with the opportunistic schedulers that nearly always serve the mobiles when their channel conditions are very good. This result also shows that the WFO scheduling has slightly better performances than the two other opportunistic schedulers. Keeping more mobiles active (cf. Fig.11b) but with a relatively lower traffic backlog (cf. Fig.8a), the WFO scheme preserves multiuser diversity and takes more advantage of it obtaining a slightly higher bit rate per subcarrier (cf. Fig. 11a).

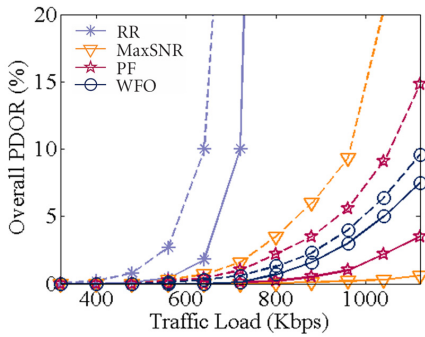


Fig. 13a. Measured QoS for close mobiles (solid lines) and far mobiles (dashed lines).

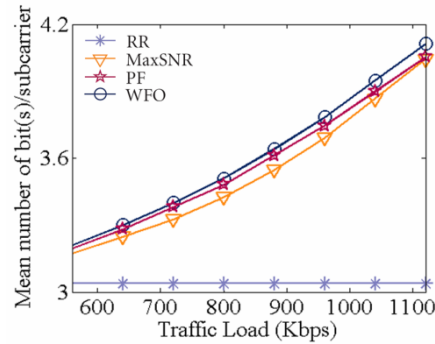


Fig. 13b. Spectral efficiency.

Fig. 13. Performances of schedulers with fixed multiuser diversity.

In the results described so far, the traffic load was varied by increasing or decreasing the number of mobiles in the system, which modified the multiuser diversity. This exhibited the opportunistic behaviour of the schedulers and especially their ability to take advantage of the multiuser diversity brought with the increase of the number of mobiles. We also studied the ability of each scheduler to take profit of the multiuser diversity brought by a given number of users. In Fig.12, we provide complementary results obtained in a context where the traffic load variation is done through just increasing the mobile bit rate requirement and keeping a constant number of users (10 mobiles). The results in Fig. 12a show that, as previously, WFO outperforms the other scheduling schemes. With its weighted algorithm, WFO dynamically adjusts the priorities of the mobiles and ensures a completely fair allocation. WFO is the only one which allows reaching higher traffic loads with an acceptable PDOR for all mobiles. Additionally, even if the traffic load increases without variation in the number of mobiles, WFO keeps more mobiles active across the time than the other schemes and takes better advantage of the multiuser diversity. The analysis of Fig. 12b confirms that WFO maximizes the average bit rate per subcarrier.

5. Conclusion

Opportunistic schedulers take benefit of multiuser and frequency diversity. They preferably allocate the resources to the active mobile(s) with the most favourable channel conditions at a given time. This maximizes the system throughput of OFDM wireless networks. Three major algorithms have emerged: MaxSNR, PF and more recently WFO. However, in spite of their high performances in terms of system throughput maximization, both MaxSNR and PF suffer of severe fairness deficiencies owing to unequal spatial positioning of the mobiles. This issue is resolved with WFO which appears as the best current opportunistic scheduler. WFO jointly considers the transmission conditions, the currently measured/experienced QoS and the QoS targets of the mobiles in the bandwidth allocation process. With an original weighted system that introduces dynamic priorities between the flows, it dynamically favors the flows that go through a critical period and always attributes the adequate priorities for improved QoS support. Keeping a maximum number of flows active

across time but with relatively low traffic backlogs, WFO is designed for best profiting of the multi-user diversity taking advantage of the dynamics of the multiplexed traffics. Preserving the multiuser diversity, WFO takes a maximal benefit of the opportunistic scheduling technique and maximizes the system capacity. Additionally, this also achieves a time uniform fair allocation of the resource units to the flows ensuring short term fairness. This higher layers/MAC/PHY cross-layer approach better conceals the system capacity maximization, fairness objectives and the full support of multimedia services with adequate QoS.

6. References

- Hoymann, C. (2005). Analysis and performance evaluation of the OFDM-based metropolitan area network IEEE 802.16. *Computer Networks*, Vol. 49, No. 3, pp. 341-363, ISSN: 1389-1286
- Andrews, M.; Kumaran, K.; Ramanan, K. Stolvar, A. & whiting P. (2001). Providing quality of service over a shared wireless link. *IEEE Communications Magazine*, Vol. 39, No.2, pp. 150-154, ISSN: 0163-6804
- Van de Beek, J.-J. ; Borjesson, P.O. ; Boucheret, M.-J ; Landstrom, D. ; Arenas, J.-M. Odling, P.; Ostberg, C.; Wahlgvist, M. & Wilson, S.K. (1999). A time and frequency synchronization scheme for multiuser OFDM. *IEEE J.Sel. Areas Commun*, Vol.17, No.11, pp 1900-1914, ISSN: 0733-8716
- Li, Y.G; Seshadri, N. & Ariyavisitakul, S. (1999). Channel estimation for ofdm systems with transmitter diversity in mobile wireless channels. *IEEE J.Sel. Areas Commun*, Vol.17, No.3, pp 461-471, ISSN: 0733-8716
- Truman, T.E. & Brodersen, R.W (1997). A measurement-based characterization of the time variation of an indoor wireless channel, *proceedings of Int. Universal Personal Communications Record (ICUPC)*, pp. 25-32, ISBN: 0-7803-3777-8, San Diego, CA, USA, October 1997
- Knopp, R. & Humblet, P. (1995). Choi, J. (1996). Information capacity and power control in single-cell multiuser communications, *proceedings of IEEE Conference on Communications (ICC)*, pp 331-335, ISBN: 0-7803-2486-2, Seattle, WA, USA, june 1995
- Wong, C.Y.; Cheng, R.S.; Lataief, K.B. & Murch, R.D. (1999). Multiuser OFDM with adaptative subcarrier, bit and power allocation. *IEEE J.Sel. Areas Commun*, Vol.17, No.10, pp 1747-1758, ISSN: 0733-8716
- Wang, X. & Xiang, Y. (2006). An OFDM-TDMA/SA MAC protocol with QoS constaints for broadband wireless LANs. *ACM/Springer Wireless Networks*, Vol. 12, No. 2, pp 159-170, ISSN: 1022-0038
- Viswanath, P.; Tse, D.N.C. & Laroia, R. (2002). Opportunistic beamforming using dumb antennas. *IEEE Transactions on Information Theory*, Vol. 48, No.6, pp. 1277-1294, ISSN: 0018-9448
- Kim, H.; Kim, K.; Han, Y. & Lee, J. (2002). An efficient scheduling algorithm for QoS in wireless packet data transmission, *proceedings of IEEE Int. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp 2244-2248, ISBN: 0-7803-7589-0, Lisboa, Portugal, September 2002

- Anchun, W.; Liang, X.; Xjiin, S.X. & Yan, Y. (2003). Dynamic resource management in the fourth generation wireless systems, *proceedings of IEEE Int. Conference on Communication Technology (ICCT)*, pp 1095-1098, ISBN: 7-5635-0686-1, Beijing, China, April 2003
- Svedman, P.; Wilson, K. & Ottersen, B. (2004). A QoS-aware proportional fair scheduler for opportunistic OFDM, *proceedings of IEEE Int. Vehicular Technology Conference (VTC)*, pp 558-562, ISBN: 0-7803-8521-7, Los angeles, CA, USA, September 2004
- Kim, H.; Kim, K.; Han, Y. & Yun, S. (2004). A proportional fair scheduling for multicarrier transmission systems, *proceedings of IEEE Int. Vehicular Technology Conference (VTC)*, pp. 409-413, ISBN: 0-7803-8521-7, Los angeles, CA, USA, September 2004
- Choi, J.-G. & Bahk, S. (2007). Cell-throughput analysis of the proportional fair scheduler in the single-cell environment. *IEEE Transactions on Vehicular Technology*, vol. 56, No.2, pp. 766-778, ISSN: 0018-9545
- Gueguen, C. & Baey, S. (2008). Compensated proportional fair scheduling in multiuser OFDM wireless networks, *proceedings of IEEE Wireless and Mobile Computing, Networking and Communications (WIMOB)*, pp119-125, ISBN: 978-0-7695-3393-3, Avignon, France, October 2008
- Holtzman, J. (2001). Asymptotic analysis of proportional fair algorithm, *proceedings of IEEE Int. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 33-37, ISBN: 0-7803-7244-1, San Diego, CA, USA, October 2001
- Gueguen, C. & Baey, S. (2008). Weighted fair opportunistic scheduling for multimedia QoS support in multiuser OFDM wireless networks, *proceedings of IEEE Int. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1-6, ISBN: 978-1-4244-2643-0, Cannes, France, September 2008
- Gueguen, C. & Baey, S. (2008). An efficient and fair scheduling scheme for multiuser OFDM wireless networks, *proceedings of IEEE Int. Wireless Communications and Networking Conference (WCNC)*, pp. 1610-1615, ISBN: 978-1-4244-1997-5, Las Vegas, NV, USA, April 2008
- Gueguen, C. & Baey, S. (2008). Scheduling in OFDM wireless networks without tradeoff between fairness and throughput, *proceedings of IEEE Int. Vehicular Technology Conference (VTC)*, pp. 1-5, ISBN: 978-1-4244-1721-6, Calgary, Canada, September 2008
- Gueguen, C. & Baey, S. (2009). A Fair Opportunistic Access Scheme for Multiuser OFDM Wireless Networks. *EURASIP Journal on Wireless Communications and Networking. Special issue on "Fairness in Radio Resource Management for Wireless Networks"*. Volume 2009 (2009), Article ID 726495, pp. 70-83
- Parsons, J.D (1992). *The Mobile Radio Propagation Channel*, Wiley, ISBN: 978-0-471-98857-1
- Baey, S. (2004). Modeling MPEG4 video traffic based on a customization of the DBMAP, *proceedings of Int. Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*, pp. 705-714, ISBN: 1-56555-284-9, San Jose, California, USA, July 2004
- Brady, P. (1969). A model for generating on-off speech patterns in two-conversation. *Bell System Technical Journal*, vol. 48, No.1, pp. 2445-2472

Bidirectional Cooperative Relaying

Prabhat Kumar Upadhyay and Shankar Prakriya
*Indian Institute of Technology Delhi
India*

1. Introduction

Modern radio communication systems aim to enhance throughput and reliability in wireless networks with limited resources. Wireless mobile communication over a radio channel is limited by multipath, fading, path loss, shadowing, and interference. Spatial diversity techniques are widely adopted to combat fading and other channel impairments. Cooperative communications have been recently developed to harness spatial diversity even with single-antenna terminals. The distributed terminals cooperate by relaying each other's message in order to realize a virtual antenna array and achieve cooperative diversity. Cooperative relaying has become a promising technique for enhancing coverage, reliability and throughput of wireless networks with stringent spectrum and power constraints. They have found applications in wireless cellular, ad hoc/sensor networks, WiFi/WiMAX, etc.

Dual-hop three-terminal channels wherein a relay terminal assists in the communication between source and destination terminals through some cooperation protocol are of particular interest. Relaying can be performed in either full-duplex or half-duplex mode. Full-duplex relaying allows the radios to receive and transmit simultaneously using the same frequency channel and hence achieves higher spectral efficiency. However, the large difference in power levels of the transmit and receive signals (typically 100-150 dB) makes its implementation practically difficult. In half-duplex mode, the reception and transmission at the radios are performed in time/frequency/code division orthogonal channels. Half-duplex systems are therefore practically feasible. The major drawback of half-duplex relaying is a substantial loss in spectral efficiency. This is because half of the channel resources are allocated to the relay for cooperation, which reduces the overall data rate.

While much research has focused on exploiting cooperative diversity, little effort has been directed towards improving spectral efficiency under half-duplex constraints. The authors in (Rankov & Wittneben, 2007) propose a new two-phase two-way relaying protocol where a bidirectional connection between two terminals is established with one half-duplex relay. Under this scheme, two connections are realized in the same physical channel, thereby improving the spectral efficiency. Referred to as the Two-Way Relay Channel (TWRC) in literature, this pragmatic approach has become a focus of extensive recent research. An example of a TWRC is the downlink and uplink in wireless mobile networks whereby both the base station and the mobile station need to communicate via an assisting relay station due to the lack of a reliable direct link. This is advantageous in the case when the mobile is highly shadowed or near the cell edge. It is important to note that even when a direct link of sufficient quality is available, it cannot be utilized in two-phase TWRC because otherwise both terminals need to transmit and receive simultaneously in the same phase.

In a separate remarkable development, the emergence of network coding (Ahlsweide et al., 2000) has changed the way communication networks are designed. Network coding allows the intermediate nodes to combine and code the data from multiple sources in order to enhance the overall network throughput. Originally proposed for wired communication networks, there has recently been much interest in applying network coding to wireless relay networks (Hao et al., 2007). In view of the spectral efficiency loss due to half-duplex mode, a coded bidirectional relaying scheme with three transmission phases has been proposed independently in (Wu et al., 2005) and (Larsson et al., 2006). The authors in (Kim et al., 2008) compared and analyzed the performance of various half-duplex bidirectional relaying protocols. The idea of network coding has been further exploited for the bidirectional cooperation in (Hausl & Hagenauer, 2006); (Baik & Chung, 2008); (Cui et al., 2008a); (Cui et al., 2008b). It has been shown in (Katti et al., 2007b) that wireless two-way relaying coupled with network coding achieves higher data rates.

Two-way or bidirectional relaying is flexible to allow various physical-layer transmission techniques. A lot of research is in progress on topics like TWRC capacity region or achievable rate region (Oechtering et al., 2008), channel estimation (Zhao et al., 2008); (Gao et al., 2008), multi-hop relaying (Vaze & Heath Jr., 2008), resource allocation (Agustin et al., 2008), distributed space-time coding (Cui et al., 2008c), distributed relay selection (Ding et al., 2009) and the like, using various physical layer signalling techniques, OFDM for example (Ho et al., 2008); (Jitvanichphaibool et al., 2008). Also, the bidirectional relaying scheme has been extended to the multi-user scenario (Chen & Yener, 2008); (Eslami & Wittneben, 2008). Multiple-Input Multiple-Output (MIMO) bidirectional relaying (Unger & Klein, 2007); (Gunduz et al., 2008) is a hot research area and is often envisioned to further improve the link reliability and bidirectional throughput of wireless systems.

The aim of this chapter is to present, in a unified fashion, the state-of-the-art in this new area of bidirectional cooperative communication, to elaborate on the recent analytical findings and their significance, to support them with various simulation results, and to discuss future areas of research.

2. Cooperative Communications

Cooperative communication systems seek to enhance the link capacity and transmission reliability through cooperation between distributed radios. They exploit the broadcasting nature of the wireless medium and allow single-antenna terminals to cooperate through relaying. The conventional form of cooperation is multi-hopping, where a source communicates with a destination via a series of dedicated relays. It is mainly used to combat signal attenuation in long-range communication and it does not provide any diversity advantages. The key issue in cooperative communications is resource sharing among network nodes. A three-terminal network acts as a fundamental unit in cooperative communication and has been widely studied in the literature.

The three-terminal relay channel model, introduced in (Meulen, 1971), comprises a source T_1 , a destination T_2 , and a dedicated relay R (as shown in Figure 1). The relay aids in communicating information from source to destination without actually being an information source or sink. It was assumed that all nodes operate in the full-duplex mode, so the system can be viewed as a Broadcast Channel (BC) at the source, and a Multiple Access Channel (MAC) at the destination. The upper and lower bounds on the capacity of the non-faded relay channel, derived in (Meulen, 1971), were improved significantly in (Cover & Gamal, 1979). Later, models with multiple relays have been investigated in cooperation literature. However, the

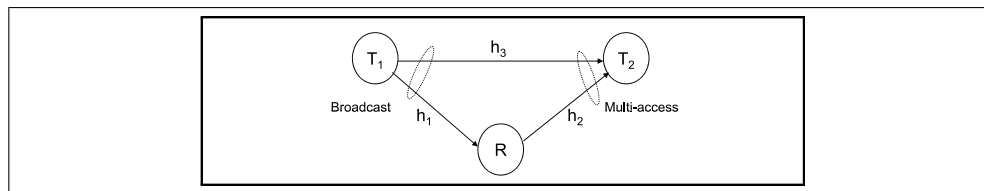


Fig. 1. The wireless relay channel.

recent developments are motivated by the user cooperation (Sendonaris et al., 2003) and cooperative diversity (Laneman et al., 2004) in a fading channel. The authors in (Sendonaris et al., 2003) introduced user cooperation by allowing the relay to transmit its own independent information. Cooperative diversity introduced in (Laneman et al., 2004) is realized by relaying and user cooperation. They proposed different cooperative diversity protocols and analyzed their performance in terms of outage probability. The terms Amplify-&-Forward (AF) and Decode-&-Forward (DF) were introduced in their work.

2.1 Cooperative Relaying Protocols

Consider a three-terminal wireless network as shown in Figure 1 in which terminal T_1 wants to transmit data to terminal T_2 with the help of a relay terminal R . In view of cellular network, T_1 and R might be mobile stations and T_2 might be a base station. Under cooperative relaying strategy, T_1 and a suitable R can share their resources, such as power and bandwidth, to transmit the information of T_1 . This cooperation might provide diversity because, even if the direct link between T_1 and T_2 is severely faded, the information might be successfully transmitted via R . It is assumed that the relay node operates in half-duplex mode and has no

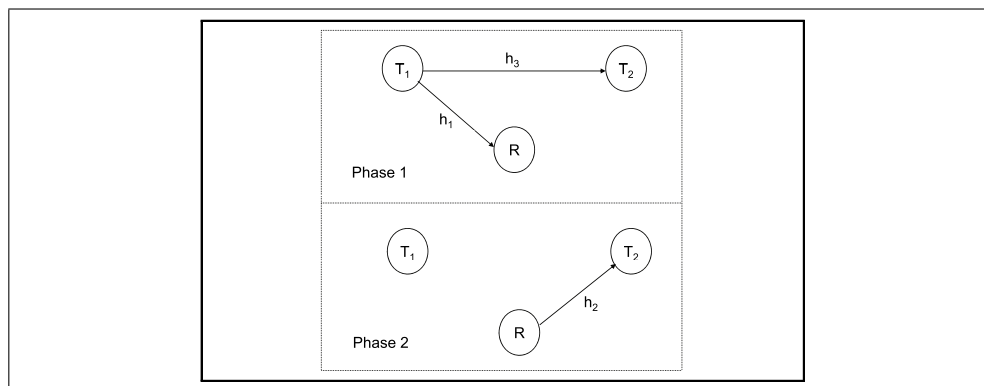


Fig. 2. Orthogonal relay transmission.

message of its own to transmit. Figure 2 illustrates the channel allocation for time-division approach with two orthogonal phases to ensure half-duplex operation. In phase 1, T_1 sends information to its destination T_2 and the information is also received by the relay R at the same time. In phase 2, the relay R can help the source T_1 by forwarding or retransmitting the information to the destination. However, there is 50% loss in spectral efficiency because of transmission in two phases.

The source and relay nodes can share their resources on the basis of some cooperation strategies to achieve the highest throughput possible for any given coding scheme. Based on different signal processing schemes employed at the relays, the cooperative relaying methods are classified into fixed relaying and adaptive relaying (Laneman et al., 2004). For fixed relaying, the relay can amplify its received signal subject to its power constraint, or decode, re-encode, and then retransmit the messages, referred to (respectively) as Amplify-&-Forward (AF) or Decode-&-Forward (DF). This scheme has the advantage of easy implementation, but the disadvantage of low spectral efficiency. This is because half of the channel resources are allocated to the relay for transmission. Adaptive relaying schemes build upon fixed relaying and adapt based upon Channel State Information (CSI) between cooperating terminals (selective relaying) or upon limited feedback from the destination (incremental relaying). Selective relaying allows transmitting terminals to select a suitable cooperative or non-cooperative action based on the measured SNR between them. If the received SNR at the relay exceeds a certain threshold, the relay performs DF operation on the message. Otherwise, if the channel between T_1 and R has severe fading such that SNR falls below the threshold, the relay idles. Incremental relaying improves upon the spectral efficiency of both fixed and selective relaying by exploiting limited feedback from the destination and relaying only when direct link from source to destination has an SNR below a threshold.

2.2 Outage Analysis and Diversity Gain

When the channel is time-varying, the channel capacity has different notions depending on the different fading states. Ergodic (Shannon) capacity is an appropriate capacity metric for channels that vary quickly, or where the channel is ergodic over the time period of interest. It can be evaluated by averaging the mutual information over all possible channel realizations. An alternate outage capacity notion is suitable for applications where the data rate cannot depend on channel variations (except in outage states, where no data are transmitted). It is a measure of data rate that can be supported by a system with a certain error probability. To investigate the diversity gain, the performance of relaying protocols is characterised in terms of outage probability. Assume frequency flat slow fading channel with CSI knowledge at the receivers only. Perfect synchronization among the terminals is also assumed. Considering a baseband-equivalent discrete-time channel model, the transmissions in time slot k can be expressed as

$$y_r[k] = h_1[k]x_1[k] + n_1[k] \quad (1)$$

$$y_2[k] = h_3[k]x_1[k] + n_3[k] \quad (2)$$

where $x_1[k]$ is the transmitted signal from T_1 , $y_r[k]$ and $y_2[k]$ are the received signals at the relay and T_2 respectively, h_i captures the effects of path-loss, shadowing and frequency nonselective fading, $n_i \sim \mathcal{CN}(0, \sigma^2)$ is the Additive White Gaussian Noise (AWGN) which captures the effects of receiver noise and other forms of interference in the system, where $i \in \{1, 2, 3\}$. Throughout the chapter, we use $h_i[k]$ and h_i interchangeably for brevity. The relay processes $y_r[k]$ and relays the information by transmitting $x_r[k]$. The signal received at T_2 in time slot $k+1$

$$y_2[k+1] = h_2[k+1]x_r[k] + n_2[k+1]. \quad (3)$$

As a function of the fading coefficients h_i (modeled as zero-mean, independent, circularly symmetric complex Gaussian random variables with variances $\sigma_{h_i}^2$), the mutual information

for a protocol is a random variable I . For a target rate \mathcal{R} , $I < \mathcal{R}$ denotes the outage event and $\Pr[I < \mathcal{R}]$ denotes the outage probability (Laneman et al., 2004). The maximum average mutual information between input and output in direct transmission, achieved by independent identically distributed (i.i.d.) zero-mean, circularly symmetric complex Gaussian inputs, is given by

$$I_D = \log \left(1 + \gamma |h_3|^2 \right) \quad (4)$$

where $\gamma = P_1/\sigma^2$ is defined as SNR without fading and P_1 is the average transmit power of terminal T_1 . For Rayleigh fading, $|h_3|^2$ is exponentially distributed with parameter $1/\sigma_{h3}^2$, the outage probability derived in (Laneman et al., 2004) is given as

$$\Pr[I_D < \mathcal{R}] = 1 - \exp \left(-\frac{2^{\mathcal{R}} - 1}{\gamma \sigma_{h3}^2} \right) \sim \frac{1}{\gamma} \frac{2^{\mathcal{R}} - 1}{\sigma_{h3}^2}. \quad (5)$$

The direct transmission does not achieve any diversity gain as is obvious from γ^{-1} dependence of outage probability in Equation (5).

2.3 AF Relaying

In this protocol, the relay amplifies the received signal in the first time slot according to its available average transmit power and forwards a scaled signal in the second time slot to the destination terminal. To remain within its power constraint, an amplifying relay must use gain

$$g[k] \leq \sqrt{\frac{P_r}{P_1 |h_1|^2 + \sigma^2}} \quad (6)$$

which is inversely proportional to the received power. Thus the relay transmits the signal $x_r[k+1] = g[k]y_r[k]$ with the power P_r in the second time slot. This scheme can be viewed as repetitive coding from two distributed transmitters T_1 and R , except that the relay R amplifies the noise in its received signal. The destination T_2 can decode its received signal y_2 by suitably combining the signals from the two time slots. This protocol produces an equivalent one-input two-output complex Gaussian noise channel with different noise levels in the outputs. The SNR received at the destination is the sum of the SNRs from T_1 and R links. The maximum average mutual information between the input and the two outputs, achieved by i.i.d. complex Gaussian inputs, is given by

$$I_{AF} = \frac{1}{2} \log \left[1 + \gamma |h_3|^2 + \frac{\gamma |h_2 g h_1|^2}{(1 + |h_2 g|^2)} \right]. \quad (7)$$

Note that g is a function of h_1 . Notations $g[k]$ and g are used interchangeably throughout for brevity. The outage probability can be approximated at high SNR (Laneman et al., 2004) as

$$\Pr[I_{AF} < \mathcal{R}] \sim \left(\frac{\sigma_{h1}^2 + \sigma_{h2}^2}{2\sigma_{h3}^2 (\sigma_{h1}^2 \sigma_{h2}^2)} \right) \left(\frac{2^{2\mathcal{R}} - 1}{\gamma} \right)^2. \quad (8)$$

The pre-log factor 1/2 in Equation (7) is due to half-duplex relaying which needs two channel uses to transmit the information from source to destination. The outage behavior decays as γ^{-2} , which indicates that fixed AF protocol offers diversity gain of 2.

2.4 DF Relaying

In this scheme the relay processes its received signal $y_r[k]$ in the first time slot to obtain an estimate $\hat{x}_1[k]$ of the source transmitted signal. Under a repetition-coded scheme, the relay transmits the signal $x_r[k+1] = \hat{x}_1[k]$ in the second time slot. Although fixed DF relaying has the advantage over AF relaying in reducing the effects of additive noise at the relay, it entails the possibility of forwarding erroneously detected symbols to the destination. Therefore it is required that both the relay and destination decode the entire codeword without error. This leads to the expression of maximum average mutual information I_{DF} between T_1 and T_2 as the minimum of the two maximum rates, one at which the relay R can reliably decode the source message, and the other at which the destination T_2 can reliably decode the source message given repeated transmissions from the source and relay. This implies that

$$I_{DF} = \frac{1}{2} \min \left\{ \log \left(1 + \gamma |h_1|^2 \right), \log \left(1 + \gamma |h_3|^2 + \gamma |h_2|^2 \right) \right\}. \quad (9)$$

Here it is obvious that the performance of this system is limited by the worst link among the T_1 - T_2 and T_1 - R . The outage probability can be obtained for high SNR (Laneman et al., 2004) as

$$\Pr[I_{DF} < \mathcal{R}] \sim \frac{1}{\sigma_{h1}^2} \frac{2^{2\mathcal{R}} - 1}{\gamma}. \quad (10)$$

The γ^{-1} behavior in Equation (10) indicates that fixed DF protocol does not provide diversity gain for large SNR.

2.5 Numerical Results

We compare the outage analysis results of AF and DF relaying protocols with direct transmission. We consider the case of statistically symmetric networks in which the Rayleigh fading channel variances are identical i.e., $\sigma_{hi}^2 = 1$. The noise variance σ^2 is assumed to be unity. We realized 10000 random channels using Monte Carlo simulation. Figure 3 shows the outage probabilities versus SNR in dB for low spectral efficiency (roughly 2 bps/Hz). The diversity order of 2 achieved by AF protocol is clear from the steeper curve slope in Figure 3. Also the fixed DF relaying curve indicates no diversity gain and hence does not have any advantage

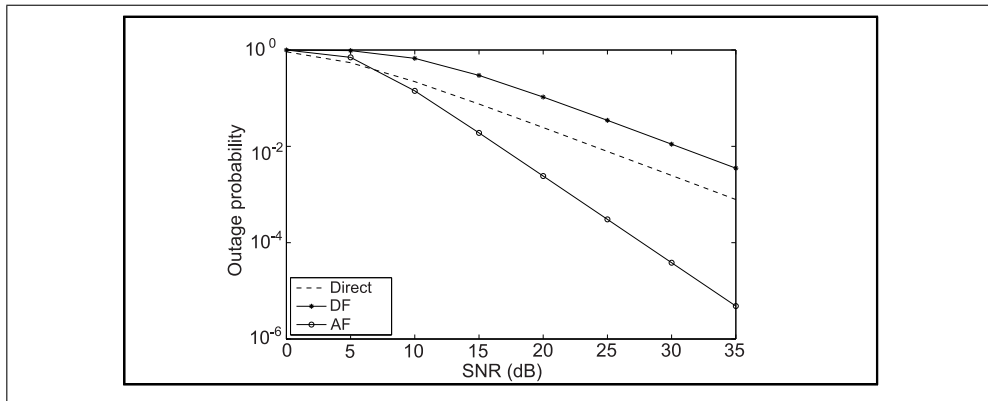


Fig. 3. Outage probabilities versus SNR in the low spectral efficiency regime.

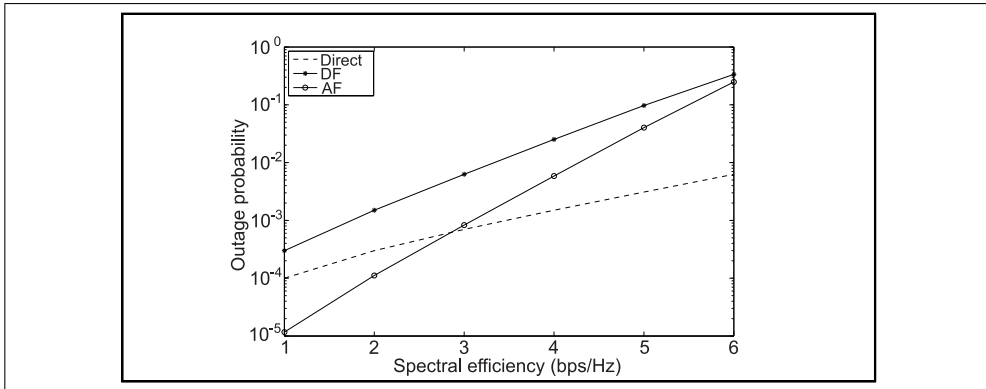


Fig. 4. Outage probabilities versus spectral efficiency for high SNR.

over direct transmission. In Figure 4, the outage probabilities are depicted as functions of spectral efficiency \mathcal{R} for a fixed SNR of 35 dB. It is clear from Figure 4 that the performance of fixed AF and DF protocols generally degrade with increasing rate \mathcal{R} . It degrades faster for AF scheme because of the inherent loss in spectral efficiency. Again, fixed DF protocol does not have any diversity advantage over direct transmission. At sufficiently high rate \mathcal{R} , direct transmission becomes more efficient than cooperative relay communication. So we can conclude that half-duplex operation requires double channel resources compared to direct transmission for a given rate and hence leads to larger effective SNR losses for increasing rate. However the performance enhancements in low spectral efficiency regime can be translated into decreased transmit power for the same reliability.

3. Bidirectional Relaying

Although unidirectional or one-way communication has been extensively considered in the literature, there is a lot of interest in recent years on bidirectional or two-way communication. In two-way communication, two terminals simultaneously transmit their messages to each other and the messages interfere with each other. The Two-Way Communication Channel (TWC) was first studied by Shannon, who derived inner and outer bounds on the capacity region (Shannon, 1961). He used a restricted two-way channel in which the encoders of both terminals do not cooperate, and the transmitted symbols at one terminal only depend on the message to be transmitted at that terminal (and not on the previously received symbols). He showed that the inner bound coincides with the capacity region of the restricted two-way channel. Later, the two-way communication problem was investigated for the full-duplex relay channel, and the achievable rate regions were derived in (Rankov & Wittneben, 2006), (Avestimehr et al., 2008), (Nam et al., 2008) and references therein. Further TWC has been exploited in (Rankov & Wittneben, 2007), as TWRC, in order to mitigate the spectral efficiency loss of cooperative protocols under half-duplex relaying. Recall that cooperative protocols can provide higher outage capacity but not ergodic capacity because of use of orthogonal time slots for relaying. Our goal here is to analyze spectrally efficient (measured in bits per channel use) transmission schemes for the half-duplex bidirectional relay channel. Presently the TWRC protocol has drawn much interest from both academic and industrial communities owing to its potential application in wireless networks.

3.1 One-Way Relay Channel (OWRC)

Consider a wireless channel in which two nodes T_1 and T_2 wish to exchange independent messages with the help of a relay node R . Once again, we assume that all terminals operate in half-duplex fashion. Therefore the relay terminal cannot receive and transmit simultaneously on the same channel resource; it receives a signal on a first hop, applies signal processing and retransmits the signal on a second hop. More importantly, there is no reliable direct link between T_1 and T_2 due to shadowing, large separation between them, or use of low power signaling. This is feasible in practice when the users are geographically separated, and the signals received from each other are very weak. This is the case when two distant land stations communicate with a satellite, or two mobile users located on opposite sides of a building communicate with the same base station on top of the building. When there is no direct connection between the two wireless terminals, relays are essential to enable communication.

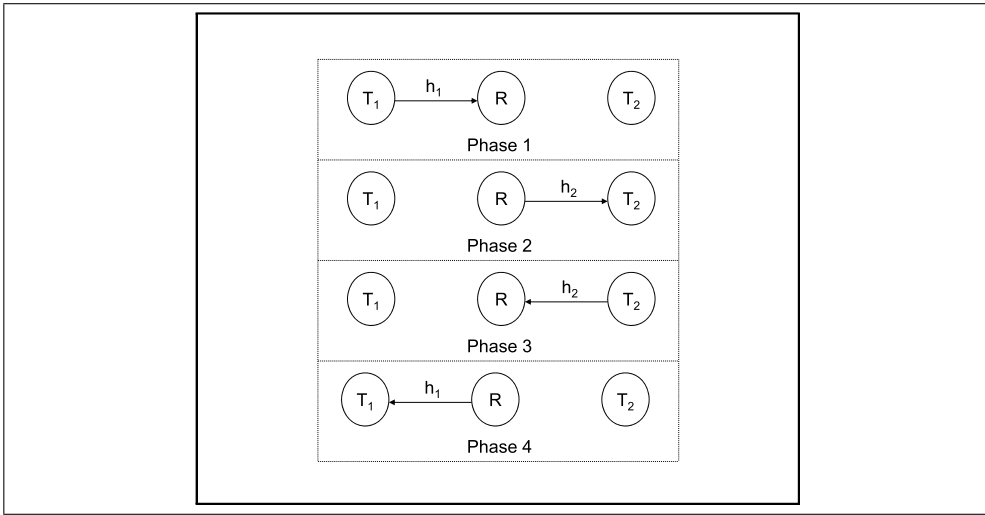


Fig. 5. Four-phase one-way relaying for bidirectional cooperation.

For a one-way relaying approach, two resources are required for the transmission from T_1 to T_2 via R and two resources are required for the transmission from T_2 to T_1 via R , leading to an overall requirement of four resources. This is a four phase protocol (Kim et al., 2008) whereby transmissions $T_1 \rightarrow R$, $R \rightarrow T_2$, $T_2 \rightarrow R$, and $R \rightarrow T_1$ occur in four consecutive phases, as illustrated in Figure 5.

3.1.1 AF-OWRC

The source terminal T_1 transmits in the first time slot an information symbol to the relay terminal R . The relay amplifies the received symbol (including noise) according to its available average transmit power and forwards a scaled signal in the second time slot to the destination terminal T_2 (Rankov & Wittneben, 2007). In time slot k , the relay receives

$$y_r[k] = h_1[k]x_1[k] + n_r[k] \quad (11)$$

where h_1 is the complex channel gain between source and relay (first hop), $x_1 \sim \mathcal{CN}(0, P_1)$ is the transmit symbol of the source, and $n_r \sim \mathcal{CN}(0, \sigma_r^2)$ is the AWGN at the relay. The relay scales $y_r[k]$ by

$$g[k] = \sqrt{\frac{P_r}{P_1|h_1[k]|^2 + \sigma_r^2}} \quad (12)$$

where P_r is the average transmit power of the relay. Depending on the amount of channel knowledge at the relay, different choices for the relay gain are possible. In time slot $k+1$, the destination receives

$$y_2[k+1] = h_2[k+1]g[k]h_1[k]x_1[k] + h_2[k+1]g[k]n_r[k] + n_2[k+1] \quad (13)$$

where h_2 is the complex channel gain between relay and destination (second hop) and $n_2 \sim \mathcal{CN}(0, \sigma_2^2)$ is the AWGN at the destination. The information rate of this scheme for i.i.d. fading channels $h_1[k]$ and $h_2[k]$ is given by (Rankov & Wittneben, 2007)

$$I_{AF} = \frac{1}{2}E \left\{ \log \left(1 + \frac{P_1|h_2g h_1|^2}{\sigma_2^2 + \sigma_r^2|h_2g|^2} \right) \right\} \quad (14)$$

where $E\{\cdot\}$ denotes the expectation with respect to the channels h_1 and h_2 . The pre-log factor $1/2$ follows because of the two channel uses needed to transmit the information from T_1 to T_2 .

3.1.2 DF-OWRC

In this scheme, the relay decodes the message sent by the source, re-encodes it (by using the same or a different codebook), and forwards the message to the destination. In time slot k , the relay receives

$$y_r[k] = h_1[k]x_1[k] + n_r[k]. \quad (15)$$

After decoding and retransmission, the destination receives in time slot $k+1$

$$y_2[k+1] = h_2[k+1]x_r[k+1] + n_2[k+1] \quad (16)$$

where $x_r \sim \mathcal{CN}(0, P_r)$ is the transmit symbol of the relay. The information rate of this scheme for i.i.d. fading channels $h_1[k]$ and $h_2[k]$ is given by (Rankov & Wittneben, 2007)

$$I_{DF} = \frac{1}{2} \min \left[E \left\{ \log \left(1 + \frac{P_1|h_1|^2}{\sigma_r^2} \right) \right\}, E \left\{ \log \left(1 + \frac{P_r|h_2|^2}{\sigma_2^2} \right) \right\} \right]. \quad (17)$$

This rate is exactly the ergodic capacity of the conventional half-duplex cooperative relay channel with no direct connection. Compared to a bidirectional communication between T_1 and T_2 without two-hop relaying, the number of required resources is doubled. Therefore this protocol is spectrally inefficient and does not take full advantage of the broadcast nature of the wireless channel.

3.2 Two-Way Relay Channel (TWRC)

The two-way relaying protocol (Rankov & Wittneben, 2007) is an effective means to increase the spectral efficiency of a half-duplex relay network. As illustrated in Figure 6, messages of nodes T_1 and T_2 are delivered to nodes T_2 and T_1 respectively in two phases, named the Multiple Access Channel (MAC) and Broadcast Channel (BC) phase. In the first (MAC) phase,

T_1 and T_2 transmit their signals to the relay node at the same time. After receiving the signals, the relay node performs appropriate signal processing and broadcasts the resulting signal to both nodes T_1 and T_2 in the second (BC) phase. At each node, its symbols contribute self interference but can clearly be canceled (because they are known). The channels in the forward direction are assumed to be the same as in the backward direction i.e., channel reciprocity is assumed.

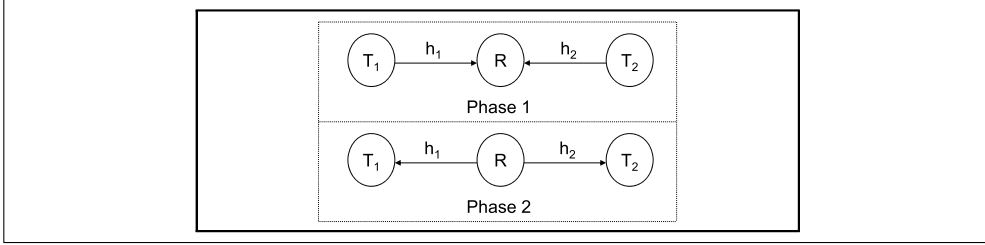


Fig. 6. Two-phase two-way relaying for bidirectional cooperation.

3.2.1 AF-TWRC

In this scheme, both terminals T_1 and T_2 transmit their symbols to relay R in the same time slot k using the same bandwidth. The relay then receives

$$y_r[k] = h_1[k]x_1[k] + h_2[k]x_2[k] + n_r[k] \quad (18)$$

where the symbols $x_1[k] \sim \mathcal{CN}(0, P_1)$ and $x_2[k] \sim \mathcal{CN}(0, P_2)$ are the i.i.d. transmit symbols of terminals T_1 and T_2 respectively. The relay scales the received signal by

$$g[k] = \sqrt{\frac{P_r}{P_1|h_1[k]|^2 + P_2|h_2[k]|^2 + \sigma_r^2}} \quad (19)$$

in order to meet its average transmit power constraint. It then broadcasts the signal in the next time slot to both terminals T_1 and T_2 . The signals received at terminals T_1 and T_2 are

$$y_2[k+1] = h_2[k+1]g[k]h_1[k]x_1[k] + h_2[k+1]g[k]h_2[k]x_2[k] + h_2[k+1]g[k]n_r[k] + n_2[k+1] \quad (20)$$

$$y_1[k+1] = h_1[k+1]g[k]h_2[k]x_2[k] + h_1[k+1]g[k]h_1[k]x_1[k] + h_1[k+1]g[k]n_r[k] + n_1[k+1]. \quad (21)$$

Assuming channel reciprocity for h_1 and h_2 . Unless mentioned otherwise, we assume that h_i remains static at least over two time slots. Since nodes T_1 and T_2 know their own transmitted symbols, they can subtract the back-propagating self-interference prior to decoding, assuming perfect knowledge of the corresponding channel coefficients. The sum-rate is given by (Rankov & Wittneben, 2007)

$$I_{AF(sum)} = \frac{1}{2}E \left\{ \log \left(1 + \frac{P_1|h_2g h_1|^2}{\sigma_2^2 + \sigma_r^2|h_2g|^2} \right) \right\} + \frac{1}{2}E \left\{ \log \left(1 + \frac{P_2|h_2g h_1|^2}{\sigma_1^2 + \sigma_r^2|h_1g|^2} \right) \right\}. \quad (22)$$

The transmission in each direction suffers still from the pre-log factor $1/2$. However, the half-duplex constraint can here be exploited to establish a bidirectional connection between two terminals and to increase the sum rate of the network.

3.2.2 DF-TWRC

Consider now a two-way communication between terminals T_1 and T_2 via a half-duplex DF relay R . In time slot k both terminals T_1 and T_2 transmit their symbols to relay R . In this MAC phase, the relay receives

$$y_r[k] = h_1[k]x_1[k] + h_2[k]x_2[k] + n_r[k], \quad (23)$$

decodes the symbols $x_1[k]$ and $x_2[k]$ and transmits $x_r[k+1] = \sqrt{\beta}x_1[k] + \sqrt{1-\beta}x_2[k]$ in the next time slot (BC phase). The received signals at T_2 and T_1 are

$$y_2[k+1] = h_2[k+1]x_r[k+1] + n_2[k+1] \quad (24)$$

$$y_1[k+1] = h_1[k+1]x_r[k+1] + n_1[k+1]. \quad (25)$$

The relay uses an average transmit power of βP_r for the forward direction and $(1-\beta)P_r$ for the backward direction. Since T_1 knows $x_1[k]$ and T_2 knows $x_2[k]$, these symbols (back-propagating self-interference) can be subtracted at the respective terminals prior to decoding of the symbol transmitted by the partner terminal. We assume that the relay decodes $x_1[k]$ and $x_2[k]$ without errors. The sum-rate is given (Rankov & Wittneben, 2007) by

$$I_{DF(sum)} = \max_{\beta} \min(I_{MA}, I_1(\beta) + I_2(1-\beta)) \quad (26)$$

$$\text{where } I_{MA} = \frac{1}{2}C \left(\frac{P_1|h_1|^2 + P_2|h_2|^2}{\sigma_r^2} \right)$$

$$I_1(\beta) = \frac{1}{2} \min \left\{ C \left(\frac{P_1|h_1|^2}{\sigma_r^2} \right), C \left(\frac{\beta P_r|h_2|^2}{\sigma_2^2} \right) \right\}$$

$$I_2(1-\beta) = \frac{1}{2} \min \left\{ C \left(\frac{P_2|h_2|^2}{\sigma_r^2} \right), C \left(\frac{(1-\beta)P_r|h_1|^2}{\sigma_1^2} \right) \right\}$$

$$\text{where } C(x) = E \{ \log(1+x) \}.$$

In the absence of CSI knowledge in the BC phase, $\beta = \frac{1}{2}$ is used by the relay. Note that in fast fading channels, the channel coefficients change from phase to phase, and reliable CSI may not be available. In other case β may be optimally chosen to maximize the sum rate. The choice of β will depend on the amount of channel knowledge available (CSI or its statistics), and applicable path losses in the links.

3.3 Simulation Results

We compute the achievable rates of the one-way and two-way relaying schemes by Monte Carlo simulations. We consider a fixed symmetric network in which the relay is equidistant from the two terminals. The Rayleigh fading channel gains are modeled as $h_i \sim \mathcal{CN}(0, 1)$. The AWGN variances are chosen as $\sigma_1^2 = \sigma_2^2 = \sigma_r^2 = \sigma^2$ and the transmit powers $P_1 = P_2 = P/2$ and $P_r = P$ such that the network consumes in each time slot an average power of P . The SNR

is defined as the ratio P/σ^2 . Over 10000 random channels were used to average the rates in Figures 7 and 8.

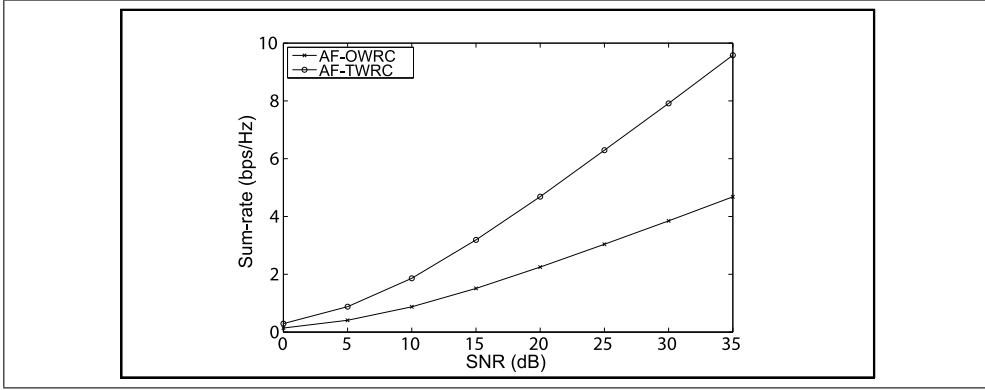


Fig. 7. Sum rate for two-way half-duplex AF relaying protocol.

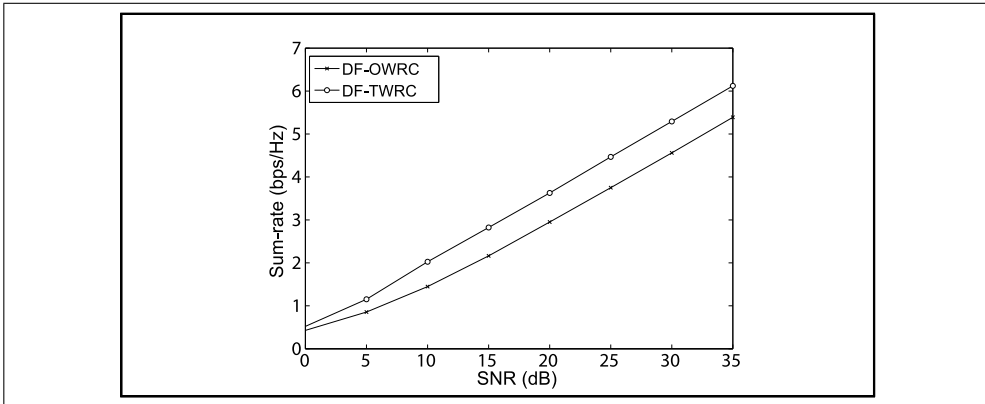


Fig. 8. Sum rate for two-way half-duplex DF relaying protocol.

We compare the sum rate of the two-way AF and two-way DF protocols with their one-way counterparts in Figures 7 and 8 respectively. We observe that both two-way protocols, AF and DF, achieve sum-rates that are larger than the rates of their one-way counterparts. Moreover, AF scheme has more pronounced improvement compared with DF. For the two-way symmetric case considered here, the DF protocol is worse than AF protocol because the sum rate is dominated by the multiple-access sum rate. Intuitively we can say that for an asymmetric channel scenario, when the relay is in the vicinity of one terminal T_1 or T_2 (and thereby experiences a stronger channel gain than the other terminal), the DF scheme achieves the maximum sum rate which can be further improved by optimal choice of β .

4. Resource Allocation

The performance of wireless relay networks can be significantly improved by efficient management of available radio resources. Mostly, resource management via power allocation is employed. We have discussed bidirectional relaying in the previous section, static resource allocation was assumed where all transmission phases are of same duration and all terminals have individual power constraints with balanced rates. Dynamic resource allocation has been investigated in (Agustin et al., 2008) in terms of phase durations, individual and sum-average power, and data rate. The system model employs a DF relay that applies superposition coding and takes into account the traffic asymmetry. It is assumed that the transmission is performed in frames of length v with N channel uses and normalized bandwidth of unity. The duration of the two phases (MAC and BC) are denoted by v_1 and v_2 respectively. The two power constraints considered are maximum power and sum-average power (both denoted by P). The first constraint assumes that all terminals transmit with power P , whereas in second case the total average power used by the three terminals is considered to be P . The mutual information of different links assuming equal noise power at all terminals is given by

$$I_{1r}(P_1) = N \log \left(1 + \frac{P_1 |h_1|^2}{\sigma^2} \right) \quad (27)$$

$$I_{2r}(P_2) = N \log \left(1 + \frac{P_2 |h_2|^2}{\sigma^2} \right) \quad (28)$$

$$I_{MAC}(P_1, P_2) = N \log \left(1 + \frac{P_1 |h_1|^2}{\sigma^2} + \frac{P_2 |h_2|^2}{\sigma^2} \right) \quad (29)$$

where I_{1r} and I_{2r} represent mutual information of $T_1 - R$ and $T_2 - R$ links respectively, and I_{MAC} is the mutual information at the relay when both terminals transmit simultaneously in the MAC phase. The signal received by the relay in MAC phase is given by

$$y_r[k] = \begin{cases} h_1[k]x_1[k] + h_2[k]x_2[k] + n_r[k] & \text{for } 0 \leq k \leq v_1 N \\ 0 & \text{for } v_1 N \leq k \leq N. \end{cases} \quad (30)$$

Under superposition coding, the DF relay forwards one signal $x_r[k]$ intended to each destination by distributing the total power between them as

$$x_r = \sqrt{\frac{\beta_1 P_r}{P_1}} x_1 + \sqrt{\frac{\beta_2 P_r}{P_2}} x_2 \quad (31)$$

where β_1 and β_2 indicate the fraction of power allocated to each signal. The signal received by each destination in second phase is given by

$$y_1[k] = \begin{cases} 0 & \text{for } 0 \leq k \leq v_1 N \\ h_1[k]x_r[k] + n_1[k] & \text{for } v_1 N \leq k \leq N \end{cases} \quad (32)$$

$$y_2[k] = \begin{cases} 0 & \text{for } 0 \leq k \leq v_1 N \\ h_2[k]x_r[k] + n_2[k] & \text{for } v_1 N \leq k \leq N. \end{cases} \quad (33)$$

The optimal selection of phase duration, data rate of each terminal are found as the maximization of the following problem (Agustin et al., 2008):

$$\arg \max_{v, R_1, R_2, P_1, P_2, P_r} \theta_1 R_1 + \theta_2 R_2 \quad \text{s.t.} \quad \begin{cases} (R_1, R_2) \in \rho(v) & \text{for } 0 \leq \ell(v) \leq 1 \\ \varphi(P_1, P_2, P_r) \leq P & \text{for } \zeta(R_1, R_2) = \kappa \end{cases} \quad (34)$$

where R_1, R_2 represents the rate transmitted by terminal T_1, T_2 respectively, \mathbf{v} is a vector that contains the duration of different phases, function $\ell(\mathbf{v})$ defines the linear connection between duration of phases, $\rho(\mathbf{v})$ denotes the achievable rate region for a given \mathbf{v} , function $\varphi(P_1, P_2, P_r)$ represents a combination of the transmitted power by the terminals considering the power constraints, function $\zeta(R_1, R_2)$ indicates a linear dependence between data rates R_1 and R_2 . The achievable rate region boundary can be attained with optimum phase and rate selection (by adjusting the parameters ϑ_1 and ϑ_2).

The achievable rate region $\rho(\mathbf{v})$ for the two-way DF protocol, described under MAC (Cover & Thomas, 1991), is given by

$$\rho(v_1, v_2) = \begin{cases} R_1 \leq \min \{v_1 I_{1r}(P_1), v_2 I_{2r}(\beta_1 P_r)\} \\ R_2 \leq \min \{v_1 I_{2r}(P_2), v_2 I_{1r}(\beta_2 P_r)\} \\ R_1 + R_2 \leq v_1 I_{MAC}(P_1, P_2) \end{cases} \quad (35)$$

with $\ell(v_1, v_2) = v_1 + v_2$. For terminals transmitting with their maximum power P , the maximum power constraint can be expressed as

$$\varphi_{max}(P_1, P_2, P_r) = \{P_1 = P, P_2 = P, P_r = P\}. \quad (36)$$

The power distribution at the relay satisfies $0 \leq \beta_1 + \beta_2 \leq 1$. Hence the power allocation can be optimized at the relay only. For sum-average power constraint, each terminal uses a fraction of total power P which is controlled by variables δ and β as follows

$$\varphi_{avg}(P_1, P_2, P_r) = \left\{ P_1 = \frac{\delta_1 P}{v_1}, P_2 = \frac{\delta_2 P}{v_1}, P_r = \frac{P}{v_2} \right\}. \quad (37)$$

It has been shown in (Agustin et al., 2008) that the sum-average power constraint must satisfy

$$P_1 v_1 + P_2 v_1 + \varsigma P_r v_2 = P \quad (38)$$

where $\varsigma = \beta_1 + \beta_2$ so that $\delta_1 + \delta_2 + \beta_1 + \beta_2 = 1$. The data rates achieved on each link are connected through

$$\zeta(R_1, R_2) = R_1 - \kappa R_2 \leq 0 \quad (39)$$

where κ is a positive real number accounts for the traffic asymmetry.

Under the sum-average power constraint the optimization problem for resource allocation is convex and has a unique solution. However for maximum power constraint, the problem has to be transformed into a convex one by introducing some auxiliary variables [see (Agustin et al., 2008) and references therein].

5. Coded Bidirectional Relaying

So far we have discussed the TWRC from cooperative communication perspectives with a major objective being compensation to make up for the half-duplex loss. In a separate but significant development, the authors in (Ahlsweide et al., 2000) have proposed the concept of network coding in which intermediate network nodes are allowed not only to route but also to mix and code the incoming data from multiple links. This reduces the amount of data transmissions in the network (thus improving the overall network throughput). Originally, the network coding concept was proposed for wired communication networks. Later it was applied to wireless communication networks by exploiting the broadcasting nature of wireless medium [it was used for relay networks for the first time in (Hao et al., 2007)]. Network

coding has been proven to be a very effective solution to overcome the interuser interference in wireless networks because of its ability to combining the different signals instead of separating them from a traditional viewpoint.

Traditionally simultaneous transmission from T_1 and T_2 was avoided in order to simplify the medium access control, and to avoid the interference at the relay R . Thereby four phases were required to perform one round of information exchange between T_1 and T_2 through R . However, by applying the idea of network coding, the authors in (Wu et al., 2005) proposed a scheme to reduce the number of required phases from four to three as illustrated in Figure 9. In this scheme, T_1 first transmits during first phase the message x_1 to R consisting of bits $b_1(1), \dots, b_1(N)$ with N denoting the message length in bits, which are decoded. During the second phase, T_2 transmits to R the message x_2 consisting of bits $b_2(1), \dots, b_2(N)$, which R decodes. In the third phase, R broadcasts to T_1 and T_2 a new message x_r consisting of bits $b_r(n)$'s, $n = 1, \dots, N$, obtained by bit-wise exclusive-or (XOR) operation over $b_1(n)$'s and $b_2(n)$'s i.e., $b_r(n) = b_1(n) \oplus b_2(n), \forall n$. Since T_1 knows $b_1(n)$'s, T_1 can recover its desired message x_2 by first decoding $b_r(n)$ and then obtaining $b_2(n)$'s of x_2 as $b_2(n) = b_1(n) \oplus b_r(n), \forall n$. Similarly, T_2 can recover x_1 . The same type of three-phase coded bidirectional relaying scheme was proposed independently in (Larsson et al., 2006). The resulting pre-log factor with respect to the sum-rate of this three-phase coded scheme is thus $2/3$ compared to $1/2$ for conventional half-duplex scheme. In this protocol, if a reliable direct link is possible, then the scheme may gain in additional diversity and often better coverage as discussed in (Kim et al., 2008).

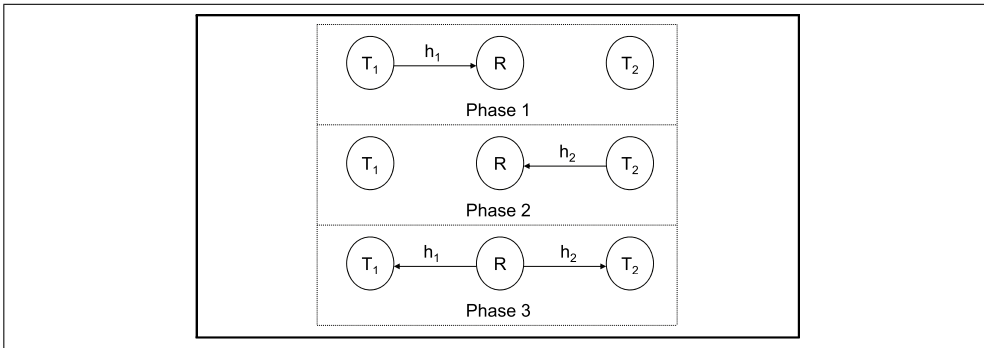


Fig. 9. Three-phase two-way relaying.

In (Popovski & Yomo, 2006); (Popovski & Yomo, 2007a); (Popovski & Yomo, 2007b), the authors reduce the number of required phases from three to two by allowing T_1 and T_2 to transmit simultaneously to R during the first phase, thereby eliminating the need for the second phase. This corresponds to the MAC phase of DF-TWRC (Rankov & Wittneben, 2007). The scheme proposed in (Katti et al., 2007a) is named as Analog Network Coding (ANC), while that in (Zhang et al., 2006) is referred to as Physical-layer Network Coding (PNC). These schemes differ in their relay operations, which are Amplify-and-Forward (AF) and Estimate-and-Forward (EF), respectively. In ANC, R simply amplifies the mixed signal received simultaneously from T_1 and T_2 and then broadcasts it to both. By subtracting the back-propagating self-interference, both T_1 and T_2 are able to receive their intended messages. Thus ANC scheme is similar to the AF-TWRC (Rankov & Wittneben, 2007). Compared to ANC, PNC (Zhang et al., 2006) performs more sophisticated operations than AF at R . Instead of decoding

messages x_1 from T_1 and x_2 from T_2 separately in two different phases like in (Wu et al., 2005), the EF method estimates at R the bitwise XORs between $b_1(n)$'s and $b_2(n)$'s from the mixed signal received, and re-encodes the decoded bits into a new broadcasting message x_r . Each of T_1 and T_2 then recovers the other's message by the same decoding method discussed in (Wu et al., 2005).

Although the schemes proposed in these works are similar to AF- and DF-TWRC, they are inspired by network coding. The principle of network coding has been further investigated for the TWRC in (Hausl & Hagenauer, 2006); (Baik & Chung, 2008); (Cui et al., 2008a); (Cui et al., 2008b). It has been shown in (Katti et al., 2007b) that joint relaying and network coding achieves higher data rates as compared to routing at the relay. In (Kim et al., 2008), the authors compared the different half-duplex bidirectional DF relaying protocols and derived their performance bounds. Note that two-phase TWRC does not exploit spatial diversity advantage like conventional approach. Including the direct link would provide diversity gain but at the cost of spectral efficiency. Even more recently, the authors in (Li et al., 2009) analyze the outage performance of two-phase AF- and DF-TWRC under half-duplex constraint. They derived the exact closed-form expressions for the outage probabilities by considering network coding at the relay for DF case. Furthermore, they propose an adaptive bidirectional relaying protocol which switches between AF and DF to minimize the outage probability of the system. TWRC coupled with network coding is thus developing as a promising technology to combat interference and to improve throughput in wireless networks.

6. MIMO Bidirectional Relaying

It is well known that Multiple-Input Multiple-Output (MIMO) communication systems have the ability to enhance the channel capacity and link reliability without requiring an increase in power or bandwidth. In (Unger & Klein, 2007) it is proposed to extend two-way relaying to terminals with multiple antennas (leading to MIMO-TWRC). They investigate the average performance of MIMO-TWRC by using multiple antennas at the relay terminal only. The proposed scheme exploits the fact that the relay R is a receiver as well as a transmitter in the dual-hop case, and hence assumes CSI at the R is not unreasonable. Like in a Time Division Duplex (TDD) system, CSI for receive and transmit processing can be obtained by directly estimating the channel from T_1 to the R and from T_2 to the R in the first phase, and then exploiting channel reciprocity in the second phase. Thereafter, the relay can perform spatial filtering to its receive and transmit signal. In (Han et al., 2008) the average sum rate improvement of two-way relaying is analyzed by deriving an upper and a lower bound for average sum rate of two-way relaying which was not derived in (Rankov & Wittneben, 2007). They also extend the work to the case when the source terminal and the destination terminal have two antennas each and the relay has only one antenna (in order to implement Alamouti's scheme). The proposed scheme achieves higher average sum rate compared to the single antenna case, and furthermore both the source and destination terminals achieve diversity of order two. The authors in (Zhang et al., 2009) analyze the capacity region for the ANC/AF-based TWRC with linear processing (beamforming) at the relay with multiple antennas. They have also shown that the ANC/AF-based TWRC have a capacity gain over the DF-based TWRC for sufficiently large channel correlations and equal MAC and BC phase-durations.

In (Hammerstrom et al., 2007) the authors further extend the two-way relaying scheme of (Rankov & Wittneben, 2007) to the case when multiple antennas are used at all terminals (assuming the knowledge of transmit CSI at the DF relay). Figure 10 shows a set up for MIMO two-way relaying where all the terminals are equipped with $M > 1$ antennas. It is assumed

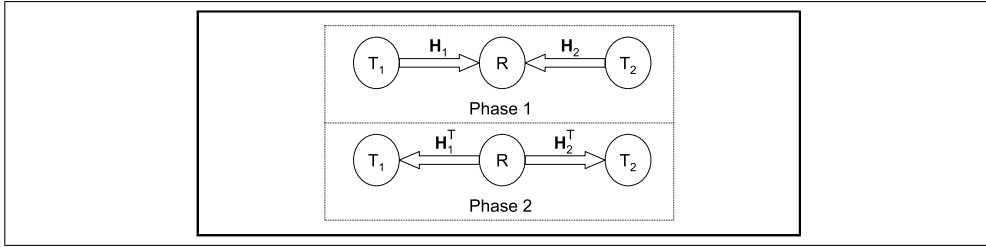


Fig. 10. MIMO two-way relay channel.

that both T_1 and T_2 perfectly know $(\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_1^T$ and $\mathbf{H}_2^T)$ in the receiving mode, but not in the transmit mode. The relay R on the other hand has knowledge of both receive and transmit CSI. This is a reasonable assumption since the relay has to estimate the channels $(\mathbf{H}_1$ and $\mathbf{H}_2)$ for decoding in the first phase and exploiting channel reciprocity in the second phase (like in TDD system). Terminal T_1 transmits vector \mathbf{x}_1 to T_2 whereas T_2 transmits vector \mathbf{x}_2 to T_1 respectively in two phases. Frequency flat slow fading and perfect synchronization is assumed between all terminals. The signal received at the relay in the first phase is given by

$$\mathbf{y}_r = \mathbf{H}_1 \mathbf{x}_1 + \mathbf{H}_2 \mathbf{x}_2 + \mathbf{n}_r \quad (40)$$

where \mathbf{H}_1 and \mathbf{H}_2 are the $M \times M$ channel matrices (with each element being i.i.d. complex Gaussian with zero mean and unit variance) that remain constant during the block transmission, \mathbf{x}_1 and \mathbf{x}_2 are $M \times 1$ symbol vectors with power P_1 and P_2 respectively, and $\mathbf{n}_r \sim \mathcal{CN}(0, \sigma_r^2 \mathbf{I}_M)$ is the $M \times 1$ complex AWGN vector.

During first phase, the relay decodes the messages from both terminals T_1 and T_2 . Using a Gaussian codebook, the achievable rates of both terminals are theoretically described by the MIMO-MAC (Cover & Thomas, 1991), which imposes constraints on the individual first-phase rates $R_{1,I}$ and $R_{2,I}$, as well as the first-phase sum rate $R_{1,I} + R_{2,I}$ for successful decoding at the destination terminal:

$$R_{1,I} \leq I_{1,I} = \log \left| \mathbf{I} + \frac{P_1}{M\sigma_r^2} \mathbf{H}_1 \mathbf{H}_1^H \right| \quad (41)$$

$$R_{2,I} \leq I_{2,I} = \log \left| \mathbf{I} + \frac{P_2}{M\sigma_r^2} \mathbf{H}_2 \mathbf{H}_2^H \right| \quad (42)$$

$$R_{1,I} + R_{2,I} \leq I_I = \log \left| \mathbf{I} + \frac{\frac{P_1}{M} \mathbf{H}_1 \mathbf{H}_1^H + \frac{P_2}{M} \mathbf{H}_2 \mathbf{H}_2^H}{\sigma_r^2} \right|. \quad (43)$$

In the second phase, the relay applies bit-level XOR precoding on decoded messages. The relay therefore broadcasts the vector \mathbf{x}_r with power P_r . The received signals at T_1 and T_2 are given by

$$\mathbf{y}_1 = \mathbf{H}_1^T \mathbf{x}_r + \mathbf{n}_1 \quad (44)$$

$$\mathbf{y}_2 = \mathbf{H}_2^T \mathbf{x}_r + \mathbf{n}_2 \quad (45)$$

where $\mathbf{n}_1 \sim \mathcal{CN}(0, \sigma_1^2 \mathbf{I}_M)$ and $\mathbf{n}_r \sim \mathcal{CN}(0, \sigma_2^2 \mathbf{I}_M)$ are the complex AWGN $M \times 1$ vectors at T_1 and T_2 respectively. Since both destinations have to be able to decode \mathbf{x}_r , the maximum data rate in the second phase is given by

$$I_{II} = \min \{I_{1,II}, I_{2,II}\} \quad (46)$$

where

$$I_{1,II} = \log \left| \mathbf{I} + \frac{1}{\sigma_1^2} \mathbf{H}_1^T \mathbf{\Lambda}_r \mathbf{H}_1^* \right| \quad (47)$$

$$I_{2,II} = \log \left| \mathbf{I} + \frac{1}{\sigma_2^2} \mathbf{H}_2^T \mathbf{\Lambda}_r \mathbf{H}_2^* \right| \quad (48)$$

where $\mathbf{\Lambda}_r = \mathbb{E} \{ \mathbf{x}_r \mathbf{x}_r^H \}$ and $\text{trace}(\mathbf{\Lambda}_r) = P_r$. The maximum sum-rate of this MIMO two-way relaying scheme is given by (Hammerstrom et al., 2007)

$$R_{sum} = \frac{1}{2} \min \{ I_I, \min \{ I_{1,I}, I_{2,II} \} + \min \{ I_{2,I}, I_{1,II} \} \}. \quad (49)$$

The above rate expression can be optimized by exploiting CSI knowledge at the relay subject to the relay transmit power constraint. This is achieved by maximizing the data rate in the second phase as follows

$$\begin{aligned} I_{II,opt} &= \max_{\mathbf{\Lambda}_r} \min \{ I_{1,II}, I_{2,II} \} \\ \text{s.t. } &\text{trace}(\mathbf{\Lambda}_r) = P_r. \end{aligned} \quad (50)$$

This optimization problem is independent of first phase data rates and can be solved by semidefinite programming method by assuming $\mathbf{\Lambda}_r$ to be positive semidefinite [see (Hammerstrom et al., 2007) and references therein].

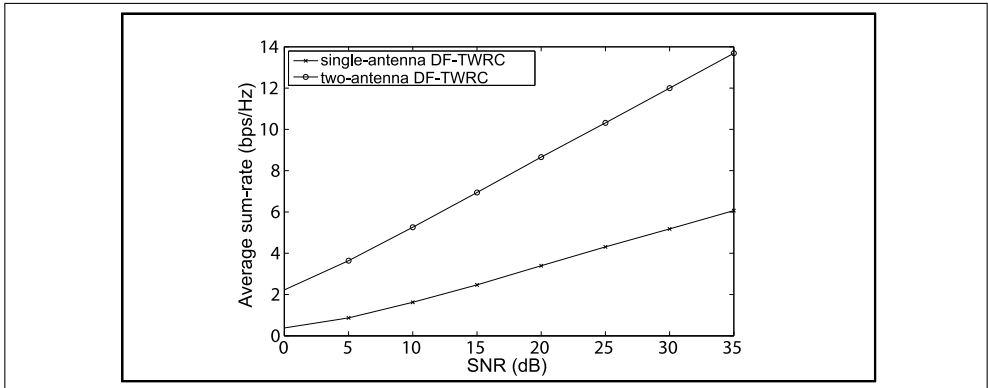


Fig. 11. Average sum-rate of two-antenna DF-TWRC with XOR precoding compared to one-antenna case.

Figure 11 compares the average sum-rate obtained (assuming Gaussian codebook) for DF-TWRC using XOR precoding with one and two antennas at the terminals T_1 and T_2 . Rayleigh

fading is assumed, and the elements of the channel matrices \mathbf{H}_1 and \mathbf{H}_2 are zero mean and unit variance complex Gaussian random variables. All nodes use the same transmit power $P_1 = P_2 = P_r = P$ and are assumed to have the same noise variance $\sigma_1^2 = \sigma_2^2 = \sigma_r^2 = \sigma^2$. The SNR is defined as P/σ^2 . We simulated 10000 random channels for each value in Figure 11. We observe that there is considerable improvement in sum-rate by using two antennas at each node compared to the single antenna case.

Further, the authors in (Hammerstrom et al., 2007) compared two approaches of combining the messages in the second phase at the relay (the superposition coding and the XOR precoding) and showed that MIMO-TWRC achieves substantial improvement in spectral efficiency compared to conventional relaying with or without transmit CSI at the relay. They also showed that the difference in sum-rate compared to the case where no CSIT is used increases with increasing ratio between number of relay antennas and number of node antennas. Also XOR precoding always achieves higher minimum user rates than superposition coding if CSIT is used. In (Oechtering & Boche, 2007) the authors propose transmit strategies in a MIMO two-way DF relaying scenario with individual power constraints. The optimal relay transmit strategy is given by two point-to-point water-filling solutions which are coupled by the relay power distribution. The diversity-multiplexing trade-off analysis for the MIMO-TWRC is dealt in (Gunduz et al., 2008). In (Yang & Chun, 2008), the transmission rate is improved by using the generalized Schur decomposition-based MIMO-TWRC.

7. Bidirectional Relaying with Multiple Relays

This section extends the theory of single-relay two-way communication to one level up in the network hierarchy by employing multiple relays. In two-way multiple relay channel two terminals T_1 and T_2 exchange information with the help of M relay terminals in two phases. Dedicated multiple relays can be utilized to relay copies of the transmitted information to the destination such that each copy experiences independent channel fading, hence providing diversity gain to the system. Such communication strategy is best suited for applications in wireless ad-hoc networks, cellular scenarios, and wireless backhaul interconnections.

7.1 Distributed Space-Time Coding

The idea of space-time coding devised for MIMO systems can be applied to a wireless relay network [see (Jing & Hassibi, 2006)] by having the relays that cooperate distributively. The concept of distributed space-time coding (Jing & Hassibi, 2006) is investigated for two-way multiple relay channel in (Cui et al., 2008c). The authors in (Cui et al., 2008c) propose a new type of relaying scheme called *partial* DF for distributed TWRC where each relay removes part of the noise before relaying information in the broadcast phase. They suggest two-way relaying protocols using Linear Dispersion (LD) codes that operate over two time slots. In this scheme, two terminals T_1 and T_2 communicate through multiple relays R_i , $i = 1, \dots, M$. Each half-duplex terminal is equipped with a single antenna. Terminal T_j , $j \in [1, 2]$, transmits the signal vector $\mathbf{s}_j = [s_{j1}, \dots, s_{jT}]^T$ where $s_{jt} \in \mathcal{A}_m$, $t = 1, \dots, T$, \mathcal{A}_m is a finite constellation with average power of unity, and T is the length of each time slot. Hence, $E\{\mathbf{s}_j^H \mathbf{s}_j\} = T$. The average power of terminal T_j is P_j and each relay has the equal power P_r/M so that the total power of all the relays is P_r . The noise variance is assumed to be unity at every node. During the first phase, both terminals T_1 and T_2 transmit their message to the relays. The signal received by the i th relay is given by

$$\mathbf{y}_{ri} = \sqrt{2P_1}h_{1i}\mathbf{s}_1 + \sqrt{2P_2}h_{2i}\mathbf{s}_2 + \mathbf{n}_{ri} \quad (51)$$

where $h_{ji} \sim \mathcal{CN}(0, 1)$ represents the channel gain between terminal T_j and relay R_i , \mathbf{n}_{ri} is the $T \times 1$ vector representing the AWGN at the i th relay. The source transmit power is assumed to be $\sqrt{2P_j}$ because each source terminal transmits every two time slots. During second time slot, the i th relay processes \mathbf{y}_{ri} and transmits \mathbf{s}_{ri} scaled by g to maintain average power P_r . The signal received by j th terminal is given as

$$\mathbf{y}_j = \sum_{i=1}^M g h_{ji} \mathbf{s}_{ri} + \mathbf{n}_j, \quad j = 1, 2, \quad (52)$$

where \mathbf{n}_j is the AWGN vector at the j th terminal.

7.1.1 2-AF

In this scheme \mathbf{s}_{ri} is obtained by precoding \mathbf{y}_{ri} with a unitary matrix \mathbf{W}_{ri} and then scaled by $g = \sqrt{\frac{2P_r}{M(2P_1+2P_2+1)}}$. The signal received at terminal T_2 is given by

$$\mathbf{y}_2 = g \left(\sqrt{2P_1} \mathbf{S}_1 \mathbf{h}_1' + \sqrt{2P_2} \mathbf{S}_2 \mathbf{h}_2' \right) + \mathbf{z}_2 = g \left(\sqrt{2P_1} \mathbf{H}_1 \mathbf{s}_1 + \sqrt{2P_2} \mathbf{H}_2 \mathbf{s}_2 \right) + \mathbf{z}_2 \quad (53)$$

where $\mathbf{S}_j = [\mathbf{W}_{r1}\mathbf{s}_j, \dots, \mathbf{W}_{rM}\mathbf{s}_j]$, $\mathbf{h}_1' = [h_{11}h_{21}, \dots, h_{1M}h_{2M}]^T$, $\mathbf{h}_2' = [h_{21}^2, \dots, h_{2M}^2]^T$, $\mathbf{z}_2 = g \sum_{i=1}^M h_{2i} \mathbf{W}_{ri} \mathbf{n}_{ri} + \mathbf{n}_2$, $\mathbf{H}_1 = g \sum_{i=1}^M h_{1i} h_{2i} \mathbf{W}_{ri}$, $\mathbf{H}_2 = g \sum_{i=1}^M h_{2i}^2 \mathbf{W}_{ri}$.

Since terminal T_2 knows the back propagating signal \mathbf{s}_2 , the maximum-likelihood (ML) decoding of \mathbf{s}_1 is obtained as (Cui et al., 2008c)

$$\hat{\mathbf{s}}_1 = \arg \min_{\hat{\mathbf{s}}_1 \in \mathcal{A}_1^M} \left\| \mathbf{y}_2 - g \left(\sqrt{2P_1} \mathbf{H}_1 \hat{\mathbf{s}}_1 + \sqrt{2P_2} \mathbf{H}_2 \mathbf{s}_2 \right) \right\|^2. \quad (54)$$

Similarly, ML decoding of \mathbf{s}_2 is performed at terminal T_1 . The disadvantage of this scheme is that it amplifies the relay noise.

7.1.2 Partial DF I

This protocol overcomes the drawback of 2-AF scheme (Cui et al., 2008c) by allowing each relay R_i to first decode \mathbf{s}_1 and \mathbf{s}_2 via the ML decoder

$$\{\hat{\mathbf{s}}_{1i}, \hat{\mathbf{s}}_{2i}\} = \arg \min_{\hat{\mathbf{s}}_{1i} \in \mathcal{A}_1^M, \hat{\mathbf{s}}_{2i} \in \mathcal{A}_2^M} \left\| \mathbf{y}_{ri} - \sqrt{2P_1} h_{1i} \mathbf{W}_{ri} \hat{\mathbf{s}}_1 - \sqrt{2P_2} h_{2i} \mathbf{W}_{ri} \hat{\mathbf{s}}_2 \right\|^2. \quad (55)$$

The number of unknowns in the above equation is twice the number of equations, this making the error probability high. For this reason it has been suggested in (Cui et al., 2008c) that instead of sending $\hat{\mathbf{s}}_{1i}$ and $\hat{\mathbf{s}}_{2i}$ directly, each relay transmits

$$\mathbf{s}_{ri} = \mathbf{W}_{ri} \left(\sqrt{2P_1} h_{1i} \hat{\mathbf{s}}_{1i} + \sqrt{2P_2} h_{2i} \hat{\mathbf{s}}_{2i} \right) \quad (56)$$

scaled by $g = \sqrt{\frac{P_r}{M(P_1+P_2)}}$. Thus the relays remove noise from the received signal without dealing with the channel effects. If $Pr(\Delta \mathbf{s}_{1i}, \Delta \mathbf{s}_{2i})$ represents the pairwise error probability at the i th relay, where $\Delta \mathbf{s}_{1i} = \mathbf{s}_1 - \hat{\mathbf{s}}_{1i}$ and $\Delta \mathbf{s}_{2i} = \mathbf{s}_2 - \hat{\mathbf{s}}_{2i}$, then the ML decoding at T_2 is given as (Cui et al., 2008c)

$$\hat{\mathbf{s}}_1 = \arg \max_{\tilde{\mathbf{s}}_1 \in \mathcal{A}_1^M} \sum_{\Delta \mathbf{s}_{1i}, \Delta \mathbf{s}_{2i}} \prod_{i=1}^M Pr(\Delta \mathbf{s}_{1i}, \Delta \mathbf{s}_{2i}) \exp \left\{ -\|\mathbf{y}_2 + \mathbf{y}' - \mathbf{y}''\|^2 \right\} \quad (57)$$

where $\mathbf{y}' = g \sum_{i=1}^M \mathbf{W}_{ri} \left(\sqrt{2P_1} h_{1i} \Delta \mathbf{s}_{1i} + \sqrt{2P_2} h_{2i} \Delta \mathbf{s}_{2i} \right)$ and $\mathbf{y}'' = g \left(\sqrt{2P_1} \mathbf{H}_1 \tilde{\mathbf{s}}_1 + \sqrt{2P_2} \mathbf{H}_2 \mathbf{s}_2 \right)$.

It is difficult to implement the above decoder directly when either number M or constellation size is large. At high SNR, $\prod_{i=1}^M Pr(\Delta \mathbf{s}_{1i}, \Delta \mathbf{s}_{2i})$ is dominated by $\Delta \mathbf{s}_{1i} = \mathbf{0}, \Delta \mathbf{s}_{2i} = \mathbf{0}$. Therefore the ML decoding at terminal T_2 can be approximated as follows

$$\hat{\mathbf{s}}_1 = \arg \min_{\tilde{\mathbf{s}}_1 \in \mathcal{A}_1^M} \left\| \mathbf{y}_2 - g \left(\sqrt{2P_1} \mathbf{H}_1 \tilde{\mathbf{s}}_1 + \sqrt{2P_2} \mathbf{H}_2 \mathbf{s}_2 \right) \right\|^2. \quad (58)$$

Similarly, ML decoding at terminal T_1 can be approximated (Cui et al., 2008c).

7.1.3 Partial DF II

In both AF and *partial* DF I schemes, a weighted sum of symbols is transmitted from two terminals. This causes wastage of power since each destination knows the back-propagating signal. In *partial* DF II (Cui et al., 2008c), components are superimposed via modular arithmetic. Let the size of constellation \mathcal{A}_j be Z_j with $\mathcal{A}_j(q)$ representing the q th element of \mathcal{A}_j , where $j = 1, 2$ and $q = 0, \dots, Z_j - 1$. Consider \mathbf{u}_1 and \mathbf{u}_2 such that $\mathcal{A}_1(\mathbf{u}_1) = \mathbf{s}_1$ and $\mathcal{A}_2(\mathbf{u}_2) = \mathbf{s}_2$. With the setting $Z = \max\{Z_1, Z_2\}$, it can be assume that $Z_1 \geq Z_2$ without loss of generality. Under this protocol, each relay obtains $\hat{\mathbf{s}}_{1i}, \hat{\mathbf{s}}_{2i}$ from Equation (55) as in *partial* DF I. If $\mathcal{A}_1(\hat{\mathbf{u}}_{1i}) = \hat{\mathbf{s}}_{1i}$ and $\mathcal{A}_2(\hat{\mathbf{u}}_{2i}) = \hat{\mathbf{s}}_{2i}$ then each relay transmits

$$\mathbf{s}_{ri} = \mathbf{W}_{ri} \mathcal{A}_1(\text{mod}(\hat{\mathbf{u}}_{1i} + \hat{\mathbf{u}}_{2i}, Z)) \quad (59)$$

where “mod” stand for the componentwise modular operation and $g = g_i = \sqrt{\frac{2P_r}{M}}$. Since fading channels are considered, the probability that there exists a pair of vectors $\{\mathbf{u}_1, \mathbf{u}_2\}$ and $\{\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2\}$ such that $\sqrt{2P_1} h_{1i} \mathcal{A}_1(\mathbf{u}_1) + \sqrt{2P_2} h_{2i} \mathcal{A}_2(\mathbf{u}_2) = \sqrt{2P_1} h_{1i} \mathcal{A}_1(\hat{\mathbf{u}}_1) + \sqrt{2P_2} h_{2i} \mathcal{A}_2(\hat{\mathbf{u}}_2)$ is very small. It has been shown in (Cui et al., 2008) that the AF protocol achieves the diversity order $\min\{M, T\} \left(1 - \frac{\log \log P}{\log P}\right)$, where P is the total power of the network whereas the *partial* DF II protocol achieves a diversity order M when $T \geq M$.

7.2 Distributed Relay Selection Scheme

A distributed relay selection strategy is proposed in (Ding et al., 2009) that selects the best suited relay for realizing PNC in a dense relays network. In this transmission scheme, both T_1 and T_2 broadcast their information to all M relays simultaneously during first phase. The signal received at the i th relay R_i is given by

$$\mathbf{y}_{ri} = \sqrt{P} h_{1i} \mathbf{s}_1 + \sqrt{P} h_{2i} \mathbf{s}_2 + \mathbf{n}_{ri} \quad (60)$$

where P is the source transmission power, \mathbf{s}_j represents the unit-power signal transmitted by the j th source, and h_{ji} represents the channel gain between the j th source and the i th relay. The channel model for frequency flat Rayleigh fading is considered as

$$h_{ji} = \frac{\tilde{h}_{ji}}{\sqrt{d_{ji}^\alpha}} \quad (61)$$

where h_{ji} accounts for the channel fading characteristics due to the rich scattering environment, d_{ji} represents the distance between the j th source and the i th relay, and α is the path loss exponent. It is reasonable to assume that each relay terminal has its local channel information under channel reciprocity condition. This local channel information can be exploited in realizing a distributed strategy of relay selection to improve the system performance. For instance, consider that the relay R_b has been chosen as the best relay with corresponding channels h_{1b} and h_{2b} . In the second phase the best relay R_b performs AF operation and transmits the mixed signal given by

$$s_{rb} = \frac{\sqrt{P}h_{1b}s_1 + \sqrt{P}h_{2b}s_2 + n_{rb}}{\sqrt{P|h_{1b}|^2 + P|h_{2b}|^2 + \sigma^2}} \sqrt{P} \quad (62)$$

to the two destinations. After removing the back-propagating self interference, the signal received at j th terminal is given by

$$y_j = \frac{\sqrt{P}h_{jb}}{\sqrt{P|h_{1b}|^2 + P|h_{2b}|^2 + \sigma^2}} (\sqrt{P}h_{lb}s_l + n_{rb}) + n_j. \quad (63)$$

Therefore the mutual information between the l th source and j th destination is given by

$$I_{jl} = \log \left(1 + \frac{\gamma^2 |h_{1b}|^2 |h_{2b}|^2}{2\gamma |h_{jb}|^2 + \gamma |h_{lb}|^2 + 1} \right), \forall \quad j \neq l \quad \& \quad j, l \in [1, 2], \quad (64)$$

where $\gamma = P/\sigma^2$ represents the SNR. Relay selection (Ding et al., 2009) is performed in medium access layer. It has been claimed that the two destinations have different preferences but they do not tend to contradict each other. The relay with channels yielding large I_{12} also has channels that give a large value for I_{21} , if not exactly the maximum. The best relay is selected based on the following criterion

$$\frac{|h_{1b}|^2 |h_{2b}|^2}{2\gamma |h_{1b}|^2 + \gamma |h_{2b}|^2 + 1} \quad (65)$$

that maximizes the value of I_{12} . Then for this selected relay, the mutual information for second source, I_{21} is determined. The relay selected in such a way is suboptimal for the second source and hence some performance loss for the second source can be expected. The outage probability for this scheme (Ding et al., 2009) at high SNR is

$$Pr[I_{jl} < \mathcal{R}] = \frac{[(d_{jb}^\alpha + 2d_{lb}^\alpha)(2^{\mathcal{R}} - 1)]^M}{\gamma^M}. \quad (66)$$

It is clear from Equation (66) that the proposed transmission scheme in (Ding et al., 2009) has the advantage of diversity of order M . While the authors in (Ding et al., 2009) dealt with the relay selection scheme for the specific case of PNC-based TWRC, the problem is relevant in all TWRC based links.

8. Summary and Future Directions

Cooperative relaying has evolved in recent years as a powerful tool to enhance the reliability and throughput of wireless radio networks. The basic research challenge is to design spectrally efficient relaying schemes for better utilization of the available resources like power

and spectrum. In this chapter, we discussed several half-duplex cooperative relaying protocols and their performances. Among these, two-way relaying is envisioned as a promising protocol to save radio resources in wireless networks, whereby both up- and down-link are transmitted on the same channel resources.

There are still many open issues related to the channels investigated so far. Mostly the slow frequency flat fading scenarios has been considered in the literature, performance analysis for fast as well as frequency selective fading two-way relay channels need to be addressed. The theoretical capacity limits or achievable rate regions of TWRC still needs to be developed for clustered and distributed scenarios. The reported protocols still suffer performance loss as compared to the theoretical bounds. So better code designs with acceptable complexity need to be urgently evolved to meet the above challenge. MIMO bidirectional relaying strategies has already gained some momentum, but schemes like beamforming, distributed coding, and relay selection still need to be explored. Perfect synchronization among multiple radios is perhaps the most difficult task to perform for bidirectional traffic in a cooperative network. Cooperative relaying techniques can be expected to be adopted in future wireless systems, as it has been introduced in the IEEE 802.16j (WiMAX) standard. However, substantial research efforts are needed to construct practical systems based on bidirectional cooperation for larger wireless networks. Immense research interest is currently being focused to assess whether the cooperation technology enables the implementation of cognitive radio.

9. References

- Agustin, A.; Vidal, J. & Munoz, O. (2008). Protocols and resource allocation for the two-way relay channel with half-duplex terminals. *Submitted to IEEE Transactions on Wireless Communications*, Paper-TW-Oct-08-1427, Oct. 2008.
- Ahlsweide, R.; Cai, N.; Li, S.Y.R. & Yeung, R.W. (2000). Network information flow. *IEEE Transactions on Information Theory*, Vol. 46, No. 4, July 2000, 1204-1216.
- Avestimehr, A.S.; Sezgin, A. & Tse, D.N.C. (2008). Approximate capacity of the two-way relay channel: A deterministic approach. *Available at <http://arxiv.org/PS-cache/arxiv/pdf/0808/0808.3145v1.pdf>*, Aug. 2008.
- Baik, I.J. & Chung, S.Y. (2008). Network coding for two-way relay channels using lattices. *Proceedings of IEEE International Conference on Communications (ICC)*, 3898-3902, June 2008.
- Chen, M. & Yener, A. (2008). Multiuser two-way relaying for interference limited systems. *Proceedings of IEEE International Conference on Communications (ICC)*, 3883-3887, May 2008.
- Cover, T.M. & Gamal, A.A.E. (1979). Capacity theorems for the relay channel. *IEEE Transactions on Information Theory*, Vol. 25, No. 5, Sep. 1979, 572-584.
- Cover, T.M. & Thomas, J.A. (1991). *Elements of Information Theory*, ISBN 0-471-06259-6, New York: Wiley, 1991.
- Cui, T.; Ho, T. & Klierer, J. (2008a). Memoryless relay strategies for two-way relay channels: Performance analysis and optimization. *Proceedings of IEEE International Conference on Communications (ICC)*, May 2008.
- Cui, T.; Gao, F.F. & Tellambura, C. (2008b). Physical layer differential network coding for two-way relay channels. *Proceedings of IEEE Global Communications Conference (GLOBE-COM)*, New Orleans, LA, USA, Nov. 2008, LA, USA.

- Cui, T.; Gao, F.F.; Ho, T. & Nallanathan, A. (2008c). Distributed space-time coding for two-way wireless relay networks. *Proceedings of IEEE International Conference on Communications (ICC)*, Beijing, May 2008, China.
- Ding, Z.; Leung, K.K.; Goeckel, D. & Towsley, D. (2009). On the study of network coding with diversity. *IEEE Transactions on Wireless Communications*, Vol. 8, No. 3, Mar. 2009, 1247-1259.
- Esli, C. & Wittneben, A. (2008). One- and two-way decode-and-forward relaying for wireless multiuser MIMO networks. *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, New Orleans, Nov. 2008, LA, USA.
- Gao, F.; Zhang, R. & Liang, Y.C. (2008). On channel estimation for amplify-and-forward two-way relay networks. *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, New Orleans, Nov. 2008, LA, USA.
- Gunduz, D.; Goldsmith, A. & Vincent Poor, H. (2008). MIMO two-way relay channel: diversity-multiplexing tradeoff analysis. Available at <http://arxiv.org/PS-cache/arxiv/pdf/0812/0812.3642v1.pdf>, Dec. 2008.
- Hammerstrom, L.; Kuhn, M.; Esli, C.; Zhao, J.; Wittneben, A. & Bauch, G. (2007). MIMO two-way relaying with transmit CSI at the relay. *Proceedings of IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAW)*, June 2007, Finland.
- Han, Y.; Ting, S.H.; Ho, C.K. & Chin, W.H. (2008). High rate two-way amplify-and-forward half-duplex relaying with OSTBC. *Proceedings of IEEE Vehicular Technology Conference (VTC)*, Marina Bay, 2426-2430, May 2008, Singapore.
- Hao, Y.; Goeckel, D.; Ding, Z.; Towsley, D. & Leung, K.K. (2007). Achievable rates for network coding on the exchange channel. *Proceedings of IEEE Military Communications Conference (MILCOM)*, Orlando, 371-382, Oct. 2007, Florida.
- Hausl, C. & Hagenauer, J. (2006). Iterative network and channel decoding for the two-way relay channel. *Proceedings of IEEE International Conference on Communications (ICC)*, 1568-1573, June 2006.
- Ho, C.K.; Zhang, R. & Liang, Y.C. (2008). Two-way relaying over OFDM: Optimized tone permutation and power allocation. *Proceedings of IEEE International Conference on Communications (ICC)*, 3908-3912, May 2008.
- Jing, Y. & Hassibi, B. (2006). Distributed space-time coding in wireless relay networks. *IEEE Transactions on Wireless Communications*, Vol. 5, No. 12, Dec. 2006, 3524-3536, ISSN 1536-1276.
- Jitvanichphaibool, K.; Zhang, R. & Liang, Y.C. (2008). Optimal resource allocation for two-way relay-assisted OFDMA. *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, New Orleans, Nov. 2008, LA, USA.
- Katti, S.; Gollakota, S. & Katabi, D. (2007a). Embracing wireless interference: Analog network coding. *Proceedings of Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, 397-408, ISBN 978-1-59593-713-1, Kyoto, Feb. 2007, Japan.
- Katti, S.; Maric, I.; Goldsmith, A.; Katabi, D. & Medard, M. (2007b). Joint relaying and network coding in wireless networks. *Proceedings of IEEE International Symposium on Information Theory*, Nice, June 2007, France.
- Kim, S.J.; Mitran, P. & Tarokh, V. (2008). Performance bounds for bi-directional coded cooperation protocols. *IEEE Transactions on Information Theory*, Vol. 54, No. 11, Nov. 2008, 5235-5241.

- Laneman, J.N.; Wornell, G.W. & Tse, D.N.C. (2004). Cooperative diversity in wireless networks: Efficient protocols and outage behavior. *IEEE Transactions on Information Theory*, Vol. 50, No. 12, Dec. 2004, 3062-3080, ISSN 0018-9448.
- Larsson, P.; Johansson, N. & Sunell, K.E. (2006). Coded bidirectional relaying. *Proceedings of IEEE Vehicular Technology Conference (VTC)*, 851-855, ISBN 1-7803-9392-9, Melbourne, May 2006, Australia.
- Li, Q.; Ting, S.H.; Pandharipande, A. & Han, Y. (2009). Adaptive two-way relaying and outage analysis. *IEEE Transactions on Wireless Communications*, Vol. 8, No. 6, June 2009.
- Meulen, E.C. (1971). Three-terminal communication channels. *Adv. Appl. Probab.*, Vol. 3, No. 1, 120-154, 1971.
- Nam, W.; Chung, S.Y. & Yong H.L. (2008). Capacity bounds for two-way relay channels. *Proceedings of IEEE International Zurich Seminar on Communications*, 2008.
- Oechtering, T.J. & Boche, H. (2007). Optimal transmit strategies in multi-antenna bidirectional relaying. *Proceedings of IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, Apr. 2007, USA.
- Oechtering, T.J.; Schnurr, C.; Bjelakovic, I. & Boche, H. (2008). Broadcast capacity region of two-phase bidirectional relaying. *IEEE Transactions on Information Theory*, Vol. 54, No. 1, Jan. 2008, 454-458.
- Popovski, P. & Yomo, H. (2006). Bi-directional amplification of throughput in a wireless multi-hop network. *Proceedings of IEEE Vehicular Technology Conference (VTC)*, 588-593, ISBN 1-7803-9392-9, Melbourne, May 2006, Australia.
- Popovski, P. & Yomo, H. (2007a). Wireless network coding by amplify-and-forward for bidirectional traffic flows. *Proc. IEEE Communication Letters*, Vol. 11, No. 1, Jan. 2007, 16-18, ISSN 1089-7798.
- Popovski, P. & Yomo, H. (2007b). Physical network coding in two-way wireless relay channels. *Proceedings of IEEE International Conference on Communications (ICC)*, 707-712, ISBN 1-4244-0353-7, Glasgow, June 2007, Scotland.
- Rankov, B. & Witteben, A. (2006). Achievable rate regions for the two-way relay channel. *Proceedings of IEEE International Symposium on Information Theory*, Seattle, 2006, WA.
- Rankov, B. & Witteben, A. (2007). Spectral efficient protocols for half-duplex relay channels. *IEEE Journal on Selected Areas in Communications*, Vol. 25, No. 2, Feb. 1979, 379-389, ISSN 0733-8716.
- Sendonaris, A.; Erkip, E. & Aazhang, B.(2003). User cooperation diversity-Part I: System description. *IEEE Transactions on Communications*, Vol. 51, No. 11, Nov. 2003, 1927-1938, ISSN 0090-6778.
- Shannon, C.E. (1961). Two-way communication channels. *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, Vol. 1, 1961, 611-644, UC Press, Berkeley.
- Unger, T. & Klein, A. (2007). Linear transceive filters for relay stations with multiple antennas in the two-way relay channel. *Proceedings of 16th IST Mobile and Wireless Communications Summit*, Budapest, July 2007, Hungary.
- Vaze,R. & Heath Jr., R.W. (2008). To code or not to code in multi-hop relay channels. *available at <http://arxiv.org/abs/0805.3164>*, May, 2008.
- Wu, Y.; Chou, P. & Kung, S.Y. (2005). Information exchange in wireless networks with network coding and physical-layer broadcast. *Proceedings of 39th Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, Mar. 2005, MD.

- Yang, H.J. & Chun, J. (2008). Generalized Schur decomposition-based two-way relaying for wireless MIMO systems. *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, New Orleans, Nov. 2008, LA, USA.
- Zhang, S.; Liew, S.C. & Lam, P.P. (2006). Physical-layer network coding. *Proceedings of the 12th Annual International Conference on Mobile Computing and Networking*, 358-365, ISBN 1-59593-286-0, Los Angeles, CA, Sept. 2006, USA.
- Zhang, R.; Liang, Y.C.; Chai, C.C. & Cui, S. (2009). Optimal beamforming for two-way multi-antenna relay channel with analogue network coding. *IEEE Journal on Selected Areas in Communications*, Vol. 27, No. 5, June 2009, 699-712.
- Zhao, J.; Kuhn, M.; Wittneben, A. & Bauch, G. (2008). Self-interference aided channel estimation in two-way relaying systems. *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, New Orleans, Nov. 2008, LA, USA.

A Novel Amplify-and-Forward Relay Channel Model for Mobile-to-Mobile Fading Channels Under Line-of-Sight Conditions

Batoool Talha and Matthias Pätzold
*University of Agder
Norway*

Mobile-to-mobile (M2M) fading channels in cooperative networks can efficiently be modeled using the multiple scattering concept. In this chapter, we propose a new second-order scattering channel model referred to as the multiple-LOS second-order scattering (MLSS) channel model¹ for M2M fading channels in amplify-and-forward relay links under line-of-sight (LOS) conditions, where the received signal comprises only the single and double scattered components. In the proposed model, LOS components exist in the direct link between the source mobile station and the destination mobile station as well as the link via the mobile relay. Analytical expressions are derived for the probability density function (PDF) of the envelope and phase of M2M fading channels. It is shown mathematically that the proposed model includes as special cases double Rayleigh, double Rice, single-LOS double-scattering (SLDS), non-line-of-sight (NLOS) second-order scattering (NLSS), and single-LOS second-order scattering (SLSS) processes. The validity of all theoretical results is confirmed by simulations. Our novel M2M channel model is important for the investigation of the overall system performance in different M2M fading environments under LOS conditions.

1. Introduction

Among several emerging wireless technologies, M2M communications in cooperative networks has gained considerable attention in recent years. The driving force behind merging M2M communications and cooperative networks is its promise to provide a better link quality (diversity gain), an improved network range, and an overall increase in the system capacity. M2M cooperative wireless networks exploit the fact that single-antenna mobile stations cooperate with each other to share their antennas in order to form a virtual multiple-input multiple-output (MIMO) system in a multi-user scenario (Dohler, 2003). Thus, in such networks, cooperative diversity (Laneman et al., 2004; Sendonaris et al., 2003a;b) is achieved by relaying the signal transmitted from a mobile station to the final destination using other mobile stations in the network. However, to cope with the problems faced within the development of such systems, a solid knowledge of the underlying multipath fading channel characteristics is essential. Therefore, the aim of this chapter is to develop a flexible M2M fading channel model for relay-based cooperative networks and to analyze its statistical

¹ The material in this chapter was presented in part at the 19th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2008, Cannes, France, September 2008.

properties. This newly developed model would help communication system designers to investigate the overall performance of cooperative communication systems.

So far, M2M amplify-and-forward relay fading channels have been modeled only for some specific communication scenarios, either assuming NLOS or partial LOS propagation conditions. It has been shown in (Patel et al., 2006) that under NLOS propagation conditions, the M2M amplify-and-forward relay fading channel can be modeled as a double Rayleigh fading channel (Erceg et al., 1997; Kovacs et al., 2002). Motivated by the studies of double Rayleigh fading channels for keyhole channels (Almers et al., 2006), the so-called double Nakagami- m fading channel model has also been proposed in (Shin & Lee, 2004). Furthermore, in amplify-and-forward relay environments, the M2M fading channel under LOS conditions can be modeled as a double Rice fading channel (Talha & Pätzold, 2007a) and/or as an SLDS fading channel (Talha & Pätzold, 2007b). In addition, M2M fading channels in cooperative networks can efficiently be modeled using the multiple scattering concept (Andersen, 2002).

In multiple scattering radio propagation environments, the received signal comprises a sum of the single, double, or generally multiple scattered components. In this chapter, we model the amplify-and-forward relay fading channel as a second-order scattering channel, i.e., the sum of only the single and the double scattered components (Salo et al., 2006). The novelty of this approach is that we have extended the NLSS (Salo et al., 2006) and the SLSS (Salo et al., 2006) channel models to an MLSS channel model (Talha & Pätzold, 2008b) by incorporating multiple LOS components in all transmission links, i.e., in the direct link between the source mobile station and the destination mobile station as well as in the link via the mobile relay. Furthermore, an important feature of the proposed MLSS channel model for M2M fading channels is that it includes several other well-known channel models as special cases, e.g., the double Rayleigh model, the double Rice model, the SLDS model, the NLSS model, and the SLSS model. Here, we derive an analytical expression of the PDF of the MLSS process, along with the PDF of the corresponding phase process. The correctness of all theoretical results would be confirmed using a high-performance channel simulator. Furthermore, all presented results provide evidence that the statistics of MLSS fading channels are entirely different from the special cases discussed above.

The chapter is structured as follows: In Section 2, the reference model for the amplify-and-forward MLSS fading channel is developed. Section 3 deals with the analysis of the statistical properties of MLSS fading processes. Section 4 confirms the validity of the analytical expressions presented in Section 3 by simulations. Finally, concluding remarks are made in Section 5.

2. The MLSS Fading Channel

Under NLOS propagation conditions, the complex time-varying channel gain of the multiple scattering radio propagation channel proposed in (Andersen, 2002) can be written as

$$\chi(t) = \alpha_1 \mu^{(1)}(t) + \alpha_2 \mu^{(2)}(t) \mu^{(3)}(t) + \alpha_3 \mu^{(4)}(t) \mu^{(5)}(t) \mu^{(6)}(t) + \dots \quad (1)$$

where $\mu^{(i)}(t)$ ($i = 1, 2, 3, \dots$) is a zero-mean complex Gaussian process that represents the scattered component of the i th link and α_i ($i = 1, 2, 3, \dots$) is a real-valued constant that

determines the contribution of i th scattered component. In (1), Gaussian processes $\mu^{(i)}(t)$ are mutually independent. However, when the fading channel is modeled by taking into account only the first two terms of (1), the resulting channel is referred to as the NLSS channel (Salo et al., 2006). Here, we are presenting an extension of the NLSS channel to the MLSS channel by incorporating LOS components in a novel manner for M2M amplify-and-forward relay fading channels. The considered communication scenario determined by a source mobile station, a destination mobile station and mobile relay is shown in Fig. 1.

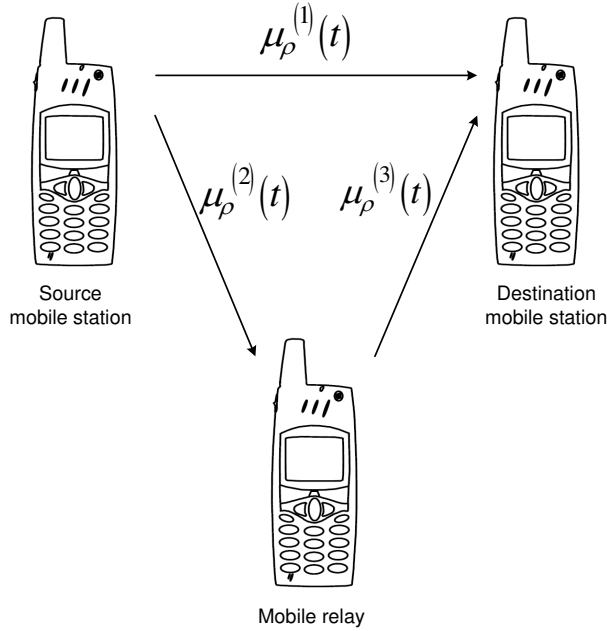


Fig. 1. The propagation scenario behind MLSS fading channels.

Starting from (1), ignoring $\mu^{(i)}(t) \forall i \geq 4$, and replacing $\mu^{(i)}(t)$ by $\mu_p^{(i)}(t)$ for $i = 1, 2, 3$, results in

$$\chi_p(t) = \mu_p^{(1)}(t) + A_{MR} \mu_p^{(2)}(t) \mu_p^{(3)}(t) \quad (2)$$

where $\alpha_1 = 1$ and $\alpha_2 = A_{MR}$. The quantity A_{MR} in (2) is referred to as the relay gain. Since we have assumed fixed gain relays in our system, it follows that A_{MR} is a real constant. Furthermore, in (2), $\mu_p^{(1)}(t)$, $\mu_p^{(2)}(t)$, and $\mu_p^{(3)}(t)$ are statistically independent non-zero-mean complex Gaussian processes, which model the individual M2M fading channel in the source mobile station to the mobile relay, the mobile relay to the destination mobile station, and the source mobile station to the destination mobile station links (see Fig. 1). Each complex Gaussian process $\mu_p^{(i)}(t) = \mu_{\rho_1}^{(i)}(t) + j\mu_{\rho_2}^{(i)}(t)$ represents the sum of the scattered component $\mu^{(i)}(t)$ and the LOS component $m^{(i)}(t)$, i.e., $\mu_p^{(i)}(t) = \mu^{(i)}(t) + m^{(i)}(t)$. The scattered component $\mu^{(i)}(t)$ is still modeled by a zero-mean complex Gaussian process $\mu^{(i)}(t) = \mu_1^{(i)}(t) + j\mu_2^{(i)}(t)$ with variance $2\sigma_i^2$. The LOS component $m^{(i)}(t) = \rho_i e^{j(2\pi f_{\rho_i} t + \theta_{\rho_i})}$ assumes a fixed amplitude ρ_i , a constant Doppler frequency f_{ρ_i} , and a constant phase θ_{ρ_i} for $i = 1, 2, 3$. Furthermore, it is obvious that

the product term $\mu_\rho^{(2)}(t) \mu_\rho^{(3)}(t)$ in (2) is a non-zero-mean complex double Gaussian process, i.e., $\mu_\rho^{(2)}(t) \mu_\rho^{(3)}(t) = \varsigma_\rho(t) = \varsigma_{\rho_1}(t) + j\varsigma_{\rho_2}(t)$. Furthermore, $\varsigma_\rho(t)$ models the overall fading in the source mobile station to the destination mobile station link via the mobile relay. It should also be noted here that the relay gain A_{MR} would just scale the mean value and variance of the complex Gaussian process $\mu_\rho^{(3)}(t)$, i.e., $m^{(3)}(t) = E\{A_{\text{MR}} \mu_\rho^{(3)}(t)\} = \rho_{A_{\text{MR}}} e^{j(2\pi f_{p_3} t + \theta_{p_3})}$ where $\rho_{A_{\text{MR}}} = A_{\text{MR}} \rho_3$, and $2\sigma_{A_{\text{MR}}}^2 = \text{Var}\{A_{\text{MR}} \mu_\rho^{(3)}(t)\} = 2(A_{\text{MR}} \sigma_3)^2$. Finally, the overall fading process consisting of the direct link between the source mobile station and the destination mobile station as well as the source mobile station to the destination mobile station link via the mobile relay results in the complex process $\chi_\rho(t) = \chi_{\rho_1}(t) + j\chi_{\rho_2}(t)$ given in (2). The absolute value of $\chi_\rho(t)$ defines the MLSS process, i.e., $\Xi(t) = |\chi_\rho(t)|$. Furthermore, the argument of $\chi_\rho(t)$ introduces the phase process $\Theta(t)$, i.e., $\Theta(t) = \arg\{\chi_\rho(t)\}$.

3. Statistical Analysis Of the MLSS Fading Channel

In this section, we derive the analytical expressions for the statistical properties of MLSS fading channels introduced in Section 2. The starting point for the derivation of the analytical expression of the PDF of MLSS fading channels, as well as the PDF of the corresponding phase process is the computation of the joint PDF $p_{\chi_{\rho_1} \chi_{\rho_2} \dot{\chi}_{\rho_1} \dot{\chi}_{\rho_2}}(u_1, u_2, \dot{u}_1, \dot{u}_2; t)$ of the stochastic processes $\chi_{\rho_1}(t)$, $\chi_{\rho_2}(t)$, $\dot{\chi}_{\rho_1}(t)$, and $\dot{\chi}_{\rho_2}(t)$ at the same time t . Throughout this chapter, the overdot indicates the time derivative. Equation (2) shows that the joint PDF $p_{\chi_{\rho_1} \chi_{\rho_2} \dot{\chi}_{\rho_1} \dot{\chi}_{\rho_2}}(u_1, u_2, \dot{u}_1, \dot{u}_2; t)$ can be written in terms of a 4-dimensional (4D) convolution integral as

$$p_{\chi_{\rho_1} \chi_{\rho_2} \dot{\chi}_{\rho_1} \dot{\chi}_{\rho_2}}(u_1, u_2, \dot{u}_1, \dot{u}_2; t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{\mu_{\rho_1}^{(1)} \mu_{\rho_2}^{(1)} \dot{\mu}_{\rho_1}^{(1)} \dot{\mu}_{\rho_2}^{(1)}}(u_1 - y_1, u_2 - y_2, \dot{u}_1 - \dot{y}_1, \dot{u}_2 - \dot{y}_2; t) p_{\varsigma_{\rho_1} \varsigma_{\rho_2} \dot{\varsigma}_{\rho_1} \dot{\varsigma}_{\rho_2}}(y_1, y_2, \dot{y}_1, \dot{y}_2; t) d\dot{y}_2 d\dot{y}_1 dy_2 dy_1 \quad (3)$$

where $p_{\mu_{\rho_1}^{(1)} \mu_{\rho_2}^{(1)} \dot{\mu}_{\rho_1}^{(1)} \dot{\mu}_{\rho_2}^{(1)}}(u_1, u_2, \dot{u}_1, \dot{u}_2; t)$ is the joint PDF of the processes $\mu_{\rho_1}^{(1)}(t)$, $\mu_{\rho_2}^{(1)}(t)$, $\dot{\mu}_{\rho_1}^{(1)}(t)$, and $\dot{\mu}_{\rho_2}^{(1)}(t)$ at the same time t . Similarly, $p_{\varsigma_{\rho_1} \varsigma_{\rho_2} \dot{\varsigma}_{\rho_1} \dot{\varsigma}_{\rho_2}}(y_1, y_2, \dot{y}_1, \dot{y}_2; t)$ in (3), represents the joint PDF of the processes $\varsigma_{\rho_1}(t)$, $\varsigma_{\rho_2}(t)$, $\dot{\varsigma}_{\rho_1}(t)$, and $\dot{\varsigma}_{\rho_2}(t)$ at the same time t . It is worth mentioning here that the processes $\mu_{\rho_i}^{(1)}(t)$, $\dot{\mu}_{\rho_i}^{(1)}(t)$, $\varsigma_{\rho_i}(t)$, and $\dot{\varsigma}_{\rho_i}(t)$ ($i = 1, 2$) are uncorrelated in pairs. Furthermore, the process pairs $\{\mu_{\rho_i}^{(1)}(t), \dot{\mu}_{\rho_i}^{(1)}(t)\}$ and $\{\varsigma_{\rho_i}(t), \dot{\varsigma}_{\rho_i}(t)\}$ ($i = 1, 2$) are statistically independent, which allows us to express $p_{\varsigma_{\rho_1} \varsigma_{\rho_2} \dot{\varsigma}_{\rho_1} \dot{\varsigma}_{\rho_2} \mu_{\rho_1}^{(1)} \mu_{\rho_2}^{(1)} \dot{\mu}_{\rho_1}^{(1)} \dot{\mu}_{\rho_2}^{(1)}}(y_1, y_2, \dot{y}_1, \dot{y}_2, u_1, u_2, \dot{u}_1, \dot{u}_2; t)$ as given in (3). The joint PDF $p_{\mu_{\rho_1}^{(1)} \mu_{\rho_2}^{(1)} \dot{\mu}_{\rho_1}^{(1)} \dot{\mu}_{\rho_2}^{(1)}}(u_1, u_2, \dot{u}_1, \dot{u}_2; t)$ can be obtained using the multivariate Gaussian distribution (see, e.g., (Simon, 2002, Eq. (3.2))). The expression for the joint PDF $p_{\varsigma_{\rho_1} \varsigma_{\rho_2} \dot{\varsigma}_{\rho_1} \dot{\varsigma}_{\rho_2}}(y_1, y_2, \dot{y}_1, \dot{y}_2; t)$ is presented in (4), where the quantity β_i ($i = 2, 3$) is the negative curvature of the autocorrelation function of the inphase and quadrature components of $\mu^{(i)}(t)$ ($i = 2, 3$). Under isotropic scattering conditions, β_i ($i = 2, 3$) can be expressed as (Akki, 1994; Pätzold, 2002)

$$\begin{aligned}
 p_{\zeta_{r1} \zeta_{r2} \zeta_{r1} \zeta_{r2}}(y_1, y_2, \dot{y}_1, \dot{y}_2; t) = & \\
 & - \frac{\left(z_1 - m_1^{(3)}(t) \right)^2 + \left(z_2 - m_2^{(3)}(t) \right)^2}{2\sigma_{A_{MR}}^2} e^{-\frac{\left(m_1^{(3)}(t) \right)^2 + \left(m_2^{(3)}(t) \right)^2}{2\beta_3}} \frac{e^{-\frac{2\beta_2 \left\{ \beta_3 (y_1^2 + y_2^2) + \beta_2 (z_1^2 + z_2^2) \right\}}{\beta_2 (z_1^2 + z_2^2)^2} \left\{ \left(m_1^{(3)}(t) \right)^2 + \left(m_2^{(3)}(t) \right)^2 \right\} + \beta_3 (y_1^2 + y_2^2) \left(m_1^{(2)}(t) + m_2^{(2)}(t) \right)}}{2\sigma_{A_{MR}}^2} \\
 & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dz_2 dz_1 \frac{(2\pi)^3 \sigma_{A_{MR}}^2 \left[\beta_3 (y_1^2 + y_2^2) + \beta_2 (z_1^2 + z_2^2) \right] e^{-\frac{2\beta_2 \left\{ \beta_3 (y_1^2 + y_2^2) + \beta_2 (z_1^2 + z_2^2) \right\}}{\beta_2 (z_1^2 + z_2^2)^2} \left\{ \left(m_1^{(3)}(t) \right)^2 + \left(m_2^{(3)}(t) \right)^2 \right\} + \beta_3 (y_1^2 + y_2^2) \left(m_1^{(2)}(t) + m_2^{(2)}(t) \right)}}{y_1^2 + y_2^2 - 2 \left(y_2 m_2^{(2)}(t) + y_1 m_1^{(2)}(t) \right) z_1 - 2 \left(y_2 m_1^{(2)}(t) - y_1 m_2^{(2)}(t) \right) z_2 + \left\{ \left(m_1^{(2)}(t) \right)^2 + \left(m_2^{(2)}(t) \right)^2 \right\} \left(z_1^2 + z_2^2 \right)} \\
 & \times e^{-\frac{y_1^2 + y_2^2 - 2 \left(y_2 m_2^{(2)}(t) + y_1 m_1^{(2)}(t) \right) z_1 - 2 \left(y_2 m_1^{(2)}(t) - y_1 m_2^{(2)}(t) \right) z_2 + \left\{ \left(m_1^{(2)}(t) \right)^2 + \left(m_2^{(2)}(t) \right)^2 \right\} \left(z_1^2 + z_2^2 \right)}{2\sigma_{A_{MR}}^2 \left(z_1^2 + z_2^2 \right)}} \\
 & \times e^{-\frac{y_1^2 + y_2^2 - 2 \left(y_2 m_2^{(2)}(t) + y_1 m_1^{(2)}(t) \right) z_1 - 2 \left(y_2 m_1^{(2)}(t) - y_1 m_2^{(2)}(t) \right) z_2 + \left\{ \left(m_1^{(2)}(t) \right)^2 + \left(m_2^{(2)}(t) \right)^2 \right\} \left(z_1^2 + z_2^2 \right)}{2\beta_2 \left(z_1^2 + z_2^2 \right)} \frac{e^{-\frac{\beta_3 (y_1^2 + y_2^2) + \beta_2 (z_1^2 + z_2^2)}{\beta_2 (z_1^2 + z_2^2)^2} \left\{ \left(m_1^{(3)}(t) \right)^2 + \left(m_2^{(3)}(t) \right)^2 \right\} + \beta_3 (y_1^2 + y_2^2) \left(m_1^{(2)}(t) + m_2^{(2)}(t) \right)}}{2\beta_2 \left(z_1^2 + z_2^2 \right) \left\{ m_1^{(3)}(t) \left(y_2 y_1 - y_1 y_2 + 2m_2^{(2)}(t) y_1 z_1 - 2m_1^{(2)}(t) y_2 z_1 \right) + m_2^{(3)}(t) \left(y_1 y_1 + y_2 y_2 - 2m_1^{(2)}(t) y_1 z_1 - 2m_2^{(2)}(t) y_2 z_1 \right) \right\} + \beta_3 (y_1^2 + y_2^2) \left(y_1^2 + y_2^2 \right)} \\
 & \times e^{-\frac{z_2^2 \left\{ m_1^{(2)}(t) m_1^{(3)}(t) y_1 + m_2^{(2)}(t) y_1 + m_2^{(2)}(t) m_1^{(3)}(t) y_2 - m_1^{(2)}(t) y_2 - m_1^{(3)}(t) m_2^{(2)}(t) y_2 \right\} - 2y_2 \left(m_2^{(2)}(t) z_1 - m_1^{(2)}(t) z_2 \right) + 2y_1 \left(-m_1^{(2)}(t) z_1 + m_2^{(2)}(t) y_2 \right)}{\beta_3 (y_1^2 + y_2^2) + \beta_2 (z_1^2 + z_2^2)^2} \frac{e^{-\frac{2\beta_2 \left(z_1^2 + z_2^2 \right)}{\beta_2 (z_1^2 + z_2^2)^2} \left\{ \beta_3 (y_1^2 + y_2^2) + \beta_2 (z_1^2 + z_2^2) \right\}}}{e} \\
 & \times e
 \end{aligned} \tag{4}$$

$$\begin{aligned}
p_{\Xi\Xi\Theta\Theta}(x, \dot{x}, \theta, \dot{\theta}; t) = & \frac{x^2 e^{-\frac{x^2}{2\sigma_1^2}} e^{-\frac{(m_1^{(2)}(t))^2 + (m_2^{(2)}(t))^2}{2\beta_2}} e^{-\frac{(m_1^{(3)}(t))^2 + (m_2^{(3)}(t))^2}{2\beta_3}} e^{-\frac{(m_1^{(1)}(t))^2 + (m_2^{(1)}(t))^2}{2\beta_1}}}{(2\pi)^4 \sigma_1^2 \sigma_2^2 v_{\text{AMR}}} \\
& \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dz_2 dz_1 dy_2 dy_1 \frac{e^{-\frac{(z_1 - m_1^{(3)}(t))^2 + (z_2 - m_2^{(3)}(t))^2}{2\sigma_2^2 v_{\text{AMR}}} e^{-\frac{(y_1 + m_1^{(1)}(t))^2 + (y_2 + m_2^{(1)}(t))^2}{2\sigma_1^2}} e^{-\frac{r(y_1 + m_1^{(1)}(t)) \cos \theta + r(y_2 + m_2^{(1)}(t)) \sin \theta}{\sigma_1^2}}}{\beta_1(z_1^2 + z_2^2) + \beta_3(y_1^2 + y_2^2) + \beta_2(z_1^2 + z_2^2)^2} \\
& \times e^{-\frac{y_1^2 + y_2^2 - 2(y_2 m_2^{(2)}(t) + y_1 m_1^{(2)}(t)) z_1 - 2(y_2 m_1^{(2)}(t) - y_1 m_2^{(2)}(t)) z_2 + \left\{ (m_1^{(2)}(t))^2 + (m_2^{(2)}(t))^2 \right\} \left\{ (m_1^{(1)}(t))^2 + (m_2^{(1)}(t))^2 \right\}}{2\sigma_2^2(z_1^2 + z_2^2)}} e^{2\beta_1 \left\{ \beta_1(z_1^2 + z_2^2) + \beta_3(y_1^2 + y_2^2) + \beta_2(z_1^2 + z_2^2)^2 \right\}} \\
& \times e^{\frac{\beta_2^2(z_1^2 + z_2^2)^2 \left[\beta_3 \left\{ (m_1^{(1)}(t))^2 + (m_2^{(1)}(t))^2 \right\} + \beta_1 \left\{ (m_1^{(3)}(t))^2 + (m_2^{(3)}(t))^2 \right\} + \beta_3 \beta_1 \left\{ (m_1^{(2)}(t))^2 + (m_2^{(2)}(t))^2 \right\} \right] + \beta_3(y_1^2 + y_2^2) + \beta_1(z_1^2 + z_2^2)}{2\beta_3 \beta_2 \beta_1 \left\{ \beta_1(z_1^2 + z_2^2) + \beta_3(y_1^2 + y_2^2) + \beta_2(z_1^2 + z_2^2)^2 \right\}}} \\
& \times e^{\frac{z_2^2 \left\{ m_1^{(2)}(t) m_1^{(3)}(t) y_1 + m_2^{(2)}(t) m_2^{(3)}(t) y_2 - m_1^{(2)}(t) y_2 - m_1^{(3)}(t) y_2 - z_1 \left(m_1^{(1)}(t) m_1^{(2)}(t) + m_2^{(1)}(t) m_2^{(2)}(t) \right) - z_2 \left(m_1^{(1)}(t) m_1^{(2)}(t) - m_1^{(1)}(t) m_2^{(2)}(t) \right) \right\}}{\beta_1(z_1^2 + z_2^2) + \beta_3(y_1^2 + y_2^2) + \beta_2(z_1^2 + z_2^2)^2}} \\
& \times e^{\frac{z_2 \left\{ 2z_1 \left(m_1^{(2)}(t) m_1^{(3)}(t) y_1 - m_1^{(2)}(t) y_1 - m_1^{(3)}(t) y_2 - m_2^{(2)}(t) m_2^{(3)}(t) y_2 \right) - z_1^2 \left(m_2^{(1)}(t) m_2^{(2)}(t) - m_1^{(1)}(t) m_2^{(2)}(t) \right) + \left(m_1^{(1)}(t) m_2^{(3)}(t) \right) y_1 \right\}}{\beta_1(z_1^2 + z_2^2) + \beta_3(y_1^2 + y_2^2) + \beta_2(z_1^2 + z_2^2)^2}} \\
& \times e^{\frac{-z_1^2 \left(m_1^{(2)}(t) m_2^{(3)}(t) y_2 - m_1^{(2)}(t) y_1 - m_2^{(2)}(t) m_1^{(3)}(t) y_2 \right) - z_1 \left(m_1^{(1)}(t) m_1^{(3)}(t) y_1 + m_2^{(1)}(t) y_2 - m_1^{(1)}(t) m_2^{(3)}(t) y_2 - m_1^{(1)}(t) m_2^{(2)}(t) \right) + \left(m_1^{(1)}(t) m_2^{(3)}(t) \right) y_2}{\beta_1(z_1^2 + z_2^2) + \beta_3(y_1^2 + y_2^2) + \beta_2(z_1^2 + z_2^2)^2}} \\
& \times e^{\frac{-z_1^3 \left(m_1^{(1)}(t) m_1^{(2)}(t) + m_2^{(1)}(t) + m_2^{(3)}(t) \right) (y_1 z_1 - y_1 z_2) + m_2^{(3)}(t) (y_1 z_1 + y_2 z_2) + \left(m_2^{(1)}(t) + m_2^{(2)}(t) z_1 + m_1^{(2)}(t) z_2 \right) (z_1^2 + z_2^2) \left\{ x \dot{\theta} \cos \theta + x \sin \theta \right\}}{\beta_1(z_1^2 + z_2^2) + \beta_3(y_1^2 + y_2^2) + \beta_2(z_1^2 + z_2^2)^2}} \\
& \times e^{\frac{\left\{ m_2^{(3)}(t) (y_1 z_2 - y_2 z_1) + m_1^{(3)}(t) (y_1 z_1 + y_2 z_2) + \left(m_1^{(1)}(t) + m_1^{(2)}(t) z_1 - m_2^{(2)}(t) z_2 \right) (z_1^2 + z_2^2) \left\{ x \cos \theta - x \dot{\theta} \sin \theta \right\} - \left(m_2^{(1)}(t) m_2^{(3)}(t) + m_1^{(1)}(t) m_2^{(3)}(t) \right) y_2 z_2 \right\}}{\beta_1(z_1^2 + z_2^2) + \beta_3(y_1^2 + y_2^2) + \beta_2(z_1^2 + z_2^2)^2}} \\
& \times e^{\frac{(z_1^2 + z_2^2) \left\{ x^2 + (x \dot{\theta})^2 \right\}}{\beta_1(z_1^2 + z_2^2) + \beta_3(y_1^2 + y_2^2) + \beta_2(z_1^2 + z_2^2)^2}} \\
& \times e^{\frac{2 \left\{ \beta_1(z_1^2 + z_2^2) + \beta_3(y_1^2 + y_2^2) + \beta_2(z_1^2 + z_2^2)^2 \right\}}{2\beta_3 \left\{ \beta_1(z_1^2 + z_2^2) + \beta_3(y_1^2 + y_2^2) + \beta_2(z_1^2 + z_2^2)^2 \right\}}}, \quad x \geq 0, |\dot{x}| < \infty, |\theta| \leq \pi, |\dot{\theta}| < \infty. \quad (5)
\end{aligned}$$

$$\beta_2 = 2(\sigma_2\pi)^2 (f_{\max_1}^2 + f_{\max_2}^2) \quad (5a)$$

$$\beta_3 = 2(\sigma_{A_{MR}}\pi)^2 (f_{\max_2}^2 + f_{\max_3}^2) . \quad (5b)$$

The symbols f_{\max_1} , f_{\max_2} , and f_{\max_3} appearing in (5a) and (5b) denote the maximum Doppler frequency caused by the motion of the source mobile station, the mobile relay, and the destination mobile station, respectively. Substituting the joint PDF $p_{\mu_{\rho_1}^{(1)} \mu_{\rho_2}^{(1)} \dot{\mu}_{\rho_1}^{(1)} \dot{\mu}_{\rho_2}^{(1)}}(u_1, u_2, \dot{u}_1, \dot{u}_2)$ and (4) in (3), applying the concept of transformation of random variables (Papoulis & Pillai, 2002), and doing tedious algebraic manipulations, the joint PDF $p_{\Xi \dot{\Xi} \Theta \dot{\Theta}}(x, \dot{x}, \theta, \dot{\theta}; t)$ of the processes $\Xi(t)$, $\dot{\Xi}(t)$, $\Theta(t)$, and $\dot{\Theta}(t)$ can be derived. The resulting joint PDF $p_{\Xi \dot{\Xi} \Theta \dot{\Theta}}(x, \dot{x}, \theta, \dot{\theta}; t)$ is presented in (5), which is of fundamental importance, because it provides the basis for the computation of the PDF, the level-crossing rate (LCR), and the average duration of fades (ADF) of MLSS processes $\Xi(t)$ as well as the PDF of the phase process $\Theta(t)$. Using (5), the analytical expressions of the LCR and the ADF of the MLSS processes $\Xi(t)$ have been derived in (Talha & Pätzold, 2008a). In (5), the quantity β_1 is given by $\beta_1 = 2(\sigma_1\pi)^2 (f_{\max_1}^2 + f_{\max_3}^2)$ (Akki, 1994; Pätzold, 2002).

3.1 PDF of the MLSS Process

The joint PDF $p_{\Xi\Theta}(x, \theta; t)$ of the MLSS process $\Xi(t)$ and the phase process $\Theta(t)$ can be obtained by solving the integrals over the joint PDF $p_{\Xi \dot{\Xi} \Theta \dot{\Theta}}(x, \dot{x}, \theta, \dot{\theta}; t)$ according to

$$p_{\Xi\Theta}(x, \theta; t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{\Xi \dot{\Xi} \Theta \dot{\Theta}}(x, \dot{x}, \theta, \dot{\theta}; t) d\dot{\theta} d\dot{x} \quad (6)$$

for $x \geq 0$ and $|\theta| \leq \pi$. Solving (6) results in the following expression

$$p_{\Xi\Theta}(x, \theta; t) = \frac{x e^{\frac{x^2}{2\sigma_1^2}}}{(2\pi)^2 \sigma_1^2 \sigma_2^2 \sigma_{A_{MR}}^2} \int_0^{\infty} \int_0^{\pi} \frac{\omega}{v} e^{-\frac{(\omega/v)^2 + \rho_2^2}{2\sigma_2^2}} e^{-\frac{v^2 + \rho_{A_{MR}}^2}{2\sigma_{A_{MR}}^2}} e^{-\frac{g_1(\omega, \psi; t)}{2\sigma_1^2}} e^{-\frac{x g_3(\omega, \theta, \psi; t)}{\sigma_1}} \times I_0 \left(\sqrt{g_2(\omega, v, \psi; t)} \right) d\psi d\omega dv, \quad x \geq 0, |\theta| \leq \pi \quad (7)$$

where

$$g_1(\omega, \psi; t) = \omega^2 + \rho_1^2 + 2\rho_1\omega \cos(\psi - 2\pi f_{\rho_1}t - \theta_{\rho_1}) \quad (8a)$$

$$g_2(\omega, v, \psi; t) = \left(\frac{\rho_2\omega}{\sigma_2^2 v} \right)^2 + \left(\frac{\rho_{A_{MR}}v}{\sigma_{A_{MR}}^2} \right)^2 + \frac{2\rho_2\rho_{A_{MR}}\omega}{\sigma_2^2 \sigma_{A_{MR}}^2} \cos[\psi - 2\pi(f_{\rho_2} + f_{\rho_3})t - (\theta_{\rho_2} + \theta_{\rho_3})] \quad (8b)$$

$$g_3(\omega, \theta, \psi; t) = \frac{\rho_1 \cos(\theta - 2\pi f_{\rho_1}t - \theta_{\rho_1}) + \omega \cos(\theta - \psi)}{\sigma_1} . \quad (8c)$$

In (7), $I_0(\cdot)$ is the zeroth-order modified Bessel function of the first kind (Gradshteyn & Ryzhik, 2000).

The PDF $p_{\Xi}(x)$ of the MLSS fading process $\Xi(t)$ can be obtained by integrating (7) over θ in the interval $[-\pi, \pi]$. Hence,

$$p_{\Xi}(x) = \frac{x e^{-\frac{x^2}{2\sigma_1^2}}}{2\pi\sigma_1^2\sigma_2^2\sigma_{A_{MR}}^2} \int_0^{\infty} \int_{-\pi}^{\pi} d\nu d\omega \frac{\omega}{\nu} e^{-\frac{(\omega/\nu)^2 + \rho_2^2}{2\sigma_2^2}} e^{-\frac{\nu^2 + \rho_{A_{MR}}^2}{2\sigma_{A_{MR}}^2}} \int_{-\pi}^{\pi} d\psi e^{-\frac{g_4(\omega, \psi)}{2\sigma_1^2}} I_0\left(\frac{x}{\sigma_1^2} \sqrt{g_4(\omega, \psi)}\right) I_0\left(\sqrt{g_5(\omega, \nu, \psi)}\right), x \geq 0 \quad (9)$$

where

$$g_4(\omega, \psi) = \omega^2 + \rho_1^2 + 2\rho_1\omega \cos(\psi) \quad (10a)$$

$$g_5(\omega, \nu, \psi) = \left(\frac{\rho_2\omega}{\sigma_2^2\nu}\right)^2 + \left(\frac{\rho_{A_{MR}}\nu}{\sigma_{A_{MR}}^2}\right)^2 + \frac{2\rho_2\rho_{A_{MR}}\omega}{\sigma_2^2\sigma_{A_{MR}}^2} \cos(\psi). \quad (10b)$$

It is worth mentioning that the joint PDF $p_{\Xi\Theta}(x, \theta; t)$ in (6) is dependent on time t . Nevertheless, the PDF $p_{\Xi}(x)$ in (9) is independent of time t showing that MLSS processes $\Xi(t)$ are first order stationary. From the PDF $p_{\Xi}(x)$ of MLSS fading processes $\Xi(t)$, the following special cases can be obtained.

Substituting $A_{MR} = 1$, $\rho_1 = \rho_2 = \rho_3 = 0$, and taking the limit $\sigma_1^2 \rightarrow 0$ in (9), reduces the PDF of MLSS processes to the PDF of double Rayleigh processes (see, e.g., (Kovacs et al., 2002; Patel et al., 2006))

$$p_{\Xi}(x) \Big|_{\substack{A_{MR}=1 \\ \rho_1, \rho_2, \rho_3=0 \\ \sigma_1^2 \rightarrow 0}} = \frac{x}{\sigma_2^2\sigma_{A_{MR}}^2} K_0\left(\frac{x}{\sigma_2\sigma_{A_{MR}}}\right), \quad x \geq 0. \quad (11)$$

For the special case when $A_{MR} = 1$, $\rho_1 = 0$, and $\sigma_1^2 \rightarrow 0$, the PDF of MLSS processes given in (9) reduces to the PDF of double Rice processes (Talha & Pätzold, 2007a)

$$p_{\Xi}(x) \Big|_{\substack{A_{MR}=1 \\ \rho_1=0 \\ \sigma_1^2 \rightarrow 0}} = \frac{x}{\sigma_2^2\sigma_{A_{MR}}^2} \int_0^{\infty} \frac{1}{\nu} e^{-\frac{(x/\nu)^2 + \rho_2^2}{2\sigma_2^2}} e^{-\frac{\nu^2 + \rho_{A_{MR}}^2}{2\sigma_{A_{MR}}^2}} I_0\left(\frac{x\rho_2}{\nu\sigma_2^2}\right) I_0\left(\frac{\nu\rho_{A_{MR}}}{\sigma_{A_{MR}}^2}\right) d\nu, \quad x \geq 0. \quad (12)$$

Similarly, substituting $A_{MR} = 1$, $\rho_2 = \rho_3 = 0$, and $\sigma_1^2 \rightarrow 0$ in (9) allows us to write the PDF of SLDS processes in the form (Talha & Pätzold, 2007b)

$$p_{\Xi}(x) \Big|_{\substack{A_{MR}=1 \\ \rho_2, \rho_3=0 \\ \sigma_1^2 \rightarrow 0}} = \begin{cases} \frac{x}{\sigma_2^2\sigma_{A_{MR}}^2} I_0\left(\frac{x}{\sigma_2\sigma_{A_{MR}}}\right) K_0\left(\frac{\rho_1}{\sigma_2\sigma_{A_{MR}}}\right), & x < \rho_1 \\ \frac{x}{\sigma_2^2\sigma_{A_{MR}}^2} K_0\left(\frac{x}{\sigma_2\sigma_{A_{MR}}}\right) I_0\left(\frac{\rho_1}{\sigma_2\sigma_{A_{MR}}}\right), & x \geq \rho_1 \end{cases} \quad (13)$$

The PDF of NLSS processes (see, e.g., (Salo et al., 2006)) can be derived from the PDF of MLSS processes by substituting $\rho_1 = \rho_2 = \rho_3 = 0$ in (9), i.e.,

$$p_{\Xi}(x) \Big|_{\rho_1, \rho_2, \rho_3=0}^{A_{MR}=1} = \frac{x}{\sigma_1^2 \sigma_2^2 \sigma_{A_{MR}0}^2} \int_0^{\infty} \omega e^{-\frac{\omega^2}{2\sigma_1^2}} K_0\left(\frac{\omega}{\sigma_2 \sigma_{A_{MR}}}\right) I_0\left(\frac{\omega}{\sigma_1^2} x\right) d\omega, \quad x \geq 0. \quad (14)$$

Finally, solving (9) for $\rho_2 = \rho_3 = 0$, the PDF of SLSS processes (Salo et al., 2006) is obtained as

$$p_{\Xi}(x) \Big|_{\rho_2, \rho_3=0}^{A_{MR}=1} = \frac{x}{2\pi \sigma_1^2 \sigma_2^2 \sigma_{A_{MR}0}^2} \int_0^{\pi} \int_0^{\pi} d\theta d\omega e^{-\frac{\omega^2}{2\sigma_1^2}} e^{-\frac{g_5(x, \theta)}{2\sigma_1^2}} K_0\left(\frac{\omega}{\sigma_2 \sigma_{A_{MR}}}\right) I_0\left(\frac{\omega}{\sigma_1^2} \sqrt{g_6(x, \theta)}\right), \quad x \geq 0 \quad (15)$$

where

$$g_6(x, \theta) = x^2 + \rho_1^2 - 2x\rho_1 \cos \theta. \quad (16)$$

3.2 PDF of the Phase Process

Integrating (7) over x in the interval $[0, \infty)$ results in the following expression for the PDF $p_{\Theta}(\theta; t)$ of the phase process $\Theta(t)$

$$p_{\Theta}(\theta; t) = \int_0^{\infty} \int_{-\pi}^{\pi} \frac{\omega}{v} e^{-\frac{(\omega/v)^2 + \rho_2^2}{2\sigma_2^2}} e^{-\frac{v^2 + \rho_{A_{MR}}^2}{2\sigma_{A_{MR}}^2}} e^{-\frac{g_1(\omega, \psi; t)}{2\sigma_1^2}} I_0\left(\sqrt{g_2(\omega, v, \psi; t)}\right) \left[1 + \sqrt{\frac{\pi}{2}} g_3(\omega, \theta, \psi; t)\right] \\ \times e^{\frac{1}{2} g_3^2(\omega, \theta, \psi; t)} \left\{1 + \Phi\left(\frac{g_3(\omega, \theta, \psi; t)}{\sqrt{2}}\right)\right\} d\psi d\omega dv, \quad |\theta| \leq \pi \quad (17)$$

where $g_1(\cdot, \cdot; t)$, $g_2(\cdot, \cdot, \cdot; t)$, and $g_3(\cdot, \cdot, \cdot; t)$ are defined in (8a), (8b), and (8c), respectively. In (17), $\Phi(\cdot)$ represents the error function (Gradshteyn & Ryzhik, 2000, Eq. (8.250.1)). From (17), it is obvious that the phase process $\Theta(t)$ is not strict sense stationary because the density $p_{\Theta}(\theta; t)$ is a function of time t . This time dependency is due the Doppler frequency f_{ρ_i} of the LOS component $m^{(i)}(t)$ ($i = 1, 2, 3$). However, for the special case when $f_{\rho_i} = 0$, $\rho_i \neq 0$ ($i = 1, 2, 3$), the phase process $\Theta(t)$ becomes first order stationary.

The PDF $p_{\Theta}(\theta; t)$ of the phase process $\Theta(t)$ given in (17) reduces to the PDF of the phase process corresponding to double Rayleigh (Talha & Pätzold, 2007a), double Rice (Talha & Pätzold, 2007a), SLDS (Talha & Pätzold, 2007b), NLSS, and SLSS processes by selecting ρ_1 , ρ_2 , ρ_3 , and σ_1^2 in a similar fashion as described in Subsection 3.1.

4. Numerical Results

In this section, we will provide sufficient evidence to support the validity of the analytical expressions presented in Section 3 with the help of simulations. Furthermore, for the sake of completeness, a detailed comparison of the PDF $p_{\Xi}(x)$ of MLSS processes $\Xi(t)$ to that of the special cases mentioned above will be presented. Similarly, a comparison of the PDF $p_{\Theta}(\theta)$ of the phase process $\Theta(t)$ with other phase PDFs will also be presented. The concept of sum-of-sinusoids (Pätzold, 2002) was employed to simulate the underlying uncorrelated Gaussian noise processes of the overall MLSS process $\Xi(t)$. The number of sinusoids required to simulate the inphase and quadrature components of Gaussian processes $\mu^{(i)}(t)$ was

selected to be 20 and 21, respectively. Furthermore, the simulation model parameters were computed using the generalized method of exact Doppler spread (GMEDS₁) (Pätzold & Hogstad, 2006). The maximum Doppler frequencies, i.e., $f_{\max 1}$, $f_{\max 2}$, and $f_{\max 3}$ were set to 91 Hz, 75 Hz, and 110 Hz, respectively. Furthermore, for simplicity, the amplitudes of the three LOS components ρ_1 , ρ_2 , and ρ_3 are assumed to be equal, i.e., $\rho_1 = \rho_2 = \rho_3 = \rho$. The quantities σ_1^2 , σ_2^2 , σ_3^2 , and the relay gain A_{MR} are set to 1, unless stated otherwise.

The results presented in Figs. 2 and 3 show a good fitting between the analytical and simulation results. In Fig. 2, the PDF $p_{\Xi}(x)$ of MLSS processes $\Xi(t)$ is compared to that of classical Rayleigh, classical Rice, double Rayleigh, double Rice, SLDS, NLSS, and SLSS processes for $\rho = 1$, where f_{ρ_1} , f_{ρ_2} , and f_{ρ_3} were set to 166 Hz, 185 Hz, and 201 Hz, respectively. It can be observed that the maximum value of the PDF $p_{\Xi}(x)$ of MLSS processes $\Xi(t)$ is lower than that of all the other processes under consideration for the same value of ρ . However, the PDF $p_{\Xi}(x)$ of MLSS processes $\Xi(t)$ has a higher spread as compared to the spreads of the above mentioned processes for the same value of ρ . The same trend can be seen when different values of ρ_i ($i = 1, 2, 3$) are selected. Furthermore, increasing the value of the relay gain A_{MR} causes a decrease in the maximum value and an increase in the spread of the PDF of MLSS processes $\Xi(t)$.

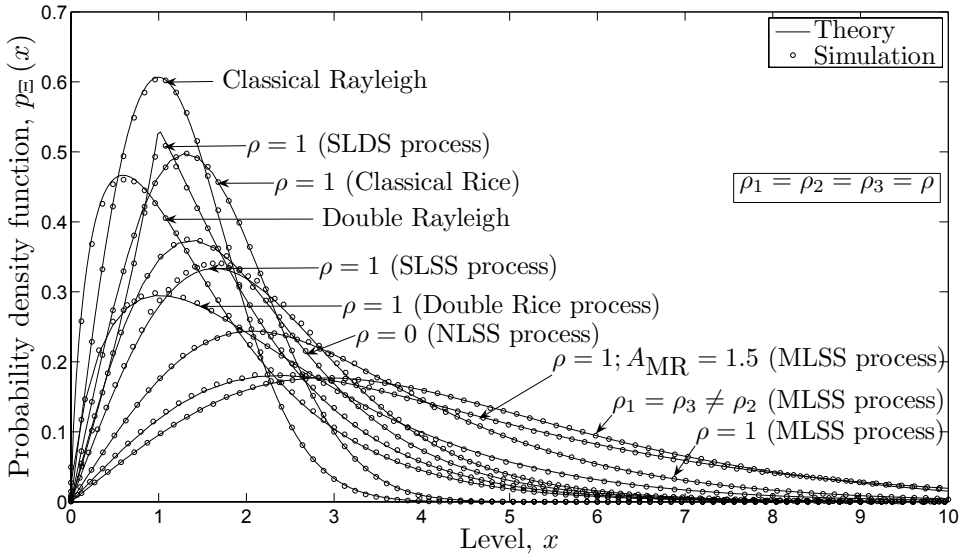


Fig. 2. A comparison of the PDF $p_{\Xi}(x)$ of the MLSS process $\Xi(t)$ with that of various other stochastic processes.

Figure 3 presents a comparison of the PDF $p_{\Theta}(\theta)$ of the phase process $\Theta(t)$ with that of the phase processes corresponding to the classical Rayleigh and double Rayleigh processes, the classical Rice and double Rice processes, the SLDS process, the NLSS process, and the SLSS process for $\rho = 1$. It should be noted that the results presented in Fig. 3 are valid for the case when f_{ρ_i} ($i = 1, 2, 3$) is set to zero. It can be observed that the PDF of the SLDS phase

process has the highest peak. The PDF $p_{\Theta}(\theta)$ of the phase process $\Theta(t)$ follows the same trend in terms of the maximum value and the spread as that of the classical Rice process for the same value of ρ . Furthermore, it is interesting to note that for the phase process $\Theta(t)$ with ρ_i ($i = 1, 2, 3$) being selected as $\rho_1 = \rho_3 = 0.5$ and $\rho_2 = 1$, the PDF $p_{\Theta}(\theta)$ is almost the same as that of the double Rice process for $\rho = 1$. Figure 3 also shows the impact of the relay gain A_{MR} on the PDF $p_{\Theta}(\theta)$ of the phase process $\Theta(t)$.

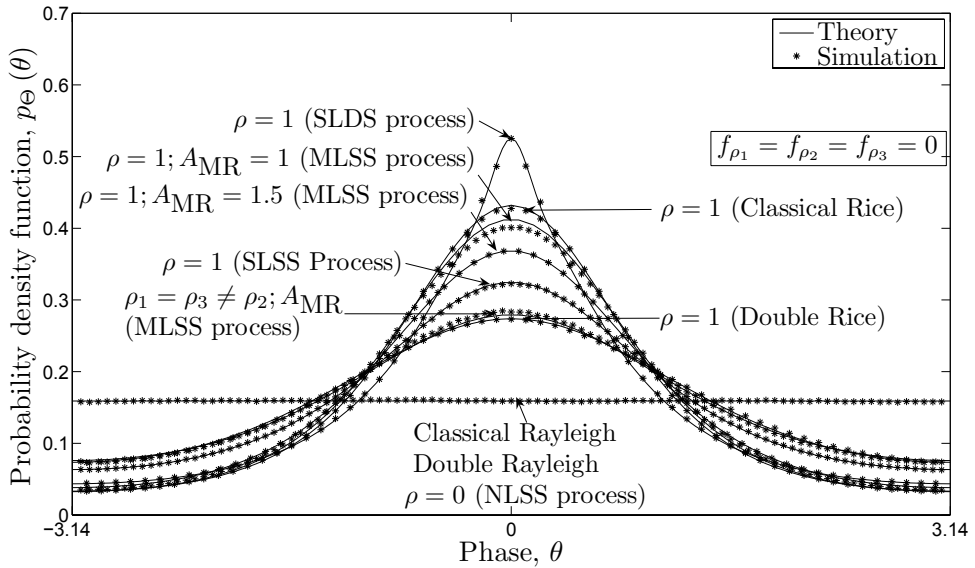


Fig. 3. A comparison of the PDF $p_{\Theta}(\theta)$ of the phase process $\Theta(t)$ with that of various other various processes.

5. Conclusion

In this chapter, we have proposed a new flexible M2M amplify-and-forward relay fading channel model under LOS propagation conditions. The novelty in the model is that we have considered the LOS components in both transmission links, i.e., the direct link between the source mobile station and the destination mobile station as well as the link via the mobile relay. By analogy with multiple scattering radio propagation channels, we have developed the MLSS fading channel as a second-order scattering channel, where the received signal comprises the single and double scattered components. Furthermore, the flexibility of the MLSS fading channel model comes from the fact that it can be reduced to double Rayleigh, double Rice, SLDS, NLSS, and SLSS channel models under certain assumptions.

This chapter also presents a deep analysis of the statistical properties of MLSS fading channels. The statistical properties studied, include the PDF of MLSS processes along with the PDF of the corresponding phase processes. Accurate analytical expressions have been derived for the above mentioned phase statistical quantities. The accuracy and validity of the analytical expressions are confirmed by simulations. The excellent fitting of the theoretical

and simulation results verifies the correctness of the derived analytical expressions. The presented results show that the statistical properties of MLSS channels are quite different from those of the processes embedded in the MLSS channel model as special cases. It is also evident from the illustrated results that the relay gain has a significant impact on the statistical properties of MLSS channels.

The statistics of a fading channel dictate the choice of the transmitter and the receiver techniques, including the detection, modulation, and coding schemes, etc. Therefore, the theoretical results presented in this chapter are quite useful for the designers of the physical layer of M2M cooperative wireless networks. Furthermore, the developed M2M channel model can be employed to investigate the overall system performance of M2M communication systems under both NLOS and LOS propagation conditions.

6. References

- Akki, A. S. (1994). Statistical properties of mobile-to-mobile land communication channels, *IEEE Trans. Veh. Technol.* **Vol. 43**(No. 4): 826–831.
- Almers, P., Tufvesson, F. & Molisch, A. F. (2006). Keyhole effect in MIMO wireless channels: measurements and theory, *IEEE Trans. Wireless Commun.* **Vol. 5**(No. 12): 3596–3604.
- Andersen, J. B. (2002). Statistical distributions in mobile communications using multiple scattering, *Proc. 27th URSI General Assembly*, Maastricht, Netherlands.
- Dohler, M. (2003). *Virtual Antenna Arrays*, Ph.D. dissertation, King's College, London, United Kingdom.
- Erceg, V., Fortune, S. J., Ling, J., Rustako, Jr, A. J. & Valenzuela, R. A. (1997). Comparison of a computer-based propagation prediction tool with experimental data collected in urban microcellular environment, *IEEE J. Select. Areas Commun.* **Vol. 15**(No. 4): 677–684.
- Gradshteyn, I. S. & Ryzhik, I. M. (2000). *Table of Integrals, Series, and Products*, 6th edn, New York: Academic Press.
- Kovacs, I. Z., Eggers, P. C. F., Olesen, K. & Petersen, L. G. (2002). Investigations of outdoor-to-indoor mobile-to-mobile radio communication channels, *Proc. IEEE 56th Veh. Technol. Conf., VTC'02-Fall*, Vol. 1, Vancouver BC, Canada, pp. 430–434.
- Laneman, J. N., Tse, D. N. C. & Wornell, G. W. (2004). Cooperative diversity in wireless networks: efficient protocols and outage behavior, *IEEE Trans. Inform. Theory* **Vol. 50**(No. 12): 3062–3080.
- Papoulis, A. & Pillai, S. U. (2002). *Probability, Random Variables and Stochastic Processes*, 4th edn, New York: McGraw-Hill.
- Patel, C. S., Stüber, G. L. & Pratt, T. G. (2006). Statistical properties of amplify and forward relay fading channels, *IEEE Trans. Veh. Technol.* **Vol. 55**(No. 1): 1–9.
- Pätzold, M. (2002). *Mobile Fading Channels*, Chichester: John Wiley & Sons.
- Pätzold, M. & Hogstad, B. O. (2006). Two new methods for the generation of multiple uncorrelated Rayleigh fading waveforms, *Proc. IEEE 63rd Semianual Veh. Tech. Conf., VTC'06-Spring*, Vol. 6, Melbourne, Australia, pp. 2782–2786.
- Salo, J., El-Sallabi, H. M. & Vainikainen, P. (2006). Statistical analysis of the multiple scattering radio channel, *IEEE Trans. Antennas Propagat.* **Vol. 54**(No. 11): 3114–3124.
- Sendonaris, A., Erkip, E. & Aazhang, B. (2003a). User cooperation diversity — Part I: System description, *IEEE Trans. Commun.* **Vol. 51**(No. 11): 1927–1938.

- Sendonaris, A., Erkip, E. & Aazhang, B. (2003b). User cooperation diversity — Part II: Implementation aspects and performance analysis, *IEEE Trans. Commun.* **Vol. 51**(No. 11): 1939–1948.
- Shin, H. & Lee, J. H. (2004). Performance analysis of space-time block codes over keyhole Nakagami-m fading channels, *IEEE Trans. Veh. Technol.* **Vol. 53**(No. 2): 351–362.
- Simon, M. K. (2002). *Probability Distributions Involving Gaussian Random Variables: A Handbook for Engineers and Scientists*, Dordrecht: Kluwer Academic Publishers.
- Talha, B. & Pätzold, M. (2007a). On the statistical properties of double Rice channels, *Proc. 10th International Symposium on Wireless Personal Multimedia Communications, WPMC 2007*, Jaipur, India, pp. 517–522.
- Talha, B. & Pätzold, M. (2007b). On the statistical properties of mobile-to-mobile fading channels in cooperative networks under line-of-sight conditions, *Proc. 10th International Symposium on Wireless Personal Multimedia Communications, WPMC 2007*, Jaipur, India, pp. 388–393.
- Talha, B. & Pätzold, M. (2008a). Level-crossing rate and average duration of fades of the envelope of mobile-to-mobile fading channels in cooperative networks under line-of-sight conditions, *Proc. 51st IEEE Globecom 2008*, New Orleans, USA, pp. 1–6. DOI 10.1109/GLOCOM.2008.ECP.860.
- Talha, B. & Pätzold, M. (2008b). A novel amplify-and-forward relay channel model for mobile-to-mobile fading channels under line-of-sight conditions, *Proc. 19th IEEE Int. Symp. on Personal, Indoor and Mobile Radio Communications, PIMRC 2008*, Cannes, France, pp. 1–6. DOI 10.1109/PIMRC.2008.4699733.

Resource Management with Limited Capability of Fixed Relay Station in Multi-hop Cellular Networks

Jemin Lee and Daesik Hong
Yonsei University
Republic of Korea

1. Introduction

The purpose of this chapter is to develop a resource management technique to utilize resource efficiently in multi-hop cellular networks. Multi-hop cellular networks have been proposed as a way to enhance throughput and extend coverage (Cho & Haas, 2004). This enhancement can in general be achieved by deploying relay stations in conventional cellular networks. The advantage of multi-hop networks arises from the reduction in the overall path loss achieved by using a relay station between a base station and a mobile station. Moreover, deeply shadowed mobile stations can be supported by using relay stations to bypass obstacles.

Even though the multi-hop transmission has advantages, it also carries a penalty: the need for additional resources to transmit data in multi-hop manner (for example, in two-hop transmission, two time slots or frequency channels for the base station-relay station link and the relay station-mobile station link) (Lee *et al.*, 2008). Hence, this penalty of multi-hop transmission is represented by 'worms', which devour resources (Ju *et al.*, 2009).

The trade-offs associated with the multi-hop networks make it difficult to assess their overall performance. For mobile stations with quality-of-service (QoS) requirements guaranteed by single-hop transmission with few resources, multi-hop transmission could end up wasting additional resources through multiple hops, even though it may provide higher end-to-end data rates. In other words, a higher end-to-end data rate does not guarantee higher efficiency in resource utilization. Hence, the amount of resources required to guarantee QoS should be considered when assessing multi-hop transmission performance and the performance can be different depending on a mobile station.

In addition, the infrastructure cost increases almost linearly with the equipment capability (Johansson *et al.*, 2004), and it is impossible to change the capability of the installed equipment flexibly according to the change of the required capability. Hence, a relay station has the limited and determined capability. Due to the limited capability, some of mobile stations cannot transmit in multi-hop if all capability of relay station has been already fully used for the other mobile stations. Hence, the resource management for assigning the

limited capability to the mobile station, who can take more advantage of multi-hop transmission, is required.

In this chapter, for utilizing the resources efficiently, the resource management considering both the different multi-hop gains of each mobile stations and the limited capability of relay station is provided. First of all, a brief explanation about multi-hop cellular networks is provided in Section 2. Then, the transmission mode selection is discussed as a way to determine whether multi-hop or single-hop transmission is the transmission mode most appropriate for minimizing the resources used to guarantee a certain QoS in Section 3. The multi-hop gain is defined as the amount of resources saved by using multi-hop transmission instead of single-hop transmission, and the elements which affect the multi-hop gain are discussed. Based on the affecting elements, two criteria for transmission mode selection are provided and the performance of them is also verified.

In Section 4, the multi-hop user admission is discussed as a way to determine which mobile station should be admitted or rejected to transmit data in multi-hop for maximizing the achievable multi-hop gain using the limited capability of relay station. The multi-hop user admission is formulated as a multi-dimensional knapsack problem, and two efficient heuristic algorithms for multi-hop user admission are introduced and the performance of those algorithms is discussed.

Finally, the structure of resource management including the transmission mode selection and the multi-hop user admission is provided in Section 5.

2. Multi-hop Cellular Networks

The network under consideration in this chapter is a downlink orthogonal frequency division multiple access (OFDMA) multi-hop cellular network. The multi-hop system adopted here is the two-hop relaying system, which is known to be the most efficient multi-hop system with respect to system capacity (Cho & Haas, 2004). In this system, data can be transmitted from a base station to a mobile station in one of two transmission modes: single-hop transmission or multi-hop transmission via a fixed relay station.

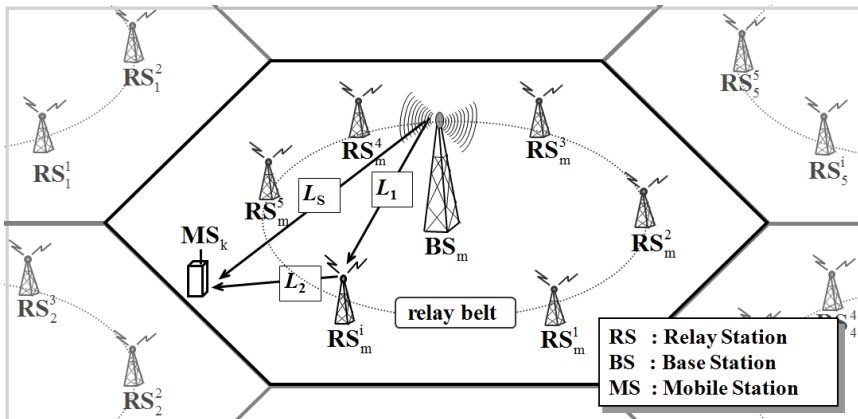


Fig. 1. Downlink multi-hop cellular networks

Fig. 1 shows an example of the connectivity for this system. In Fig. 1, the k th mobile station, MS_k , is connected with the m th base station, BS_m , and MS_k uses the i th fixed relay station in the m th cell, RS_m^i , for multi-hop transmission. A number of fixed relay stations are placed on the relay belt. All the fixed relay stations are regenerative relays, so they decode data received from the base station and then forward it to the target mobile stations. In this system, three kinds of links are formed. The base station-fixed relay station links and the fixed relay station-mobile station links occurring with two-hop transmission are denoted by the link for the first hop (L_1) and the link for the second hop (L_2), respectively. The link for single-hop transmission (L_s) also denotes the base station-mobile station link. Generally, it is assumed that L_1 has good channel condition for a line-of-sight (LOS) environment, and L_2 and L_s are in a non line-of-sight (NLOS) environment (Liu *et al.*, 2006). The LOS assumption can be satisfied by deploying fixed relay stations at selected locations, such as on top of a building.

As QoS parameters which can be handled in the physical layer, the target data rate and the target bit-error-rate (BER) can be considered. In the regenerative relay, errors generated at each hop are propagated to the next hop. The sum of the target bit-error-rates for each hop in two-hop transmission should therefore be equal to or less than the target bit-error-rate in two-hop transmission, B^T , as $\sum_{i=1}^2 B_{L_i}^T \leq B^T$ where $B_{L_i}^T$ is the target bit-error-rate on the link of the i th hop (Boyer *et al.*, 2004).

In addition, the end-to-end data rate from a base station to a mobile station is determined by the minimum data rate among the rates in each hop (Jing *et al.*, 2005). Hence, the target data rate for each hop should be equal to or greater than the target data rate, R^T , as $R_{L_i}^T \geq R^T$ where $R_{L_i}^T$ is the target data rate on the link of the i th hop.

In OFDMA systems, each subcarrier can obtain a different channel gain. So, the received signal to interference and noise ratio (SINR) on the n th subcarrier of the link L ($L \in \{L_s, L_1, L_2\}$) can be defined as

$$\Gamma_L(n) = \frac{G_L(n) \cdot P_I}{\sum_{j \neq m} \alpha_j(n) \cdot G_{I_j}(n) \cdot P_I + \eta}, \quad \forall L, \quad (1)$$

where η is the additive Gaussian noise power and $G_L(n)$ is the channel gain of link L on the n th subcarrier. I_j is the link between an interferer in the j th cell and the target node, and P_I is the transmission power of the transmitter on that link.

If it is assumed that every link in a cell utilizes different resources and resources used for transmission on a link are also reused at the same link in other cells concurrently, then the link $BS_m - RS_m^i$ and the link $BS_j - RS_j^i$ ($j \neq m$) use the same resources. Hence, in single-hop transmission, L is the link $BS_m - MS_k$, I_j is the link $BS_j - MS_k$, and $P_I = P_{BS}$, where P_{BS} is the transmission power of the base station in (1). In addition, in two-hop transmission, L is the link $BS_m - RS_m^i$, I_j is the link $BS_j - RS_m^i$, and $P_I = P_{BS}$ on the first hop. L is the link $RS_m^i -$

MS_k , I_j is the link $RS_j^i - MS_k$, and $P_j = P_{FRS}$ on the second hop where P_{FRS} is the transmission power of the fixed relay station.

Some of subcarriers are not being used when the mobile stations do not require all of subcarriers in OFDMA systems. Hence, $\alpha_j(n)$ in (1) is adopted to express the loading state of the n the subcarrier in the j th cell. $\alpha_j(n) = 1$ if the n th subcarrier is used to transfer data in the j th cell, while $\alpha_j(n) = 0$ otherwise. Therefore, the average loading state of the j th cell, ρ_j , is defined as

$$\rho_j = \frac{1}{N} \sum_{n=1}^N \alpha_j(n), \quad (2)$$

where N is the total number of subcarriers in a cell. In addition, ρ_j becomes one when there is full loading.

With adaptive modulation coding (AMC), the throughput on the link L is expressed as the function of $\Gamma_L(n)$ and B_L^T , as follows:

$$\begin{aligned} R_L(n) &= f(B_L^T, \Gamma_L(n)) \\ &= \frac{1}{T_s} \log_2 \left(1 + \frac{-1.5}{\ln(5 \cdot B_L^T)} \cdot \Gamma_L(n) \right), \forall L, n, \end{aligned} \quad (3)$$

where T_s is the symbol duration (Qiu and Chawla, 1999). The number of subcarriers on link L required to guarantee the QoS for MS_k , $C_{k,L}$, is determined by R^T and $R_L(n)$. For instance, when the channel gains of all subcarriers are equal and the average loading state for the other cells is one, then $R_L(1) = R_L(2) = \dots = R_L$, and $C_{k,L}$ can be defined as

$$C_{k,L} = \lceil R^T / R_L \rceil \forall k, L, \quad (4)$$

where $\lceil \chi \rceil$ is the least integer equal to or greater than χ . Hence, the total numbers of subcarriers of MS_k required to transmit data with a QoS guarantee by single-hop transmission and multi-hop transmission, C_k^{SH} and C_k^{MH} , are respectively defined as

$$C_k^{SH} = C_{k,L_S}, \quad C_k^{MH} = \sum_{i=1}^2 C_{k,L_i}, \forall k. \quad (5)$$

A fixed relay station has a limited capability due to the cost, the limitation of power amplifier and so on. In this chapter, the number of supportable subcarriers is considered as the capability of the fixed relay station. It means that the i th fixed relay station can support up to $N_{i,RX}$ subcarriers for receiving on L_1 (the link between the base station and the relay station) and $N_{i,TX}$ subcarriers for transmitting on L_2 (the link between the relay station and the mobile station) in a time. Hence, the total number of required subcarriers of each link should be equal or fewer than the supportable number of subcarriers of a fixed relay station in each link as follows:

$$\sum_{k \in MU_i} C_{k,L_1} \leq N_{i,RX}, \quad \sum_{k \in MU_i} C_{k,L_2} \leq N_{i,TX}, \quad (6)$$

where \overline{MU}_i is the set of mobile stations who want to use the i th fixed relay station in transmission. Fixed relay stations which have higher capabilities can generally support more subcarriers with higher total power.

3. Transmission Model Selection: Multi-hop vs. Single-hop

The multi-hop transmission needs for additional resources, but it achieves more reliable transmission than the single-hop transmission. Due to this trade-off, the multi-hop transmission cannot be better than the single-hop transmission for all users. Hence, the elements, which affect the performance of multi-hop transmission, are investigated and the transmission mode selection for more efficient resource utilization is discussed in this section.

3.1 Achievable Gain from Multi-hop Transmission

With respect to efficiency of resources, the gain associated with multi-hop transmission is achieved when subcarriers are saved by transmitting in multi-hop instead of in single-hop. Hence, the multi-hop gain, S_k , can be defined as the relative ratio of the amount of saved subcarriers to the number of required subcarriers in single-hop transmission as follows (Lee *et al.*, 2008):

$$S_k = \frac{C_k^{SH} - C_k^{MH}}{C_k^{SH}}, \quad \forall k. \quad (7)$$

This approach shows the amount of gain or loss with multi-hop transmission. Thus, if S_k has a positive value, that means that the multi-hop transmission is saving subcarriers with guaranteeing the QoS requirements for the k th mobile station. On the other hand, if the value of S_k is negative, that implies that the multi-hop transmission is wasting subcarriers.

Fig. 2 presents the multi-hop gain in various environments with the assumption of the equal average loading states of other cells as $\rho_j = \rho, \forall j \neq m$ in (2). The distance between the k th mobile station and the base station is denoted by d_k and the cell radius is d_{cell} . System parameters in Table 1 are used for simulations. Fig. 2 and the formulas from (1) to (5) show that multi-hop gain is affected by three elements: loading state of other cells, location of the mobile station, and QoS requirements. Due to the long transmission distance in single-hop transmission, C_{k,L_S} is generally greater than C_{k,L_1} or C_{k,L_2} . However, for the multi-hop transmission, subcarriers for multi-hops, the sum of C_{k,L_1} and C_{k,L_2} , should be used. By this relation, the multi-hop gain is affected by the loading states of other cells and the location of the mobile station. As the loading state of other cells increases or if the mobile station is located near the cell boundary, the SINRs of the three links decrease. At a lower SINR, the achievable data rate is more sensitive to variation of SINR due to the log function as in (3). This means that the farther the SINR falls, the faster the number of subcarriers required for guaranteeing QoS increases. Hence, in this environment, C_{k,L_S} increases more rapidly than the sum of C_{k,L_1} and C_{k,L_2} , so that a bigger multi-hop gain can be achieved.

System Parameters	Values
System bandwidth	5MHz
Number of subcarriers	1024
Path loss exponent (LOS/NLOS)	2.35 / 3.76
Standard deviation of shadowing (LOS/NLOS)	3.4 dB / 8 dB
Transmission power of base station / fixed relay station	43 dBm / 40 dBm
Cell radius / Radius of the relay belt	500 m / 250 m
Number of cells	7
Number of fixed relay stations per cell	6, Symmetrically located on the relay belt
Number of mobile stations per cell	200, Uniformly distributed
Modulation order	BPSK, QPSK, 16-,64-,128- QAM
Power control	Equal power allocation

Table 1. System parameters (for simulations of Fig. 2, Fig. 3, Fig. 5 and Fig. 6)

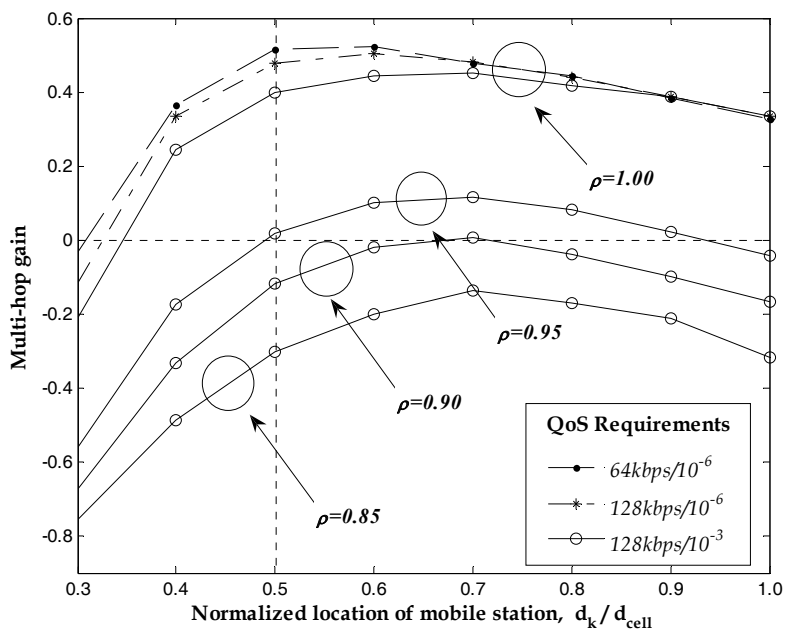


Fig. 2. Multi-hop gain based on QoS requirements (target data rate/target BER), loading state of other cells (ρ) and mobile station location (the mobile station's QoS parameters are 128kbps / 10^{-3} for all cases except for the case of $\rho = 1.0$.)

However, this increasing multi-hop gain begins to decrease when d_k is above a certain value (e.g., $d_k / d_{cell} \geq 0.7$ when $\rho = 0.95$ in Fig. 2). The reason for this is that C_{k,L_2} is also increasing rapidly due to larger interference as the mobile station moves to the cell edge, causing the difference between C_k^{SH} and C_k^{MH} to get smaller.

On the other hand, as the mobile station approaches the base station or the loading states of other cells decrease, the SINRs for all three links are increasing. When the SINR has increased enough, the required number of subcarriers becomes small and it is no longer sensitive to the variation in SINR. Hence, in this environment, all of C_{k,L_S} , C_{k,L_1} and C_{k,L_2} are small, so that the multi-hop transmission would waste subcarriers because of the usage of subcarriers for two hops. For this reason, the multi-hop transmission does not have any gain as the loading states of other cells decrease or the mobile station approaches the base station (e.g., $d_k / d_{cell} \leq 0.5$ when $\rho = 0.95$ in Fig. 2).

In addition, comparing the upper three lines in Fig. 2 shows that the multi-hop gain has different values depending on the target data rate and the target BER. Even the multi-hop gains for these three cases become similar when the mobile station approaches the cell boundary because of the large amount of interference, a lower target BER induces a higher multi-hop gain because the reliable transmission is more important to mobile stations requiring lower target BERs. The multi-hop gain also varies depending on the target data rate due to the subcarrier allocation process.

In this subsection, the affecting elements on the multi-hop gain, the loading state of other cells, the location of mobile station, and QoS requirement, have been discussed. Since the multi-hop transmission can save (or waste) resources according to those elements, the multi-hop transmission should be used selectively for efficient utilization of resources. Hence, transmission mode selection is required to determine the appropriate transmission mode for each mobile station: multi-hop transmission or single-hop transmission.

3.2 Mechanism for Transmission Mode Selection

The number of required subcarriers changes depending on the transmission mode, the QoS requirements, and the channel condition. To save subcarriers, whichever transmission mode requires fewer subcarriers should be the one selected. Hence, the subcarrier-based criterion, $\Lambda_{k,S}$, is defined as follows (Lee *et al.*, 2008):

$$\Lambda_{k,S} = C_k^{SH} - C_k^{MH}, \forall k. \quad (8)$$

If $\Lambda_{k,S}$ is greater than zero, then the selected transmission mode becomes the multi-hop transmission; otherwise, single-hop transmission mode is selected.

In the case where the SINRs for each subcarrier are different, C_k^{SH} and C_k^{MH} cannot be calculated before the transmission mode is determined because the frequency bands for single-hop and multi-hop transmission may be different. The subcarrier allocation which determines the number of required subcarriers should therefore be performed after determination of the transmission mode. In this case, the average number of required subcarriers $\overline{C_k^\varphi}$ can be used instead of C_k^φ to calculate $\Lambda_{k,S}$ where $\varphi \in \{SH, MH\}$. $\overline{C_k^{SH}}$ and

$\overline{C_k^{MH}}$ are respectively defined as $\overline{C_k^{SH}} = \left\lceil R^T / \overline{R_{L_S}} \right\rceil$, and $\overline{C_k^{MH}} = \sum_{i=1}^2 \left\lceil R^T / \overline{R_{L_i}} \right\rceil, \forall k$, where $\overline{R_L} = (1/N) \cdot \sum_{n=1}^N R_L(n)$.

As simpler way to select transmission mode, the distance based criterion can be used. As discussed in Section 3.1, the multi-hop gain changes depending on the locations of mobile stations. Positive gain could be obtained when a mobile station is located near the base station. Hence, the transmission mode can be determined using the following criterion (Lee *et al.*, 2008):

$$\Lambda_{k,D} = d_k - d_{ref}, \forall k. \quad (9)$$

In (9), d_k is $\sqrt{(x_0 - x_k)^2 + (y_0 - y_k)^2}$ where (x_0, y_0) and (x_k, y_k) are the coordinates (x and y) of a base station's location and the mobile station's location obtained by a positioning system (e.g., global positioning system), respectively, and d_{ref} is the reference distance for the transmission mode selection. If d_k is greater than d_{ref} , $\Lambda_{k,D}$ has a positive value and the multi-hop transmission is selected. This means that the multi-hop transmission is applied for mobile stations located outside of the circle with radius d_{ref} .

The performance of the transmission mode selection can be expressed as blocking probability, which is the ratio of the number of unsupportable mobile stations, K_U , to the total number of mobile stations which access the system, K_T , or K_U / K_T as Fig. 3. In this figure, TMS-S and TMS-D represent the transmission mode selections using the subcarrier-based criterion ($\Lambda_{k,S}$) and the distance-based criterion ($\Lambda_{k,D}$), respectively.

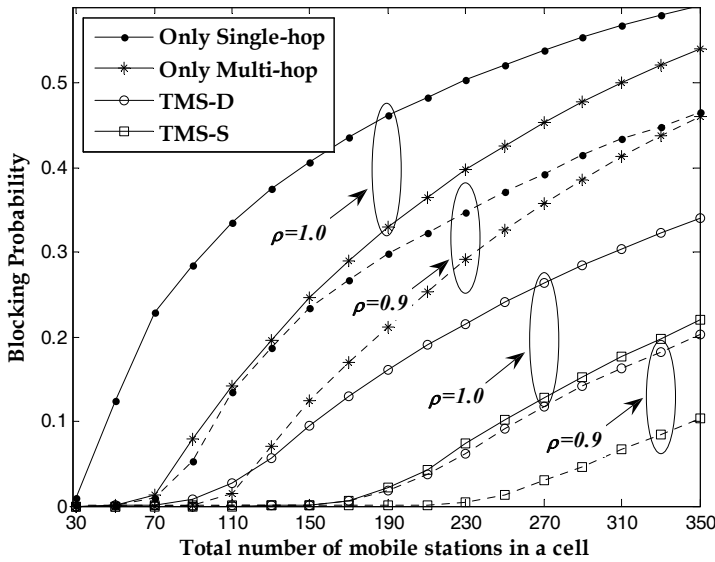


Fig. 3. Comparison of blocking probabilities based on total number of mobile stations in a cell

In Fig. 3, TMS-D and TMS-S demonstrate better performance, regardless of the loading states of other cells, than the cases where either single-hop transmission or the multi-hop transmission is applied without the selection process. This means that more mobile stations can be supported in a cell with lower blocking probability when the transmission mode selection is applied. Specifically, when the loading states of other cells are 1.0 and the blocking probability is 0.1, only 45 mobile stations and 95 mobile stations can be supported using conventional single-hop transmission and the multi-hop transmission, respectively. The number of supportable mobile stations increases to 150 with TMS-D and 250 with TMS-S.

4. Multi-hop User Admission for Relay Stations with Limited Capabilities

Based on the transmission mode selection in Section 3.2, *the multi-hop user* and *the single-hop user*, which denote the mobile stations which select the multi-hop transmission and the single-hop transmission, respectively, are determined. Since all fixed relay stations have the limited capabilities, a fixed relay station may not be able to support all of multi-hop users. When the required subcarriers of multi-hop users are beyond the number of supportable subcarriers in (6), some multi-hop users cannot receive data in multi-hop. Hence, to utilize the limited capability of the fixed relay station efficiently, the multi-hop user admission is required to allow the multi-hop users, which achieve high multi-hop gain, to use a fixed relay station. Therefore, in this section, the problem of multi-hop user admission is discussed and the multi-hop user admission algorithms are presented.

4.1 Formulation of Multi-hop User Admission

The multi-hop user admission strategy is formulated to determine the admitted multi-hop users to use the fixed relay station among all multi-hop users for maximizing the total multi-hop gains which can be obtained from the admitted multi-hop users as follows:

$$\begin{aligned}
 \max \quad & \sum_{i=1}^F \sum_{k \in MU_i} x_k \cdot S_k \\
 \text{s.t} \quad & \Omega_{1,i} : \sum_{k \in MU_i} x_k \cdot C_{k,L_1} \leq N_{i,RX}, \quad \forall i \\
 & \Omega_{2,i} : \sum_{k \in MU_i} x_k \cdot C_{k,L_2} \leq N_{i,TX}, \quad \forall i \\
 & \Omega_3 : x_k \in \{0,1\}, \quad \forall k,
 \end{aligned} \tag{10}$$

where F is the total number of fixed relay stations in a cell, and the x_k is the indicator of admission. If MS_k is admitted to transmit in multi-hop using the fixed relay station, then $x_k = 1$, otherwise $x_k = 0$ and MS_k should transmit in single-hop.

For each fixed relay station, there are two constraints from the limited capabilities: $\Omega_{1,i}$ and $\Omega_{2,i}$ based on the i th fixed relay station. The Lagrangian of multi-hop user admission problem, $Lag(u, u')$, can be defined as follows:

$$\begin{aligned}
 Lag(u, u') &= \sum_{i=1}^F \left(\sum_{k \in MU_i} x_k \cdot S_k - u_i \sum_{k \in MU_i} (x_k \cdot C_{k,L_1} - N_{i,RX}) - u'_i \sum_{k \in MU_i} (x_k \cdot C_{k,L_2} - N_{i,TX}) \right) \\
 &= \sum_{i=1}^F Lag(u_i, u'_i)
 \end{aligned} \tag{11}$$

where u_i and u'_i are Lagrangian multipliers. Hence, it is shown that the primal problem can be decomposed into sub-problems $Z(u_i, u'_i)$ where $Z(u_i, u'_i) = \max Lag(u_i, u'_i)$. It means that the parallel multi-hop user admissions for each fixed relay station can work independently since the result of one fixed relay station's multi-hop user admission does not affect the other fixed relay stations' multi-hop user admission. Therefore, we can deal with the multi-hop user admission problem by decomposing into the independent multi-hop user admission problems for each fixed relay station, which can be formulated as follows:

$$\begin{aligned} \max \quad & \sum_{k \in MU_i} x_k \cdot S_k \\ \text{s.t} \quad & \Omega_{1,i}, \Omega_{2,i}, \text{ and } \Omega_3. \end{aligned} \quad (12)$$

The problem in (12) has the equivalent form with the two-dimensional knapsack problem (TDKP) which is a kind of multi-dimensional knapsack problem (MDKP) (Qiu & Chawla, 1999). The MDKP is a variant of the classical 0-1 knapsack problem (KP) with more than two knapsacks.

4.2 Algorithms for Multi-hop User Admission

As shown in Section 4.1, the multi-hop user admission strategy can be represented as the two-dimensional knapsack problem a kind of MDKP. The KP and the MDKP are proven to be NP-hard, so the optimal solutions of them cannot be obtained in a polynomial time (Akbar *et al.*, 2005). Hence, a heuristic algorithm could be used for multi-hop user admission, and two multi-hop user admission algorithms are presented in this section.

4.2.1 Balanced Link Multi-hop User Admission (BL-MUA) Algorithm

In the multi-hop user admission, the balance between the used resources on L_1 and those on L_2 is important. The reason for this is that no additional multi-hop users can be admitted when at least one of $N_{i,RX}$ on L_1 and $N_{i,TX}$ on L_2 is fully occupied for other multi-hop users. Hence, the balanced link multi-hop user admission (BL-MUA) algorithm has been proposed in (Lee *et al.*, 2007) considering the balance in admission by adopting the primal effective gradient method (PEGM) (Toyoda, 1975).

The PEGM determines the priority of admission using the new measurement of the aggregate resource. The aggregate resource is to penalize the multi-hop user which requires many subcarriers in the more loaded link. The *aggregate resource* in multi-hop user admission can be defined as $A_k = (C_{k,L_1} \cdot l_1 + C_{k,L_2} \cdot l_2) / \sqrt{l_1^2 + l_2^2}$ where l_1 and l_2 are the total numbers of used subcarriers on L_1 and L_2 by the currently admitted multi-hop users, respectively. The priority function of the BL-MUA algorithm, $U_{BL,k}$, is set by the multi-hop gain over the aggregate resources, S_k / A_k , as follows:

$$U_{BL,k} = \frac{S_k \sqrt{l_1^2 + l_2^2}}{C_{k,L_1} \cdot l_1 + C_{k,L_2} \cdot l_2}. \quad (11)$$

If at least one of $N_{i,TX}$ and $N_{i,RX}$ is insufficient for supporting all multi-hop users in \overline{MU}_i , the multi-hop user with the lowest priority is rejected to use the fixed relay station one by one. The priority function helps to balance between the loading states of two links while rejecting the multi-hop user.

4. 2. 2 Focused Link Multi-hop User Admission (FL-MUA) Algorithm

The BL-MUA algorithm attaches importance to the balance of used resources in L_1 and L_2 . However, if fixed relay stations are located in a LOS environment with the base station and the environment of L_2 is a NLOS (like a general assumption), then the number of required subcarriers to guarantee the same target data rate in L_2 is much more than that in L_1 , $C_{k,L_1} < C_{k,L_2}$. This means that more multi-hop users could not be admitted in a fixed relay station generally due to the full loading of L_2 , not that of L_1 (the supportable subcarriers in L_2 is exhausted quickly than that in L_1).

Therefore, the multi-hop user admission considering both loading states of L_1 and L_2 can be simplified to that considering only that of L_2 . In this case, the multi-hop user admission strategy becomes a simple knapsack problem, not the two-dimensional knapsack problem anymore. Hence, the focused link multi-hop user admission (FL-MUA) algorithm is proposed to focus only on the loading state of L_2 (Lee *et al.*, 2007). As the priority function in the FL-MUA algorithm, the multi-hop gain per the average number of required subcarriers in L_2 is used as

$$U_{FL,k} = S_k / C_{k,L_2}. \quad (12)$$

When the supportable subcarriers of a fixed relay station are not sufficient, the multi-hop users are excluded in a low-priority order one by one.

4. 2. 3 Procedure of multi-hop user admission algorithms

The process of the multi-hop user admission algorithms progresses independently for each fixed relay station, and the overall procedure is shown by the flow chart in Fig. 4. If multi-hop users which want to use the i th fixed relay station exist, the total number of required subcarriers for supporting all multi-hop users in L_1 and L_2 , l_1^o and l_2^o , can be calculated as

$$l_1^o = \sum_{k \in \overline{MU}_i} C_{k,L_1}, \quad l_2^o = \sum_{k \in \overline{MU}_i} C_{k,L_2}. \quad (13)$$

If $l_1^o > N_{i,RX}$ or $l_2^o > N_{i,TX}$, then the multi-hop user admission algorithm starts. First of all, the priority values of all multi-hop users are obtained based on the priority function in (11) or (12). After that, the multi-hop user with the lowest priority is selected and the indicator of admission is set to zero. The selected multi-hop user should do single-hop transmission, so the multi-hop user is excluded from the \overline{MU}_i and added to the set of single-hop users, \overline{SU} . If one or both of the capabilities in the first hop and the second hop, $N_{i,RX}$ and $N_{i,TX}$, are still insufficient after one multi-hop user exclusion, the multi-hop user admission process

progresses again. At this time, the used resources, l_1 and l_2 , are changed due to the exclusion of a multi-hop user, so the priority function of the BL-MUA algorithm should be updated. On the other hand, in the FL-MUA algorithm, the updating process is not required and the algorithm just keeps using the previously determined priorities, so it is simpler than the BL-MUA algorithm. The process of the multi-hop user admission continuously works until both of $N_{i,RX}$ and $N_{i,TX}$ are sufficient for supporting all multi-hop users in \overline{MU}_i .

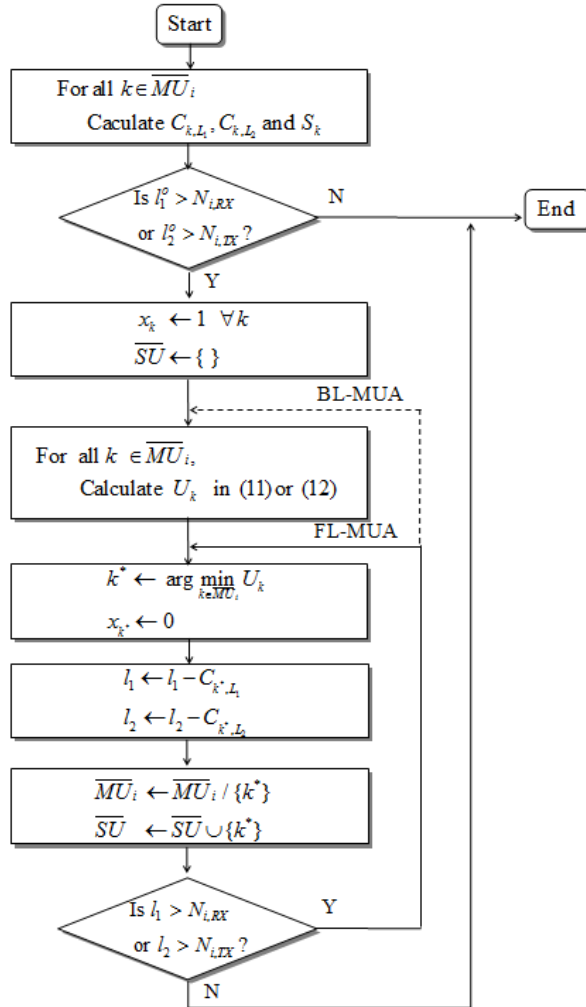


Fig. 4. Flow chart of the multi-hop user admission algorithms (BL-MUA algorithm and FL-MUA algorithm)

4.3 Performance of Multi-hop User Admission Algorithms

In this section, the performance of the BL-MUA algorithm and that of the FL-MUA algorithm are evaluated with the assumption that the deployed fixed relay stations have the same capabilities as N_R and all capabilities for transmitting and receiving are equal as

$$N_{i,RX} = N_{i,TX} = N_R, \forall i.$$

The performances of the multi-hop user algorithms are verified with two types of fixed relay stations: the fixed relay station with low capability (L-FRS) and that with high capability (H-FRS). The numbers of supportable subcarriers per fixed relay station are set to 16 for L-FRS and 64 for H-FRS, and the average loading state of other cells is set to one. The other system parameters are the same as Table 1. In addition, the performances of the multi-hop user admission algorithms are compared to the case where multi-hop users are randomly admitted without an admission algorithm (w/o MUA in Fig. 5).

In both multi-hop user admission algorithms, with the higher priority, the supportable subcarriers of a fixed relay station are used for the mobile stations which occupy fewer subcarriers with higher multi-hop gain. Hence, the number of admitted multi-hop users in a fixed relay station can be increased by the algorithms. Those are verified in Fig. 5.

Fig. 5 presents the number of admitted multi-hop users in a fixed relay station according to two types of fixed relay stations. More multi-hop users can be supported in a fixed relay station by the multi-hop user admission algorithms within the limited capability of fixed relay station compared to the case without multi-hop user admission algorithm.

Moreover, the performance difference between the multi-hop user admission algorithms and the case without the algorithm becomes more significant as the number of multi-hop users increases. When the number of multi-hop users is small, the capability of fixed relay station is generally sufficient for supporting all multi-hop users. Hence, the multi-hop user admission is not actually required and the performance of the multi-hop user admission algorithms is similar to the case without the algorithms. However, the multi-hop user admission becomes meaningful when the fixed relay station cannot support all multi-hop users because of the insufficient capability.

In addition, the performance of BL-MUA algorithm and that of FL-MUA algorithm are similar. The reason for this is from the physical characteristics of L_1 and L_2 . The number of required subcarriers in L_2 is much more than that in L_1 as $C_{k,L_2} > C_{k,L_1}$.

In this case, the priority function of the BL-MUA algorithm in (11) can be approximated to the priority function of the FL-MUA algorithm in (12) as

$$U_{BL,k} = S_k \sqrt{l_1^2 + l_2^2} / (C_{k,L_1} \cdot l_1 + C_{k,L_2} \cdot l_2) \approx S_k / C_{k,L_2} = U_{FL,k}.$$

Thus, the similar priority functions are used in both algorithms, so the total numbers of admitted users of them do not have big difference. It implies that the FL-MUA algorithm could obtain similar performance to the BL-MUA algorithm with less complexity.

In addition, in the aspect of total capacity in a cell, more mobile stations can be supported with guaranteeing their QoS requirements regardless of single-hop or multi-hop transmissions using the multi-hop user admission algorithms. This can be verified by Fig. 6 which shows the blocking probability as the number of mobile stations per cell increases.

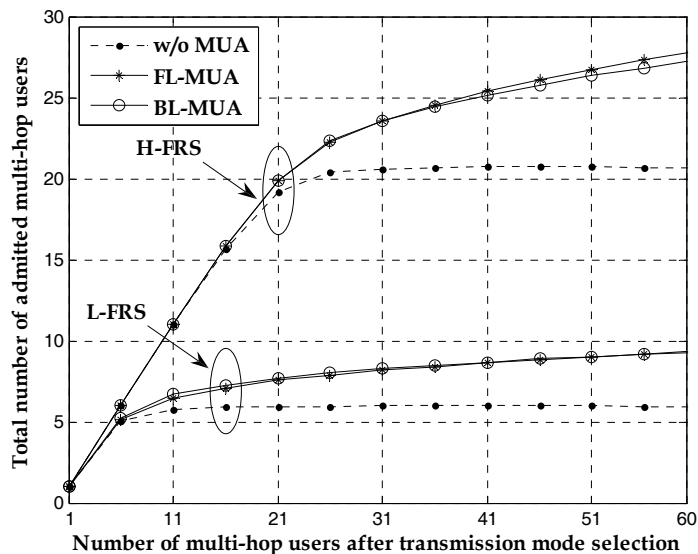


Fig. 5. The total multi-hop gain obtained from the admitted multi-hop users in a fixed relay station according to two types of fixed relay stations

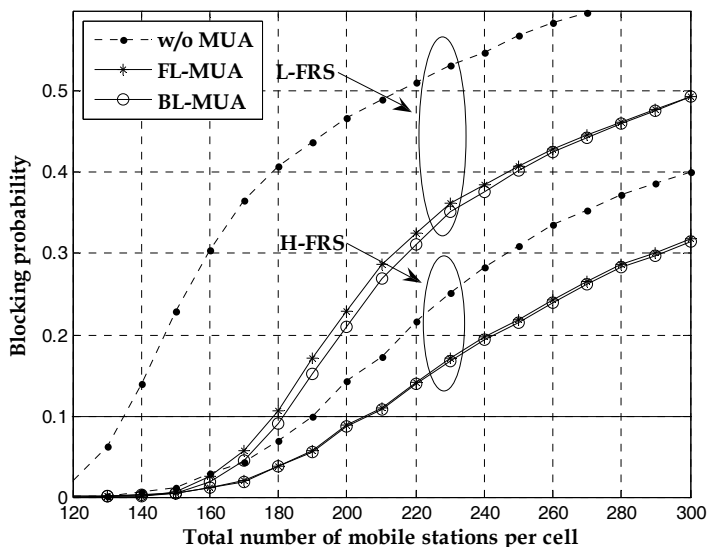


Fig. 6. The blocking probabilities according to two types of fixed relay stations

The blocking probabilities of the multi-hop user admission algorithms are smaller than that of the case without the algorithm over all range. Specifically, within 0.1 blocking probability, the case without the algorithm can support at most 135 mobile stations with L-FRSs and 185 mobile stations with H-FRSs. On the other hand, the numbers of supportable mobile stations are increased up to 180 mobile stations with L-FRSs and 223 mobile stations with H-FRSs by the multi-hop user admission algorithms. It implies that 33% and 20% of the supportable mobile stations in a cell are increased by the multi-hop user admission algorithms with L-FRSs and H-FRSs, respectively. Thus, more mobile stations can be supported in a cell with low blocking probability by the multi-hop user admission algorithms.

5. Overall Structure of Resource Management in Multi-hop Cellular Networks

In Section 3 and Section 4, the transmission mode selection and the multi-hop user admission have been discussed. The transmission mode determined through the transmission mode selection can be changed after the multi-hop user admission. The reason for this is that some mobile stations should transmit in single-hop due to the limitation of the fixed relay station's capability even though they select the multi-hop transmission in the transmission mode selection. Therefore, the final transmission mode determination and the resource allocation should be performed after the multi-hop user admission. The overall process of the resource management including the transmission mode selection and the multi-hop user admission is summarized as follows:

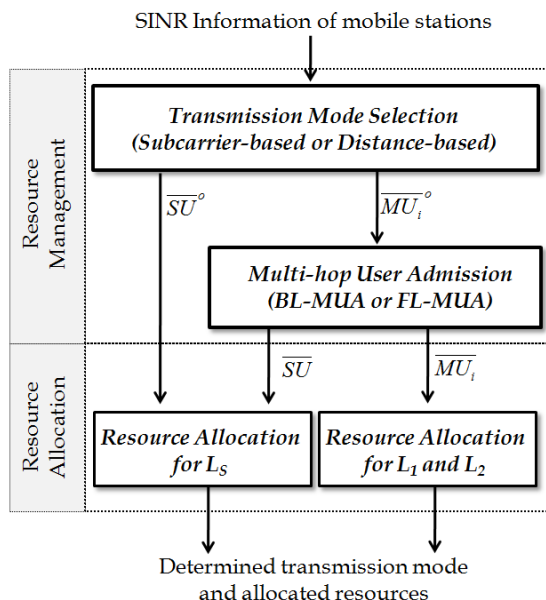


Fig. 7. The overall structure of resource management in multi-hop cellular networks

1) Once the mobile station set is defined and the required SINR information of the mobile stations are collected in a base station, the resource management process can start. The required SINR information is determined according to the criterion of the transmission mode selection and it could include the received SINRs of L_1 , L_2 and L_S .

2) During the transmission mode selection with a specific criterion (subcarrier-based or distance-based criterion), the set of multi-hop users, \overline{MU}_i^o , and the set of single-hop users, \overline{SU}^o , are determined.

3) After determining the set of multi-hop users, the multi-hop user admission can be performed. Through a multi-hop user admission algorithm (BL-MUA or FL-MUA algorithm), the mobile stations who can finally receive data in multi-hop, \overline{MU}_i and the mobile stations who should receive data in single-hop, \overline{SU} , due to the lack of the capability of the fixed relay station are determined.

4) The single-hop users in the sets of \overline{SU}^o and \overline{SU} are forwarded to the resource allocation process based on SINR information of L_S , and the multi-hop users in the set of \overline{MU}_i undergoes the resource allocation process based on SINR information of L_1 and L_2 .

After all of those processes are progressed, the final transmission mode which can maximize the multi-hop gain within the limited capability can be obtained.

According to the transmission mode selection criterion and the multi-hop user admission algorithm, the complexity of this resource management could be changed. In the transmission mode selection, the complexity of the transmission mode selection with the subcarrier-based criterion is $O(K_T N (\log p) p^2)$ where p is the required number of digits used in the operations such as square root and multiplication. It is higher than the complexity of transmission mode selection with the distance-based criterion, $O(K_T p^2)$ (Lee *et al.*, 2008). Hence, the subcarrier-based selection criterion's complexity is higher than the distance-based selection criterion, but it achieves better performance since it consider more elements for determining an appropriate transmission mode as shown in Section 3.

In the multi-hop user admission, the FL-MUA algorithm is simpler than the BL-MUA algorithm while the BL-MUA algorithm could achieve better performance as shown in Section 4.

Therefore, designers can use appropriate selection criterion and multi-hop user admission algorithm based on what the acceptable complexity level for that system is.

6. Conclusion

This chapter provides the resource management for efficient resource utilization in multi-hop cellular networks. The resource management has two parts: the transmission mode selection and the multi-hop user admission. The transmission mode selection is a way to select an appropriate transmission mode between the multi-hop transmission and the single-hop transmission for saving resources with guaranteeing QoS requirements of mobile stations. Two kinds of selection criteria, subcarrier-based and distance-based criterion, are provided after discussing the elements which affect the multi-hop gains such as the QoS requirements, the mobile station's location, and the loading states of other cells.

However, due to the limited capability of a relay station, some mobile stations cannot transmit data in multi-hop even though they select the multi-hop transmission mode. Hence, the multi-hop user admission is provided as a way to assign the limited capability of a relay station to the mobile stations which can maximize the multi-hop gain. Since the multi-hop user admission strategy is a NP-hard problem, two heuristic algorithms are provided: the BL-MUA algorithm focused on the load balance between L_1 and L_2 and the FL-MUA algorithm focused only on the load state of L_2 . Through the transmission mode selection and the multi-hop user admission, the resources can be used efficiently with supporting more mobile stations with lower blocking probability.

7. Acknowledgement

This work was supported by Korea Science and Engineering Foundation through the NRL Program (Grant R0A-2007-000-20043-0).

8. References

- Akbar, M. M.; Rahman, M. S.; Kaykobaad M. & Manning, E. G. (2005). Solving the Multidimensional Multiple-choice Knapsack Problem by Constructing Convex Hulls. *Computers & operations research*, Vol. 33, No. 5, pp. 1259-1273, ISSN 0305-0548.
- Boyer, J.; Falconer, D. D. & Yanikomeroglu, H. (2004). Multihop Diversity in Wireless Relaying Channels. *IEEE Transactions on Communications*, Vol. 52, No. 10, pp. 1820-1830, ISSN 0090-6778.
- Cho, J. & Haas, Z. J. (2004). On the Throughput Enhancement of the Downstream Channel in Cellular Radio Networks through Multihop Relaying. *IEEE Journal on Selected Areas in Communications*, Vol. 22, No. 7, pp. 1206-1219, ISSN 0733-8716.
- Jing, S.; Zhao-yang, Z.; Pei-liang, Q. & Guan-ding, Y. (2005). Subcarrier and Power Allocation for OFDMA-Based Regenerative Multihop Links. *International Conference on Wireless Communications, Networking and Mobile Computing*, Vol. 1, pp. 207-210.
- Johansson, K.; Furuskar, A.; Karlsson, P. & Zander, J. (2004). Relation between Base Station Characteristics and Cost Structure in Cellular Systems. *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*.
- Ju, H.; Oh, E. & Hong, D. (2009). Catching Resource-Devouring Worms in Next-Generation Wireless Relay Systems - Two-way Relay and Full-Duplex Relay. *IEEE Communications Magazine*, Vol. 47, No. 9, pp. 58-65, ISSN 0163-6804.
- Lee, J.; Wang, H.; Lim, S. & Hong, D. (2007). A Multi-hop User Admission Algorithm for Fixed Relay Stations with Limited Capabilities in OFDMA Cellular Networks. *IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1-5, ISBN 978-1-4244-1144-3, Athens, Sept. 2007.
- Lee, J.; Wang, H.; Seo, W. & Hong, D. (2008). QoS-guaranteed Transmission Mode Selection for Efficient Resource Utilization in Multi-hop Cellular Networks. *IEEE Transactions on Wireless Communications*, Vol. 7, Issue 10, pp. 3697-3701, ISSN 1536-1276.
- Liu, T. et al. (2006). Radio Resource Allocation in Two-hop Cellular Relaying Network, *IEEE Vehicular Technology Conference*, Vol. 1, pp. 91-95, ISBN 1-7803-9392-9, May 2006.

- Qiu, X. & Chawla, K. (1999). On the Performance of Adaptive Modulation in Cellular Systems. *IEEE Transactions on Communications*, Vol. 47, No. 6, pp. 884-895, ISSN 0090-6778.
- Toyoda, Y. (1975). A Simplified Algorithm for Obtaining Approximate Solution to Zero-one Programming Problems. *Management Science*, Vol. 21, No. 12, pp. 1417-1427, ISSN 0025-1909.

On Cross-layer Routing in Wireless Multi-Hop Networks

Golnaz Karbaschi¹, Anne Fladenmuller² and Sébastien Baey²

¹*Institut National de Recherche en Informatique et en Automatique (INRIA) – Saclay*

²*Université Pierre et Marie Curie (UPMC) – Paris
France*

1. Introduction

Wireless multi-hop networks represent a fundamental step in the evolution of wireless communications. Several new applications of such networks have recently emerged including community wireless networks, last-mile access for people, instant surveillance systems and back-haul service for large-scale wireless sensor networks, local high-speed P2P networking, or connectivity to rural/remote sites which was previously limited by cables.

Wireless multi-hop networks consist of computers and devices (nodes), which are connected by wireless communication channel, denoted as links. Since a wireless communication has a limited range, many pairs of node cannot communicate directly, and must forward data to each other via one or more cooperating intermediate nodes. Thus, in a unicast routing the source node transmits its packets to a neighboring node with which it can communicate directly. The neighboring node in turn transmits the packets to one of its neighbors and so on until the packets reach their final destination. Each node that forwards the packets are referred to as a hop and the set of the links, which are selected to transfer the packets, are called the route. Different routes from any nodes to any destinations are discovered by a distributed routing protocol in the network. Figure 1 shows an example of a wireless mesh network in which node S2 sends data traffic to the destination D via cooperation of intermediate nodes R1 and R2, while the other source node S1 sends out its data traffic to the gateway via the node A.

Wireless multi-hop networks are self-expanding networks; connectivity of the network is only due to existence of the nodes, thus the network can be expanded or decreased simply by adding or removing a node. In contrast, cellular networks are a much more expensive infrastructure since they need at least one base station to provide connectivity. Moreover, the capacity of the base station is limited and so not all the nodes in the coverage area can be connected to the network. Therefore, wireless multi-hop networks are a promising solution to expand the network easily as they allow flexibility and rapid deployment at low cost.

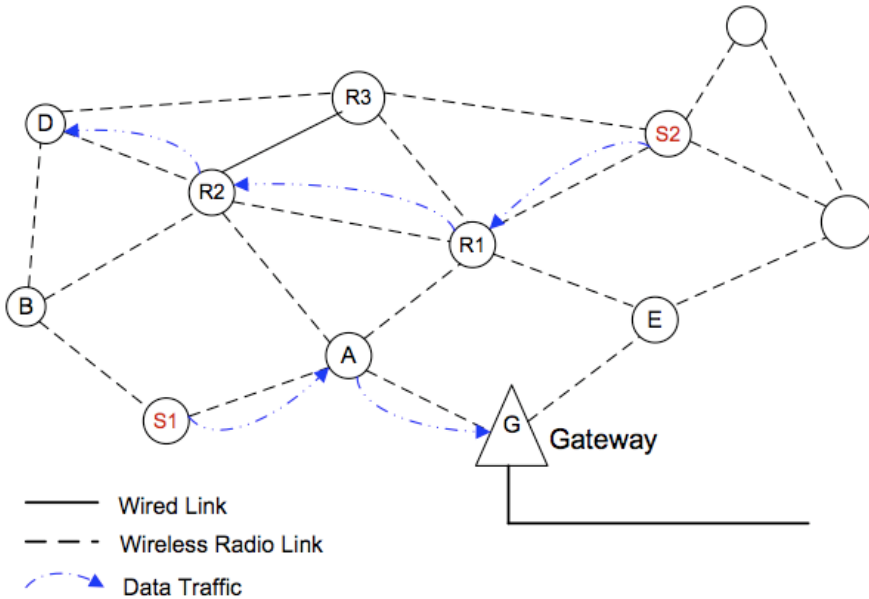


Fig. 1. An illustration of a Wireless Multi-Hop Network.

1.1 The Challenges of Wireless Multi-hop Routing in a Time Varying Environment

Routing is the fundamental issue for the multi-hop networks. A lot of routing protocols have been proposed for the wired networks and some of them have been widely used such as Routing Information Protocol (RIP) (Hedrick, 1998) and Open Shortest Path First (OSPF) (Moy, 1998). Characteristics of wireless links differ extremely from wired links. Thus, the existing routing protocols for wired networks can not work efficiently in the face of the vagaries of the radio channels and limited battery life and processing power of the devices. Moreover, the traditional routing metric of minimum-hop is not always the best solution for routing in wireless network. In the sequel, the wireless links characteristics and limitations of shortest path routing are described.

Wireless networks have intrinsic characteristics that affect intensely the performance of transport protocols. These peculiarities, which distinguish themselves from conventional wireline networks, can be summarized as follows:

- Wireless links have fundamentally low capacity. The upper band for the capacity of a wireless link follows the Shannon capacity bound.
- Signal propagation experiences large scale and small scale attenuations. Mobility of the nodes, path loss, shadowing and multi-path fading due to reflection, diffraction, scattering, absorption lead to slow and fast variations in channel quality even within the milliseconds scale (Proakis, 2004).
- The wireless medium is a broadcast medium. Therefore, in contrast to wired networks, the interference caused by other in-range traffic can unlimitedly disturb a transmission. This

causes the wireless link capacity to depend also on the sensitivity of receivers in sensing the environment as well as other links status in terms of their transmission range and power.

- Packet reception reliability over a link depends on several parameters such as modulation, source/channel coding of that link, the sensitivity of the link and the length of the packets.

Radio channels have some additional features such as asymmetrical nature and non-isotropic connectivity (Ganesan et. al, 2002; Cerpa et. al, 2003; Zhou et. al, 2004). Asymmetry of the channels means connectivity from node A to node B might differ significantly from B to A and non-isotropic connectivity means nodes geographically far away from source may get better connectivity than nodes that are geographically closer.

As a result of these characteristics, the radio cell is neither binary nor static. From the perspective of a node, the set of other nodes it can hear and the loss probability to or from these nodes vary abruptly over time with a large magnitude. This has been widely confirmed in real platforms (Couto et al, 2002; Ganesan et. al, 2002; Cerpa et. al, 2003; Couto, 2004). These random variations induce much more complexity for wireless networks to guarantee performance to transmit real-time or even critical data.

1.2 Limitations of Shortest Path Routing

Most of the existing routing algorithms use the shortest-path metric to find one or more multi-hop paths between the node pairs (Perkins & Royer, 1994; Johnson & Maltz, 1994; Park & Corson, 1997; Perkins & Belding-Royer, 2003; De Couto et. al, 2005). The advantage of this metric is its simplicity and a low overhead to the network. Once the topology is known, it is easy to find a path with a minimum number of hops between a source and a destination without additional measurement and overhead. Recent researches show that choosing the path with the smallest number of hops between nodes often leads to poor performances (De Couto et. al, 2002; De Couto et. al, 2005; Yarvis et. al, 2002).

One of the limitations of shortest path routing is that it does not capture the variable nature of wireless links. Instead, it assumes that the links between nodes either work well or do not work at all. Figure 2 shows an illustration of the different assumptions made by minimum-hop routing and link quality aware routing on the wireless links. This shows that the arbitrary choice made by minimum hop-count is not likely to select the best path among the same minimum length with widely varying qualities. Moreover, one of the current trends in wireless communication is to enable devices to operate using many different transmission rates to deal with changes in connectivity due to mobility and interference. In multi-rate wireless networks, minimum-hop works even worse. Selecting the minimum-hop paths leads to maximizing the distance travelled by each hop, but longer links are not robust enough to operate at the higher rates. Therefore, shortest path routing results in selecting the paths with the lowest rates, which degrades dramatically the overall throughput of the network.

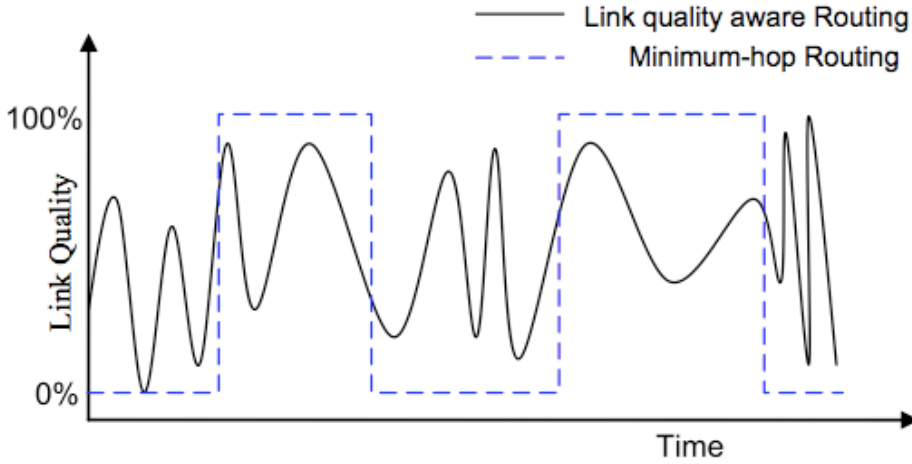


Fig. 2. Different assumptions for wireless link connectivity made by minimum-hop routing and link quality aware routing.

Furthermore, transmitting the flow over the low-rate links degrades the performance of other flows, which are transmitted over higher rate links. The main reason of this effect is that slow-speed links require larger amount of medium time to transmit a packet over the shared wireless medium and so block the other flows for a longer time. Heusse et al. denotes this problem as Performance Anomaly of 802.11b (Heusse et. al, 2003). They show analytically that a contending node with lower nominal bit rate degrades the throughput of faster contenders to even a lower bit-rate than the slowest sender. (Mahtre et. al, 2007; Razafindralambo et. al, 2008; Choi et. al, 2005) have evaluated and shown the same effect.

Another effect of multi-rate option for a minimum hop routing is that in multi-rate networks broadcast packets benefit from the longer range of low rate transmissions to reach farther nodes and so are always sent at the lowest transmission rate. Therefore, hearing the broadcast Hello messages from a node is not a good enough basis for determining that two nodes are well connected for transferring data packets at high rates. Lundgren et al. have referred to this effect as the gray-zone area (Lundgren et. al, 2002). A gray zone is the maximum area, which is covered by the broadcast messages at low rate, but not all the nodes in this area can forward the packets at high rates.

Choosing closer nodes with shorter-range links instead of minimum-hop routes can solve this problem. Consequently, minimum-hop metric has no flexibility in dealing with random quality fluctuations of the links. Link quality aware routing counters these limitations by using observation of miscellaneous parameters such as frame delivery or signal strength to select the good paths. In this approach, link quality metric is measured and observed in order to predict the near future quality of the links. This estimation is then used to determine the best route.

1.3 Cross Layer Interaction as a Solution

Typically, Open System Interconnection protocol stack (OSI) is divided into several layers which are designed independently. The interactions between adjacent layers are defined by some specific interfaces. Recently, in the quest of finding a link quality aware routing for wireless multi-hop networks, numerous link quality aware metrics have been proposed, which most of them are based on cross layer interactions between various layers of the protocol stack. Lately, there are many research efforts which show that transferring the status information between the layers can lead to a great improvement in network performance (Conti et. al, 2004; Shakkottai et. al, 2003; Goldsmith & Wicker, 1998). Recent activities of IEEE 802.11 task group in mesh networking have released IEEE 802.11s. It extends the IEEE 802.11 Medium Access Control (MAC) standard by defining an architecture and protocol that support both broadcast/multicast and unicast delivery using radio-aware metrics over self-configuring multi-hop topologies. This evolution pushes employing the cross layering technique in the real platforms in near future.

This chapter argues that the cross layering technique can be a promising solution in providing flexibility to the wireless network changes. Nevertheless evaluating the benefits of cross layer routing is often only based on the throughput, which is simplistic. Current studies generally do not consider the impact of other criteria such as response time or route flapping, which influence greatly applications performances in terms of throughput, but also mean delay, jitter and packet loss.

In the next section, a state-of-the-art of the main cross- layering metrics that have been proposed in the literature are presented. Then, the concept of reactivity for a link quality aware routing as a mean to analyse the true benefits of the cross-layer routing is introduced. Section 4 concludes this chapter.

2. Link Quality Aware Routing

Most of the primitive works in routing protocols for wireless multi-hop networks are inherited from existing routing protocols in wired networks. They devise mostly on coping with changing topology and mobile nodes (Perkins & Bhagwa, 1994; Perkins & Royer, 1999; Johnson & Maltz , 1994) and traditionally find the possible routes to any destination in the network with the minimum hop-count. As explained in Section 1.2, shortest path routing has sub-optimal performance, as they tend to include wireless links between distant nodes (De Couto et. al, 2002). A multitude of quality aware metrics have been proposed in the last decade which deal with the strict bandwidth and variable quality of wireless links and try to overcome the disadvantages of the minimum hop (MH) routing. Although most of them have been designed with the objective of increasing the transport capacity, each of them considers different QoS demands such as overall throughput, end-to-end delay, etc. Therefore, the proposed approaches for link quality aware routing can be categorized according to the aim of their design.

2.1 Wireless Network Capacity

The main purpose of efficient routing in mesh networks is improving the achieved capacity. The notion of capacity for wireless ad hoc network was defined as the maximum obtainable throughput from the network. It was first introduced by Gupta and Kumar in their seminal work (Gupta & Kumar, 2000). Network capacity for wireless multi-hop networks is

generally unknown, except for some centralized scheduling-based MAC protocols like Time Division Multiple Access (TDMA) where the problem finds a mathematical formulation.

The main finding in (Gupta, 2000) is that per-node capacity of a random wireless network with n static nodes scales as $\Theta(\frac{1}{n \log n})$. They assume a threshold-based link layer model in

which a packet transmission is successful if the received SNR at the receiver is greater than a fixed threshold. Instead of this ideal link layer model, Mhatre et al. consider a probabilistic lossy link model and show that the per-node throughput scales as only $\Theta(\frac{1}{n})$ instead of

$\Theta(\frac{1}{n \log n})$ (Mhatre & Rosenberg, 2006).

These asymptotic bounds are calculated under assumptions such as node homogeneity and random communication patterns. Therefore, some researches try to relax some of these assumptions on network configuration. Jain et al. focus on interference status among the transmitters as one of the main limiting factors on routing performance (Jain et. al, 2003). They propose to represent interference among wireless links using a conflict graph. A conflict graph shows, which wireless links interfere with each other, such that each edge in the connectivity graph is represented by a vertex and there exists an edge between two vertices if the links interfere with each other. Thus, the throughput optimization problem is posed as a linear programming problem in which upper and lower bounds of the maximum throughput are obtained by finding the maximal clique and independent set in the conflict graph.

Karnik et al. extend the conflict graph idea to a conflict set (Karnik et. al, 2007). Their rationale is that an interference model can not be binary since for a given link, generally there is a subset of links that at least one of them should be silent when the given link is transmitting. They propose a joint optimization of routing, scheduling and physical layer parameters to achieve the highest throughput. Both these proposals bring valuable achievement but as they investigate the highest capacity of a network, they have to assume TDMA instead of contention-based algorithm, which leads to probabilistic results. Hence, they implicitly assume that data transmissions are scheduled by a central entity. Therefore they may not be applied easily to more practical networks such as IEEE 802.11 with random access to the channel.

Computing the optimal throughput, despite of giving a good vision to the maximum achievable throughput in the network, may not be implementable in a real network. There are many complicated issues such as necessity of having a distributed routing/scheduling protocol, random quality for the wireless links, limited allowable overhead to the network, compatibility with MAC 802.11, etc., that motivate the researchers to find a practical solution to achieve a good performance.

2.2 Maximum Throughput Routing

Most of the work done in this area relies on the broadcasting of extra probe packets to estimate the channel quality (ex. Sivakumar et. al, 1999; De Couto et. al, 2005; Draves & Zill, 2004). However, since the quality of the wireless links depends significantly on physical settings (such as transmit rate and packet size), the probes may not reflect the actual quality of the links. The reason is that for preventing to throttle the entire channel capacity they

have to use small-sized probes at low transmission rate. Therefore the quality experienced by larger data packets at variable transmission rate is not the same as probe packets.

De Couto et al proposes a simple and effective routing metric called the expected transmission count (ETX) for 802.11-based radios employing link-layer retransmissions to recover frame losses (De Couto et. al, 2005). ETX of a wireless link is defined as the average number of transmissions necessary to transfer a packet successfully over a link. For estimating the expected number of transmission of the links, each node broadcasts periodically fixed-size probe packets. This enables every node to estimate the frame loss ratio p_f to each of its neighbors over a window time, and obtain an estimate p_r of the reverse direction from its neighbors. Then, assuming uniform distribution of error-rate over each link, the node can estimate the expected transmission count as $\frac{1}{(1 - p_f)(1 - p_r)}$. The ETX of

a path is obtained by summing up the ETX of its links. Therefore, each node picks the path that has the smallest ETX value from a set of choices. ETX has several drawbacks. First, its measurement scheme by using identical small-sized probe packets does not reflect the actual error-rate that the data packets experience over each link. The accuracy of the measurement scheme of a link quality metric has a great impact on its functionality (Karbaschi et al. 2008). Furthermore, ETX does not account the link layer abandon after a certain threshold of retransmissions. This may induce to select a path, which contains links with high loss rate. Koksall et al. introduce another version of ETX, called ENT (Effective Number of Transmissions), which deals with this problem (Koksall & Balakrishnan, 2006). ENT takes into account the probability that the number of transmissions exceeds a certain threshold and then calculates the effective transmission count based on an application requirement parameter, which limits this probability. Moreover, ETX by taking an inversion from the delivery rate to get the expected number of transmissions implicitly assumes uniform distribution for the bit error rate (BER) of the channel, which may not be correct.

The Expected Transmission Time (ETT), proposed by Draves et al. improves ETX by considering differences in link transmission rates and data packet sizes (Draves & Zill, 2004). The ETT of a link l is defined as the expected MAC layer latency to transfer successfully a packet over link l . The relation between ETT and ETX of a link l is expressed as:

$$ETT_l = ETX \frac{s_l}{r_l} \quad (1)$$

where r_l is the transmission rate of link l and s_l is the data packet size transmitted over that link. The weight of a path is simply the summation of the ETT's of the links of that path. The drawback of ETT is, as it is based on ETX, it may choose the paths, which contain the links with high loss rates. (Draves & Zill, 2004) proposes also another new metric based on ETT, which is called Weighted Cumulative Expected Transmission Time (WCETT). The purpose of this metric is finding the minimum weight path in a multi-radio network. The WCETT is motivated by observing that, enabling the nodes with multi-radio capability reduces the intra-flow interference. This interference is caused by the nodes of a path of a given flow competing with each other for channel bandwidth. For a path p , WCETT is defined as:

$$WCETT(p) = (1 - \beta) \cdot \sum_{l \in p} ETT_l + \beta \cdot \text{Max}_{1 \leq j \leq k} (X_j) \quad (2)$$

where X_j is the number of times channel j is used along path p and β is an adjustable parameter for the moving average subject to $0 \leq \beta \leq 1$. (Yang et. al, 2005) shows that WCETT is not non-isotonic and thus it is not a loop-free metric.

One of the characteristics of wireless links that can be observed is the received signal strength. It is very attractive if link quality can be reliably inferred by simply measuring the received signal strength from each received packet. Theoretically, the BER is expected to have a direct correlation with the received signal-to-noise ratio (SNR) of the packet, and the packet error rate is a function of the BER and coding. Therefore, the SNR level of the received packets has been widely used as a predictor for the loss rate of the wireless links (ex. Goff et. al, 2001; Dube et. al, 1997). (Aguayo et. al, 2004; Woo, 2004) through collecting experimental data have shown that although SNR has an impact on the delivery probability, lower values of SNR has a weak correlation with the loss rate of the links. Thus, it can not predict the quality of the links easily. In addition, (Woo, 2004) illustrates that where traffic load interference happens, collisions can affect link quality even though the received signal is very strong. The main reason is that prediction of the link quality by observing the SNR samples of the packets, counts only on the packets which are received successfully. This leads to ignoring the congestion status of the links.

A number of proposed wireless routing algorithms collect per-link signal strength information and apply a threshold to avoid links with high loss ratios (Goff et. al, 2001; Yarvis et. al, 2002). In the case that there is only one lossy route to the destination, this approach may eliminate links that are necessary to maintain the network connectivity.

2.3 Minimum Delay Routing

Some existing link quality metrics focus on finding the best paths based on the end-to-end delay associated to each path. The rationale for minimizing the paths latency is that in a fixed transmission power scenario, packets latency for reaching successfully to the other end of a link can provide an estimation of the quality of that link. The average round trip time (RTT) of the packets over each link is one of the delay based parameters representing the link quality. (Adya et. al, 2004) for instance proposes this metric. To calculate RTT, a node sends periodically a probe packet carrying its time stamp to each of its neighbors. Each neighbor immediately responds to the received probe with a probe acknowledgment which echoes its time stamp. This enables the sending node to calculate the RTT to each of its neighbors. Each node keeps an average of the measured RTT to each of its neighbors. If a probe or a response probe is lost, the average is increased to reflect this loss. A path with the least sum of RTTs is selected between any node pair.

The RTT reflects several factors, which have impact on the quality of a link. First, if a link between the nodes is lossy its average RTT is increased to give a higher weight to that link. Second, either if the sender or the neighbor is busy, the probe or its response is delayed due to queuing delay which leads to higher RTT. Third, if other nodes in the transmission range of the sender are busy, the probes experience higher delay to access the channel again resulting in higher RTT. Concisely, RTT measures the contention status and error rate of a link.

However, the small probe packets in comparison to larger data packets are rarely dropped over a lossy channel. This hides the actual bandwidth of the links. Moreover, (Draves et. al, 2004) shows that RTT can be very load-sensitive which leads to unnecessary route

instability. Load-dependency of a metric is a well-known problem in wired networks (Khanna & Zinky, 1989). To suppress the queuing delay in the RTT, Keshav, 91, proposed the packet-pair technique to measure delay of a link. In this approach to calculate the per-hop delay, a node sends periodically two probe packets back to back to each of its neighbors such that the first probe is small and the next one is large. The neighbor upon receiving the probes calculates the delay between them and then reports this delay back to the sender. The sender keeps an average from the delay samples of each neighbor and paths with lower cumulated delay are selected. This technique, by using larger packet for the second probe, reflects more accurately the actual bandwidth of the links, although it has higher overhead than RTT. Draves et. al, 2004 discusses again that packet-pair measurement is not completely free of self interference between the neighbors, although less severe than RTT.

Awerbuch et. al 2004 proposes the Medium Time Metric (MTM) which assigns a weight to each link proportional to the amount of medium time consumed by transmitting a packet on the link. It takes the inverse of the nominal rate of the links to estimate the medium time.

The variable rate of the links is determined by an auto-rate algorithm employed by the networks, such as ARF or RBAR. Existing shortest path protocols will then discover the path that minimizes the total transmission time. This metric only handles the transmission rate of the links and does not account the medium access contention and retransmission of packets at the MAC layer. Zhao et al. 2005 introduces a cross layer metric called PARMA, which aims to minimize end-to-end delay which includes the transmission delay, access delay and the queuing delay.

They consider a low saturated system with ignorable queuing delay. A passive estimation is used for the channel access delay and the transmission delay on each link is calculated as the ratio of packet length to the link speed. This metric has a good insight into estimating the total delay of each link but has simplified very much the problem of the delay calculation. For instance, it assumes that the links are error free and no packet retransmissions occur over them.

2.4 Load Balancing

In order to make routing efficiently and increase network utilization, some researchers have proposed congestion aware routing with the aim of load balancing in the network. One of the methods for spreading the traffic is using multiple non-overlapping channels. Kyasanur & Vaidya, 2006 propose a Multi-Channel Routing protocol (MCR) with the assumption that the number of interfaces per node is smaller than the number of channels. The purpose of their protocol is choosing paths with channel diversity in order to reduce the self interference between the node pairs. Moreover, they take into account the cost of interface switching latency. MCR is based on on-demand routing in a multi-channel network. While the on-demand route discovery provides strong resistance to mobility-caused link breaks, the long expected lifetimes of links in mesh networks make on-demand route discovery redundant and expensive in terms of control message overhead. Therefore, this protocol is not totally appropriate for mesh networks. Yang et al. focus more on mesh networks and propose another path weight function called Metric of Interference and Channel-switching (MIC). A routing scheme, called Load and Interference Balanced Routing Algorithm (LIBRA) is also presented to provide load balancing (Yang, 2005). MIC includes both interference and channel switching cost.

2.5 Routing with Controlling Transmission Power

Numerous works in efficient routing in multi-hop networks has focused on power control routing. The problem of power control has been investigated in two main research directions: energy-aware and interference-aware routing. In energy-aware routing approaches the objective is to find power values and routing strategy, which minimizes the consumption of power in order to maximize the battery lifetime of mobile devices. Therefore these works are suitable for sensor and ad hoc networks as in wireless mesh networks power is not a restricted constraint. Power control in interference-aware routing aims to find the optimal transmission power which gives the higher throughput or the lowest end-to-end delay. Therefore, transmission power of the nodes is controlled in order to reduce interference while preserving the connectivity. There are a lot of researches in this area. For instance Iannone et al. propose Mesh Routing Strategy (MRS) in which transmission rate, PER and interference of each link are taken into account (Iannone & Fdida, 2006). The interference is calculated based on the transmission power and number of reachable neighbors with that power level. The disadvantage of their approach is that they do not consider that links with different transmission rate have different sensibility for being disturbed by the neighbors' transmission. This effect has been taken into account in (Karnik et. al., 2008) where the authors propose a network configuration to have an optimal throughput.

The foreseen alternatives to minimum hop metric consist in establishing high quality paths, by tracking various link quality metrics in order to significantly improve the routing performances. Thus, the challenge lies in selecting *good* paths, based on a *relevant* link quality metric. However, the stability issue of link quality aware routing which can be extremely important specially in providing quality of service for jitter sensitive applications has not been addressed by the existing research efforts. In the next section, a quantitative tool to investigate the routing reactivity and its impact on applications performances is introduced.

3. Reactivity of Link Quality Aware Routing

A more reactive routing responds faster to link quality changes. This leads to detect the lossy channel faster and so, to converge to the higher quality path in a shorter time. Meanwhile, fast reacting to channel variations may produce higher path flapping and consequently higher jitter level. Therefore, there is a trade-off between providing ensured stability in selecting the paths and obtaining a high possible throughput from the network. The frequency of link quality changes may be very different for distinct wireless links (due to some factors such as fast or slow fading, nodes mobility, etc.) (Koskal & Balakrishnan, 2006; Aguayo et. al, 2004). In order to track as much as possible all the link changes and always choose a high quality path, the routing should respond accurately and as fast as possible to these changes. Response time refers to the time required by the routing agent to take into account the new link quality status.

The reactivity of the routing depends on the updating frequency of the routing tables and the sensitivity degree of the routing metric to channel variations. The updating frequency of the routing tables defines the rate at which the shortest paths are recalculated based on the current value of the link quality metrics. Although the updating period of the routing is generally longer than the time-scale of link quality variations, a shorter update period is able

to respond faster to link breakage or quality degradation and in turn will lead to a higher throughput. However, frequent changes of the selected route induce packet reordering and jitter issues. Moreover, reducing the updating period of the metric obviously produces a higher amount of routing overhead. This may overload the network and could severely degrade the network performance. The sensitivity degree of the routing metric to link quality variations is the other parameter which obviously has a great impact on the routing response time. The sensitivity degree depends on the way the set of measured parameters (frame loss, delay, SNR, ...) are mapped onto the metric. Sensitivity degree S of a link quality metric is defined as the norm of the gradient of the defined metric function with respect to the set of parameters that measures the link quality. Let q be the set of measured parameters and $m(q)$ the calculated metric based on q . The sensitivity of the metric is:

$$S_m(q) = \left\| \vec{\Delta m(q)} \right\| \quad (3)$$

With a highly sensitive metric, the variations of link conditions are intensified. The path metric, which aggregates the link metrics along a path, fluctuates faster and the probability of changing the selected route increases. The possibly resulting route flapping may cause higher jitter which for some applications is harmful as reordered or delayed packets may be considered as lost ones. This section focuses on investigating the impact of a more sensitive metric on route flapping, control overhead and real-time application performances.

3.1 Impact of the Sensitivity of a Link Quality Metric

In (Karbaschi et. al, 2008) it is shown that measurement scheme and obviously the relevance of the measured parameters have a great impact on the measurement accuracy and thus on the final result. Numerous link quality routing have been proposed in the last decade. However, in order to compare the impact of the sensitivity of two link quality metrics on the routing performance, their measurement scheme and the parameters they measure for estimating the quality of the links should be the same. No such two link quality metrics with identical observed parameters and measurement scheme can be found in the literature. Therefore, to conduct the study, two comparable and realistic link quality metrics are introduced in the following.

ARQ mechanism in 802.11b with retransmission of frames over lossy channels wastes bandwidth and causes higher end-to-end delay and interference to the other existing traffics. Therefore, the number of frame retransmissions at the MAC layer has been widely used as an estimator of the link quality (De Couto et. al, 2005; Koskal & Balakrishnan, 2006; Karbaschi et. al, 2008). Therefore, this section presents two comparable link quality metrics based on this retransmissions number.

The first link quality metric, called m_1 , is based on the FTE metric introduced in (Karbaschi et. al. 2005). Assuming that RTS/CTS is enabled for solving the hidden terminal problem, the quality and interference status of the adjacent links of a sender can be estimated by measuring the average number of required retransmissions of data and RTS frames at the MAC layer to transfer a unicast packet across a link. Therefore, each node measures the m_1 by keeping the retransmissions count of RTS and data frames over the neighbor links as follows.

Let $k_{xy}(i)$ - respectively $l_{xy}(i)$ - be the number of transmissions (including retransmissions) of the i^{th} data packet - respectively i^{th} RTS packet - over the x to y link. Thus the set of measured parameters (q) over this link will be defined as $q_{xy}(i) = \{k_{xy}(i), l_{xy}(i)\}$. The success rate in delivering two frames of data and RTS from x to y denoted by $m_1(q_{xy}^i)$ is computed as:

$$m_1(q_{xy}^i) = \frac{2}{k_{xy}(i) + l_{xy}(i)} \quad (4)$$

Referring to Equation 4, increasing the number of retransmissions of RTS or data frames reduces the value of m_1 and for a perfectly efficient link $m_1(q_{xy}^i)$ is equal to unity. If the number of retransmissions reaches a predefined threshold, the sender gives up sending the frame. In this case, $m_1(q_{xy}^i)$ is set to zero which degrades the overall average link quality very much. m_1 can be interpreted as an estimation of the success rate of transmissions over a link. Another metric, called m_2 is defined based on the same measured parameters. Assuming that the failure in transmission of data and RTS frames are independent from each other, the success rate of transmission over the link x to y can be calculated by multiplying the success probability of sending RTS and data frames as follows:

$$m_2(q_{xy}^i) = \frac{1}{k_{xy}(i)} \times \frac{1}{l_{xy}(i)} \quad (5)$$

With the same argument, if the sender gives up sending RTS or data frames, $m_2(q_{xy}^i)$ is set to zero.

Figure 3 illustrates the calculated success rate returned by m_1 and m_2 as a function of the number of data and RTS retransmissions for each sent packet over a given link (Equation 4 and 5). As shown in this figure, an interesting property of both m_1 and m_2 is that their variations over the range of RTS and data retransmission numbers is not uniform. Indeed, both metrics are much more sensitive to a given variation of its arguments k_{xy} and l_{xy} when these parameters ranges between 1 and 4 than for values ranging between 6 to 10. In other words, the quality variations of a poor link are far less reflected in the metric than the variations of a high quality link. This is desirable since if a link does not work well and there is no alternate much higher quality link, it is not worth changing the selected path. In counterparts, quality variations of a good links have a much greater impact on the throughput of that link. As a result, good quality links should be more prone to changes.

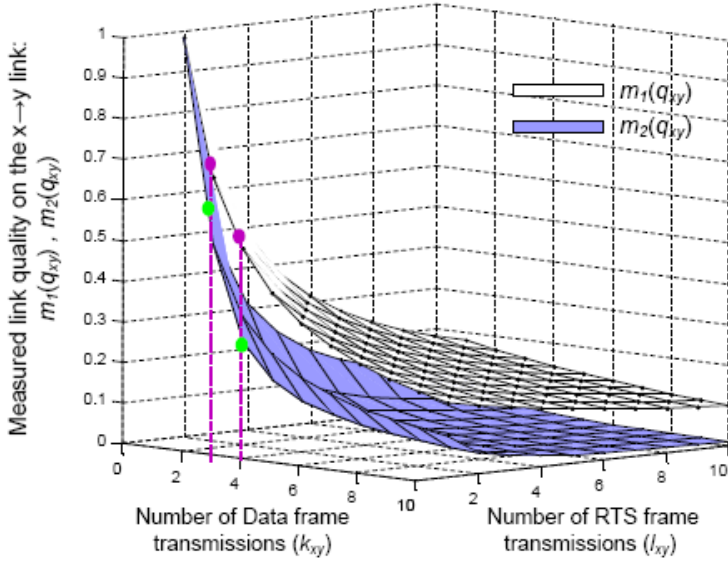


Fig. 3. Measured link quality using m_1 and m_2 function of the number of Data and RTS frames retransmissions.

From this point of view, the two metrics differ. Indeed, m_2 differentiates better than m_1 a small degradedness from a former high quality measured value. As clearly shown in Figure 3, both functions return 1 when no retransmission occurs which confirms the value of 100% success for the transmission while by lessening in link quality, m_2 drops more sharply than m_1 . For example, an increase in the number of data transmissions from 3 to 4 causes 32% decrease in m_2 and about 15 % in m_1 . Therefore, m_2 is called as Faster FTE (FFTE).

The variations of the metric with respect to quality changes can be evaluated using the sensitivity degree (Equation 3). Figure 4 compares the sensitivity degree of m_1 and m_2 using the difference $S_{m_1}(q_{xy}) - S_{m_2}(q_{xy})$. We see that for all the variations range of $k_{xy}(i)$ and $l_{xy}(i)$, S_{m_2} is larger or equal to S_{m_1} .

Consequently, m_2 has an even greater sensitivity in detecting changes in the estimated link quality than m_1 . m_2 obliges the routing agent to be more reactive and changes the selected route more often than m_1 .

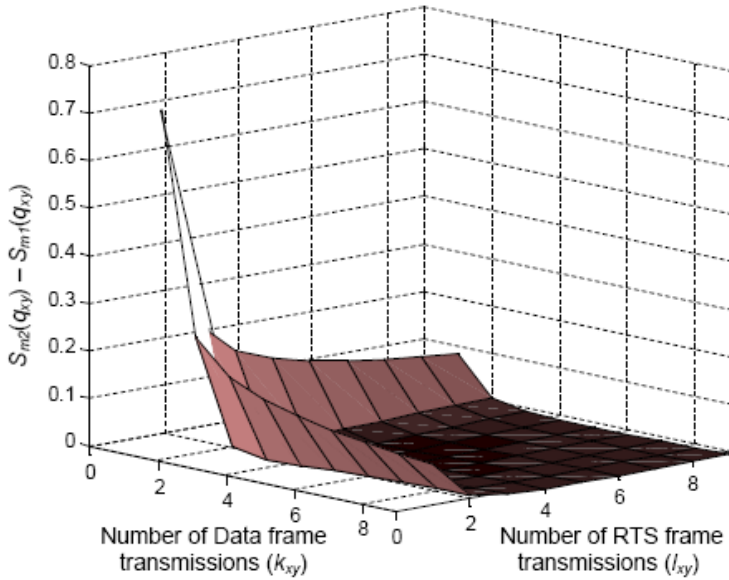


Fig. 4. Comparison of the sensitivity of the metrics m_1 and m_2 .

3.2 Simulation Study

This section presents simulation results to illustrate the performance of the link quality metrics compared to the Minimum Hop (MH) metric as the reference. Both m_1 and m_2 have been employed in DSDV (Perkins, 1994). The efficiency of the routes is estimated by multiplying the EWMA of m_1 values (or m_2) along the path towards the destination. This estimation of the link quality is piggy backed into the Hello messages that are sent in a periodic manner.

The simulations are performed under ns2.28 with the enriching the simulator in order to contain wireless channel fading effects, time variable link quality for wireless links, signal to interference and noise ratio, etc (cf. Karbaschi, 2008).

In order to show the impact of the link quality aware routing on the quality of service for a jitter-sensible flow, VoIP traffic of is modelled and multiple random connections are set in a 30-nodes random topology. VoIP is basically UDP packets encapsulating RTP packets which contain the voice data. For accurately modelling the bursty VoIP traffic, Pareto On/Off traffic is used (Dang et. al, 2004), with different transmission rates corresponding to the widely used ITU voice coders.

Firstly, the impact of the sensitivity of the metric on the performance of the DSDV is evaluated and then the resultant instability and the generated jitter are investigated. T (resp. T_0) are used as the routing update period used in the case that m_1 and m_2 (resp. MH) are implemented

In order to unify the impact of the update period on the reactivity of the routing, T and T_0 are set to 15 s. The three metrics in terms of received throughput, defined as the average number of data bits received per second, are compared in Figure 5. The result of one connection confirms that both the link quality metrics are able to transfer more bits in

comparison to the MH metric in a time duration of 2000 s. Figure 5 also shows that m_2 outperforms the two other metrics. This confirms that m_2 , the more sensitive metric, is able to find a higher throughput path faster than m_1 and so reduces the packet drops.

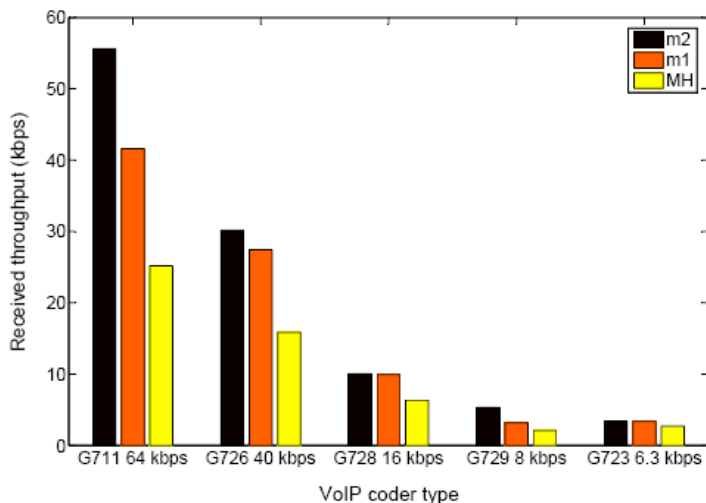


Fig. 5. Average received throughput with same routing update period ($T = T_0 = 15s$)

Comparing in Figure 6 the number of times that one dedicated flow flaps per second reveals that a link quality metric leads the routing to change the selected path more frequently. This effect is even greater when the sensitivity of the metric increases (m_2 compared to m_1). To show the impact of the metrics on the routing overhead, the average bit rate of control messages for the three metrics are compared in Figure 7. This reveals that the overhead generated by m_1 and m_2 are nearly the same and both more than MH's overhead. The reason is that the higher sensitivity of m_1 and m_2 generates more paths changes than MH. This obliges the nodes to piggy back more neighbors entries into their broadcast message and makes it larger, thus raising the routing overhead. Another cause of overhead increase is that a link quality metric needs a larger field in the control message than a hop metric (32 bits compared 16 bits). This enlarges the overall size of the control messages.

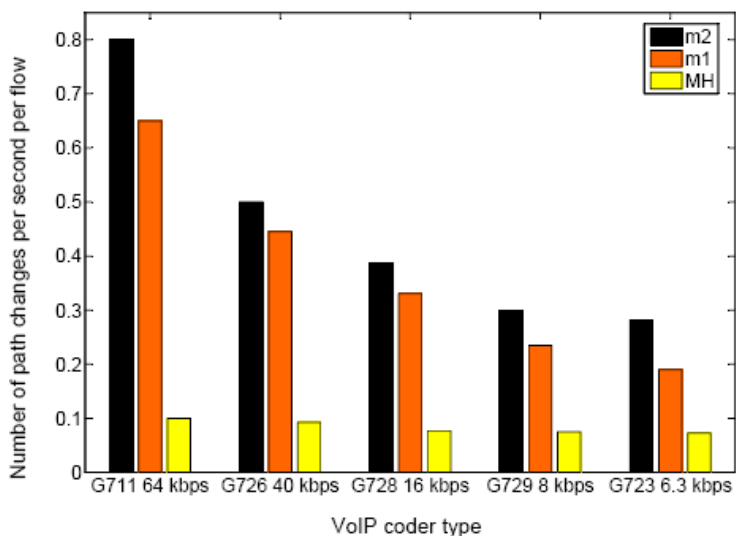


Fig. 6. Number of path changes per second per flow.

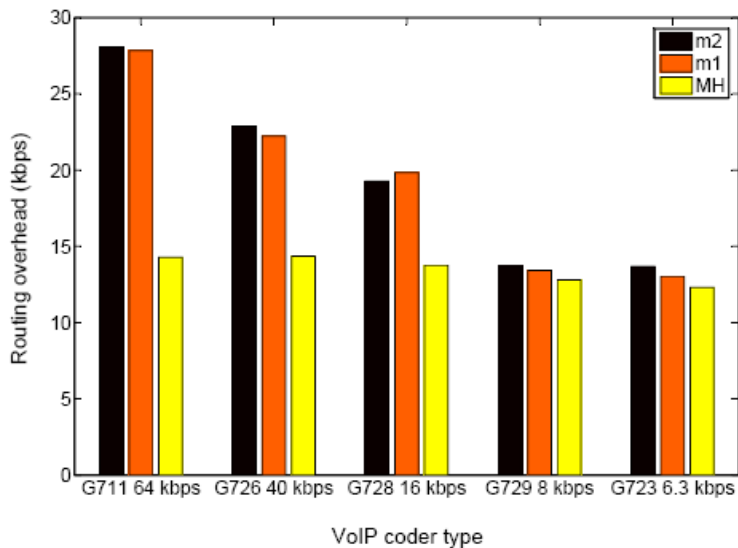


Fig. 7. Routing overhead for different VoIP coder types.

To see the efficiency of functionality of the link quality metrics, the evaluation is repeated by adjusting the amount of overhead to an identical value for the three metrics by tuning the value of T to 30 s. As explained in Section 3 this may reduce the throughput of m_1 and m_2 due to lower update rate of the metric. However, the average throughput comparison shows that m_2 still brings a much higher throughput (Figure 8).

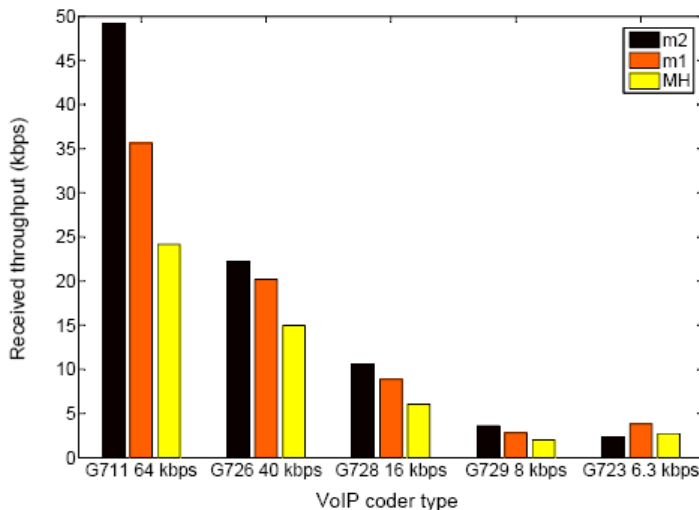


Fig. 8. Average received throughput with same overhead amount ($T = 30$ s, $T_0 = 15$ s) for different VoIP coder types.

Flapping the selected route may cause the consecutive packets to be routed through different routes. The subsequent instability of the selected path may cause a higher jitter level. Figure 9 illustrates the measured jitter per packet for the three metrics during a sample interval of a VoIP connection. Table 1 gives the mean and standard deviation of the jitter measured using the three metrics, which gives an idea of the spreading of the jitter distribution.

	MH	m_1	m_2
mean (ms)	9.5	9.6	9.43
std (ms)	20	25	30.2

Table 1. Jitter statistics

The jitter experienced using m_2 is much greater than the jitter observed with m_1 and MH. High jitter levels can have a great impact on the perceived quality in a voice conversation and as a result, many service providers now account for maximum jitter levels.

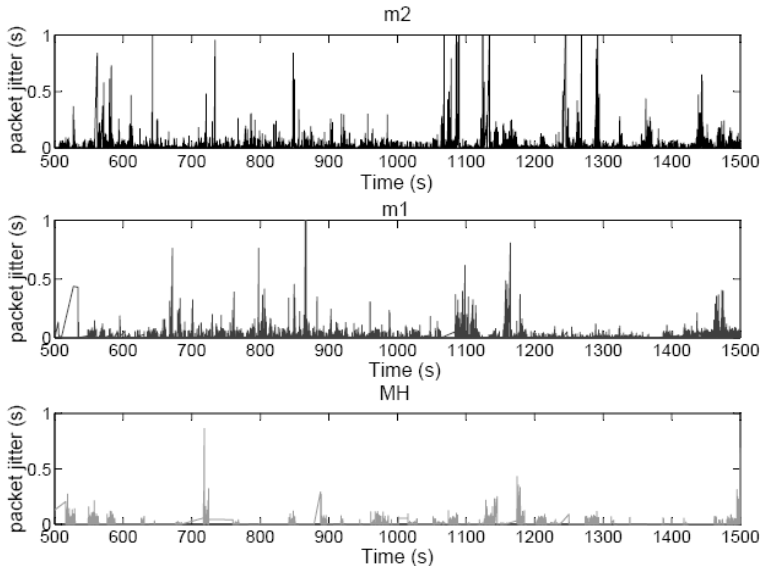


Fig. 9. Comparison of the jitter per received packet in a VoIP connection.

Most of the VoIP end-devices use a de-jitter buffer to compensate the jitter transforming the variable delay into a fixed delay (Khasnabish, 2003). Thus, high levels of jitter increase the network latency and cause a large number of packets to be discarded by the receiver. This may result in severe degradation in call quality. Therefore, real-time applications may not benefit from the higher throughput obtained with a more sensitive metric.

4. Conclusion

Wireless multi-hop networks are a promising technology to provide flexibility and rapid deployment for connecting the users at low cost. This chapter has examined the issues of link quality aware routing for wireless multi hop networks. Since the quality of wireless communications depends on many different parameters, it can vary dramatically over time and with even slight environmental changes. The peculiarity of wireless links and strong fluctuations in their quality lead to challenges in designing wireless multi hop routings. Therefore the necessity of having more efficient routing rather than the ones proposed for wired networks has arisen. Traditional hop count shortest-path routing protocols fail to provide reliable and high performance because of their blindness to under layer status. This draws lots research efforts to improve the routing performance through choosing good paths via transferring link status information from under layers.

This chapter addressed the stability issues of link quality aware routing which can be extremely important specially in providing quality of service for jitter sensitive applications. It was argued that having a reactive routing to cope with random changes of wireless links is essential. A quantitative tool for estimating the sensitivity of a link quality metric was introduced which indicated how strongly the metric reflects the quality changes. It was shown that the sensitivity has a great impact on the routing adaptivity. To illustrate this,

two comparable link quality metrics (FTE and FFTE) with different sensitivity were introduced and the routing performance observed with these metrics was compared by simulation. It was shown that having a sensitive metric can improve the routing functionality in terms of transferring a higher number of data packets through the network. However, one resulting side-effect is more oscillation in path selection. This leads to higher jitter level which a delicate application such as VoIP may not tolerate.

5. References

- Adya, A. ; Bahl, P. ; Padhye, J. ; Wolman, A. & Zhou, L (2004). A multi-radio unification protocol for ieee 802.11 wireless networks, *First International Conference on Broadband Networks (BROADNETS'04)*, ISBN 0-7695-2221-1, pp. 344–354, October 2004, San Jose, California, USA, IEEE Computer Society, 2004.
- Aguayo, D. ; Bicket, J. ; Biswas, S. ; Judd, G. & Morris, R. (2004). Link-level measurements from an 802.11b mesh network, *Proceedings of ACM Sigcomm*, Vol. 34, No. 4, pp. 121–132, Aug-Sept 2004, ACM, 1-58113-862-8/04/0008.
- Awerbuch, B. ; Holmer, D. & Rubens, H. (2004). High throughput route selection in multi-rate ad hoc wireless networks, *Kluwer Mobile Networks and Applications (MONET) Journal Special Issue on Internet Wireless Access: 802.11 and Beyond*, Vol. 2928, pp. 253–270, ISBN 3-540-20790-2, January 2004.
- Cerpa, A. ; Busek, N. & Estrin, D. (2003). Scale: A tool for simple connectivity assessment in lossy environments, *In CENS Technical Report 0021*, September 2003, experimental evaluation of wireless links.
- Conti, M. ; Maselli, G. ; Turi, G. & Giordano, S. (2004). Cross-Layering in Mobile Ad Hoc Network Design, *Computer*, Vol. 37, No. 2, pp. 48–51, Feb. 2004, doi:10.1109/MC.2004.1266295
- Choi, S. ; Park, K. & Kim, C. (2005). On the performance characteristics of WLANs: Revisited. *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pp. 97–108, ISBN 1-59593-022-1, Banff, Alberta, Canada, ACM, New York, USA.
- Dang, T. D. ; Sonkoly, B. & Molnar, S. (2004). Fractal analysis and modelling of VoIP traffic, *In International Telecommunications Network Strategy and Planning Symposium*, ISBN: 3-8007-2840-0, pp. 556–562, June 2004, DOI 10.1109/NETWKS.2004.1341826
- De Couto, D. S. J. ; Aguayo, D. ; Chambers, B. A. & Morris, R. (2002). Performance of multihop wireless networks: Shortest path is not enough. *Proceedings of the First Workshop on Hot Topics in Networks (HotNets-I)*. Princeton, New Jersey: ACM SIGCOMM, October 2002.
- De Couto, S. J. D. (2004). High-throughput routing for multi-hop wireless networks. *Ph.D. dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology*, June 2004.
- De Couto, D. S. J., Aguayo, B ; Chambers, B. A. & Morris R. (2005). A high-throughput path metric for multi-hop wireless routing, *Wireless Networks*, Volume 11 , Issue 4, July 2005, pp 419 – 434, ISSN:1022-0038.
- Draves, R. ; Padhye, J. & Zill B. (2004). Routing in multi-radio, multi-hop wireless mesh networks, *Proceedings of the 10th annual international conference on Mobile computing and Networking*, pp. 114–128, ISBN:1-58113-868-7, International Conference on

- Mobile Computing and Networking, Philadelphia, PA, USA, ACM, New York, NY, USA, 2004.
- Dube, R. ; Rais, C. D. ; Wang, K.-Y. & Tripathi, S. K. (1997). Signal stability-based adaptive routing (ssa) for ad hoc mobile networks, *IEEE Personal Communications*, Vol 4, pp 36-45, February 1997, Digital Object Identifier 10.1109/98.575990
- Ganesan, D. ; Krishnamachari, B. ; Woo, A. ; Culler, D. ; Estrin, D. & Wicker, S. (2002). Complex behavior at scale: An experimental study, *UCLA Computer Science*, Tech. Rep. TR 02-0013, 2002.
- Goldsmith A. & Wicker S. (1998). Design challenges for energy-constrained ad hoc wireless networks, *IEEE Wireless Communications*, vol. 9, no. 4, pp. 8-27, 2002.
- Goff, T. ; Abu-Ghazaleh, N. B. ; Phatak, D. S. & Kahvecioglu, R. (2001). Preemptive routing in ad hoc networks, *Proceeding of ACM/IEEE MobiCom*, pp. 43-52, ISBN 1-58113-422-3, July 2001, ACM SIGMOBILE, Rome, Italy, 2001.
- Gupta P. & Kumar P. R. (2000). The capacity of wireless networks, *IEEE Transactions on Information Theory*, March 2000, Vol. 46, No. 2, pp. 388-404, Digital Object Identifier 10.1109/18.825799.
- Hedrick, C. (1998). Routing information protocol. *Network working group*, RFC 1058, The Internet Society, June 1988.
- Heusse, M. ; Rousseau, F. ; Berger-Sabbatel, G. & Duda, A. (2003) . *Performance anomaly of 802.11b*, Proceedings of INFOCOM, Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. Vol. 2, pp. 836-843, 2003, ISBN: 0-7803-7752-4, ISSN: 0743-166X.
- Iannone, L. & Fdida, S. (2006). Evaluating a cross-layer approach for routing in wireless mesh networks, *Telecommunication Systems Journal (Springer) Special issue: Next Generation Networks - Architectures, Protocols, Performance*, vol. 31, no. 2-3, pp. 173-193, March 2006.
- Jain, K. ; Padhye, J. ; Padmanabhan, V. & Qiu L. (2006). *Impact of interference on multi-hop wireless network performance*, MobiCom '03: Proceedings of the 9th annual international conference on Mobile computing and networking, pp. 66-80, ACM Press, New York, NY, USA, 2003.
- Johnson D. B. & Maltz, D. A. (1994). *Dynamic source routing in ad-hoc wireless networks*. Proceedings of the Workshop on Mobile Computing Systems and Application, pp. 158-163, December 1994.
- Karbaschi, G. (2008). Link Quality Aware Routing in IEEE 802.11 Multi Hop Networks, *PhD Thesis dissertation, University of Pierre et Marie Curie*, April 2008.
- Karbaschi, G. & Fladenmuller, A. (2005). A link-quality and congestion-aware cross layer metric for multi-hop wireless routing, *Proceedings of IEEE Mobile Ad hoc and Sensor Systems (MASS)*, Nov. 2005.
- Karbaschi, G. ; Fladenmuller, A. & Wolfinger, B. E. (2008) *Link-quality measurement enhancement for routing in wireless mesh networks*, *Proceedings of International Symposium on a World of Wireless, Mobile and Multimedia Networks*, 2008 (WoWMoM 2008), ISBN: 978-1-4244-2099-5, pp. 1-9, Digital Object Identifier: 10.1109/WOWMOM.2008.4594842
- Keshav, S. (1991). *A control-theoretic approach to flow control*, ACM SIGCOMM Computer Communication Review, Vol. 21, no. 4, pp. 3-15, 1991, ISSN 0146-4833

- Khanna, A. & Zinky, J. (1989). *The revised arpanet routing metric*, ACM SIGCOMM Computer Communication Review, Vol. 19, No. 4, pp. 45–56, ACM New York, NY, USA 1989, ISSN 0146-4833.
- Karnik, A. ; Iyer A. & Rosenberg C. (2008). *Throughput-optimal configuration of wireless networks*, IEEE/ACM Transaction in Networking, vol 16, Issue 5, Oct. 2008, pp 1161-1174, Digital Object Identifier 10.1109/TNET.2007.909717.
- Khasnabish, B. (2003), *Implementing Voice over IP*, 1st edition, ISBN-10: 0471216666, ISBN-13: 978-0471216667, Wiley-Interscience publisher, 2003.
- Koksal C. E. & Balakrishnan H. (2006). *Quality-aware routing metrics for time-varying wireless mesh networks*, IEEE Journal on Selected Areas of Communication Special Issue on Multi-Hop Wireless Mesh Networks, vol. 24, no. 11, pp. 1984–1994, November 2006.
- Kyasanur, P. & Vaidya, N. H. (2006). Routing and link-layer protocols for multi channel multi-interface ad hoc wireless networks, *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 10, no. 1, pp. 31 –43, ISSN:1559-1662, Jan. 2006.
- Lundgren, H. ; Nordstrom, E. & Tschudin, C. (2002). Coping with communication grey zones in IEEE 802.11b based ad hoc networks. *Proceedings of IEEE Intl Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pp. 49–55, Atlanta, Ga, USA, September 2002.
- Mhatre, V. P. ; Lundgren, H. & Diot, C. (2007). Mac-aware routing in wireless mesh networks. *Proceedings of the Fourth Annual Conference on Wireless on Demand Network Systems and Services (WONS'07)*, pp. 46–49, Jan. 2007, ISBN: 1-4244-0860-1, Digital Object Identifier 10.1109/WONS.2007.340461, 2007-04-02.
- Mhatre V. & Rosenberg C. (2006). The impact of link layer model on the capacity of a random ad hoc network, *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, pp. 1688–1692, July 2006, Digital Object Identifier 10.1109/ISIT.2006.261642
- Moy, J. (1998). OSPF version 2, *network working group*, RFC 2328, The Internet Society, April 1998.
- Park V. D. & Corson M. S. (1997). A highly adaptive distributed routing algorithm for mobile wireless networks, *Proceedings of IEEE Infocom*, vol.3, pp.1405–1413, April 1997.
- Perkins, C. E. & Bhagwa, P. (1994). Highly dynamic destination-sequenced distance-vector routing (dsdv) for mobile computers," *ACM SIGCOMM Computer Communication Review*, pp. 234–244, 1994, ACM, New York, NY, USA, ISSN 0146-4833
- Perkins C. E. & Belding-Royer E. (1999). Ad-hoc on-demand distance vector routing, *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications*, pp. 90–100, New Orleans, LA, USA, Feb 1999.
- Perkins C. E. & Belding-Royer E. (2003). Ad hoc On-demand Distance Vector (aodv) routing, *Internet RFC 3561*, The Internet Society, July 2003.
- Perkins C. E. & Bhagwa, P. (1994). Highly dynamic destination-sequenced distance-vector routing (dsdv) for mobile computers. *Association for Computing Machinery's Special Interest Group on Data Communication'94*, Volume 24, Issue 4 October 1994, ISBN: 0-89791-686-7, pp. 234–244, ACM, New York, NY, USA
- Proakis, J. (2001). *Digital Communications*, McGraw-Hill, Fourth edition, ISBN 0-07-232111-3, New York.

- Razafindralambo, T. ; Guérin-Lassous, I. ; Iannone, L. & Fdida, S. (2008). Dynamic and distributed packet aggregation to solve the performance anomaly in 802.11 wireless networks. *Elsevier Computer Networks journal, Special Issue: Wireless Network Performance*, vol. 52, no. 1, pp. 77-95, 2008.
- Shakkottai, S. ; Rappaport, T. S. & Karlsson, P. C. (2003). Cross-layer design for wireless networks, *IEEE Communications Magazine*, Oct 2003, Vol. 41, o. 10, pp. 74-80, ISSN: 0163-6804.
- Sivakumar R. ; Sinha P. & Bharghavan V. (1999). Cedar: a core-extraction distributed ad hoc routing algorithm, *Communication IEEE Journal*, vol. 17, no. 8, Aug. 1999.
- Woo A. (2004). A holistic approach to multihop routing in sensor networks, *Ph.D. dissertation, University of California, Berkeley*, 2004
- Yang, Y. ; Wang, J. & Kravets, R. (2005). Interference-aware load balancing for multi-hop wireless networks, *Tech Report UIUCDCS-R-2005-2526, Department of Computer Science*, University of Illinois at Urbana-Champaign, 2005.
- Yarvis, M. Conner, W. S. Krishnamurthy, L. Chhabra, J. Elliott, B. & MainWaring A. (2002). Real-world experiences with an interactive ad hoc sensor network, *Proceedings of 31st International Conference on Parallel Processing Workshops (ICPP 2002 Workshops)*, pp. 143-151, ISBN 0-7695-1680-7, 20-23 August 2002, Vancouver, BC, Canada, IEEE Computer Society, August 2002.
- Zhao, S. ; Wu, Z. ; Acharya, A. & Raychaudhuri, D. (2005). PARMA: A Phy/Mac aware routing metric for ad-hoc wireless networks with multi-rate radios. *Proceedings of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks(WoWMoM)*, pp. 286 - 292, ISBN: 0-7695-2342-0, June 2005.
- Zhou, G. ; He, T. ; Krishnamurthy, S. & Stankovic, J. A. (2004). Impact of radio irregularity on wireless sensor networks, *Proceedings of the 2nd international conference on Mobile systems, applications, and services (MobiSys)*, pp. 125-138, June 6-9, 2004, Hyatt Harborside, Boston, Massachusetts, USA, USENIX, 2004.

Mobile WiMAX Performance Investigation*

Alessandro Bazzi, Giacomo Leonardi, Gianni Pasolini and Oreste Andrisano
*WiLab, IEIIT-BO/CNR, DEIS-University of Bologna
Italy*

1. Introduction

The IEEE802.16-2004 Air Interface standard (IEEE Std 802.16-2004, 2004), which is the basis of the WiMAX technology, is the most recent solution for the provision of fixed broadband wireless services in a wide geographical scale and proved to be a real effective solution for the establishment of wireless metropolitan area networks (WirelessMAN).

On February 2006, the IEEE802.16e-2005 amendment (IEEE Std 802.16e-2005, 2006) to the IEEE802.16-2004 standard has been released, which introduced a number of features aimed at supporting also users mobility, thus originating the so-called Mobile-WiMAX profile. Currently IEEE802.16 Task Group (TG) and WiMAX Forum are developing the next generation Mobile-WiMAX that will be defined in the future IEEE802.16m standard (Ahmadi, 2009; Li et al., 2009).

Although the Mobile-WiMAX technology is being deployed in the United States, Europe, Japan, Korea, Taiwan and in the Mideast, there are still ongoing discussions about the potential of this technology. What is really remarkable, in fact, with regard to the Mobile-WiMAX profile, is the high number of degrees of freedom that are left to manufacturers. The final decision on a lot of very basic and crucial aspects, such as, just to cite few of them, the bandwidth, the frame duration, the duplexing scheme and the up/downlink traffic asymmetry, are left to implementers. It follows that the performance of this technology is not clear yet, even to network operators.

This consideration motivated our work, which is focused on the derivation of an analytical framework that, starting from system parameters and implementation choices, allows to evaluate the performance level provided by this technology, carefully taking all aspects of IEEE802.16e into account. In particular, the analysis starts from the choices to be made at the physical layer, among those admitted by the specification, and "goes up" through the protocol pillar to finally express the application layer throughput and the number of supported voice over IP (VoIP) users, carefully considering "along the way" all characteristics of the medium access control (MAC) layer, the resource allocation strategies, the overhead introduced, the inherent inefficiencies, etc.

Let us remark that the analytical framework described in the following can be used not only as a mean to gain an insight into the IEEE802.16e performance, but, above all, to drive the choices of network operators in terms of system configuration. This is particularly true considering that beside the model derivation, here we provide criteria, equations and algorithms to make the best choices from the viewpoint of the system efficiency.

*Portions reprinted, with permission, from Proceedings of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2007 (PIMRC 2007). ©2007 IEEE.

2. IEEE802.16 overview

Before starting our analysis let us introduce the most relevant characteristics of the IEEE802.16 technology, that are recalled hereafter.

The result of the IEEE802.16 TG/WiMAX Forum activity is a complete standard family (IEEE Std 802.16-2004, 2004; IEEE Std 802.16e-2005, 2006) that specifies the air interface for both fixed and mobile broadband wireless access systems, thus enabling the convergence of mobile and fixed broadband networks through a common wide area broadband radio access technology and a flexible network architecture.

The IEEE802.16 standard family supports four transmission schemes:

- WirelessMAN-SC, which has been mainly developed for back-hauling in line-of-sight (LOS) conditions and operates in the 10 GHz - 66 GHz frequency range adopting a single carrier modulation scheme;
- WirelessMAN-SCa, which has the same characteristics of WirelessMAN-SC but operates even in non-LOS conditions in frequency bands below 11 GHz;
- WirelessMAN-OFDM, which has been developed for fixed wireless access in non-LOS conditions and adopts the orthogonal frequency division multiplexing (OFDM) modulation scheme in frequency bands below 11 GHz;
- WirelessMAN-OFDMA, which has been conceived for mobile access and adopts the orthogonal frequency division multiple access (OFDMA) scheme in the 2 GHz - 6 GHz frequency range.

Since we are interested in the mobility enhancement provided by the IEEE802.16e amendment, here we focus our attention on the WirelessMAN-OFDMA transmission scheme.

WirelessMAN-OFDMA is based on the OFDMA multiple-access/multiplexing technique which is, on its turn, based on an N_{FFT} subcarriers OFDM modulation scheme (Cimini, 1985; Van Nee & Prasad, 2000) with N_{FFT} equal to 128, 512, 1024 or 2048.

The N_{FFT} subcarriers form an OFDM symbol and can be further divided into three main groups:

- data subcarriers, used for data transmission;
- pilot subcarriers, used for estimation and synchronization purposes;
- null subcarriers, not used for transmission: guard subcarriers and DC subcarrier.

Considering sequences of OFDM symbols, it is easy to understand that transmission resources are available both in the time domain, by means of groups of consecutive OFDM symbols, and in the frequency domain, by means of groups of subcarriers (subchannels); it follows that a given mobile station can be allocated one or more subchannels for a specified number of symbols.

Several different schemes (in the following, permutation schemes) for subcarriers grouping are provided by the specification, with different possibilities for the downlink and uplink phases: among them we can cite DL-FUSC (downlink full usage of subchannels), DL-PUSC (downlink partial usage of subchannels), DL-TUSC (downlink tile usage of subchannels), or UL-PUSC (uplink partial usage of subchannels) (see the first column of table 1 for a complete list).

The minimum OFDMA time-frequency resource that can be allocated is one OFDMA-slot, which corresponds to 48 data subcarriers that can be accommodated in one, two or three OFDMA symbols, depending on which kind of permutation scheme (DL-FUSC, DL-PUSC, UL-PUSC, ...) is adopted; in particular:

Permutation scheme	Available subchannels N_{Ch}				N_{GS}
	N_{FFT}	N_{FFT}	N_{FFT}	N_{FFT}	
	128	512	1024	2048	
DOWNLINK					
DL-FUSC	2	8	16	32	1
DL-PUSC	3	15	30	60	2
DL-OptFUSC	2	8	16	32	1
DL-TUSC1	4	17	35	70	3
DL-TUSC2	4	17	35	70	3
UPLINK					
UL-PUSC	4	17	35	70	3
UL-OptPUSC	4	17	35	70	3

Table 1. Permutation schemes' parameters.

- with DL-FUSC a subchannel is constituted by 48 subcarriers in each OFDM symbol (hence, one OFDMA-slot covers one symbol);
- with DL-PUSC a subchannel is constituted by 24 subcarriers in each OFDM symbol (hence, one OFDMA-slot covers two symbols);
- with UL-PUSC a subchannel is constituted by 12 subcarriers in the first OFDM symbol, 24 subcarriers in the second OFDM symbol, 12 subcarriers in the third OFDM symbol and so on, according to the sequence 12-24-12-12-24-12....In this case one OFDMA-slot covers three symbols.

Since seven fixed combinations of modulation scheme and coding rate R_c , hereafter denoted as transmission modes, are provided by the IEEE802.16e physical layer, it follows that a single OFDMA-slot allows to transmit differently sized payloads (see table B in figure 1).

Both time division duplex (TDD) and frequency division duplex (FDD) are supported. However, the initial release of Mobile-WiMAX certification profiles only includes TDD, since it makes resource allocation more flexible (the downlink/uplink ratio can be easily adjusted to support asymmetric DL/UL traffic); for this reason, here we only consider the TDD duplexing scheme.

The TDD frame structure is depicted in the bottommost part of figure 1. Each TDD frame is divided into downlink and uplink subframes, separated by transmit/receive and receive/transmit transition gaps (TTG and RTG).

Each subframe may include "multiple zones", which means that the permutation method can be changed, thus moving, for instance, from DL-PUSC to DL-FUSC.

Focusing on the TDD frame, the first OFDM symbol of the DL subframe always carries a preamble, while a number of subsequent OFDM symbols are necessarily allocated to accommodate MAC layer control messages (FCH, DL-MAP and UL-MAP) adopting the DL-PUSC permutation scheme; similarly, a number of OFDM symbols are necessarily allocated in the UL subframe to accommodate several common signalling channels (e.g. UL Ranging, UL CQICH, UL ACK CH) adopting the UL-PUSC permutation scheme (see figure 1).

Finally, in order to correctly manage each data flow giving an acceptable quality of service to the end user, IEEE802.16e-2005 provides five different scheduling services for traffic delivery:

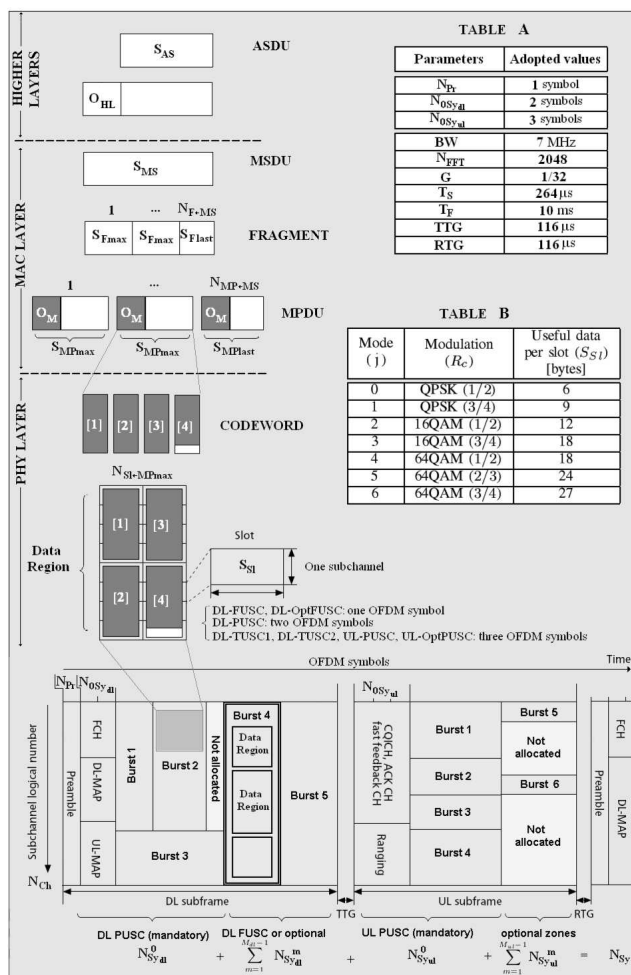


Fig. 1. IEEE802.16e WirelessMAN-OFDMA data processing and parameters setting.

- Unsolicited Grant Service (UGS),
- Real-Time Polling Service (rtPS),
- Extended Real-Time Polling Service (ertPS),
- Non-Real-Time Polling Service (nrtPS),
- Best Effort (BE).

Each scheduling service is associated with a set of quality of services (QoS) parameters: (a) maximum sustained rate, (b) minimum reserved rate, (c) maximum latency tolerance, (d) jitter tolerance and (e) traffic priority. These are the basic inputs for the service scheduler placed in the base station, which is aimed at fulfilling service specific QoS requirements. The main

differences among these services are on the uplink resource allocation; resource allocation is, in fact, defined by the base station, which cannot have a perfect knowledge of all uplink buffers in any instant. The interested reader can find a detailed description of the scheduling services in (IEEE Std 802.16-2004, 2004) and (IEEE Std 802.16e-2005, 2006).

3. Transmission resources: OFDMA-slots

In this section the amount of resources that are available for data transmission is evaluated as a function of all parameters that can be chosen by system implementers. In particular, the OFDMA-slot, which is the minimum resource available at the physical layer for data allocation, is focused and the amount of available OFDMA-slots is derived as a function of the physical layer configuration.

In order to ease the reader's task, the scheme reported in figure 2 summarizes the analytical framework outlined in this section, which lead to the assessment of the amount of available OFDMA-slots.

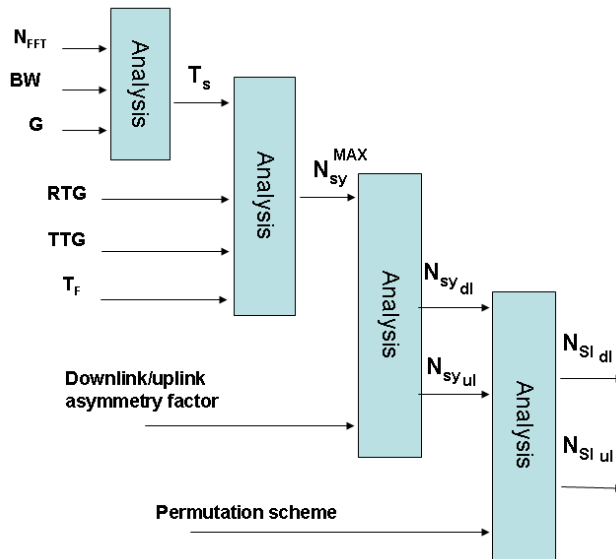


Fig. 2. Analytical framework for the derivation of the amount of OFDMA-slots per downlink/uplink subframe.

3.1 OFDM symbol duration

In order to assess the amount of resources available for data transmission at the physical layer, the OFDM symbol duration T_s must be obtained at first. T_s depends on the transmission bandwidth BW (it is typically a multiple of 1.75 MHz or 1.25 MHz), the number N_{FFT} of OFDM subcarriers (equal to 128, 512, 1024 or 2048) and the normalized (to the useful symbol duration) guard interval G (equal to 1/4, 1/8, 1/16 or 1/32).

Given BW , the value of an auxiliary parameter n (called sampling factor), introduced by the specification, can be immediately derived: in particular, $n = 28/25$ if BW is a multiple of 1.25 MHz, 1.5 MHz, 2 MHz or 2.75 MHz, otherwise $n = 8/7$.

Once BW and n are known, we can derive the sampling frequency F_s , which is defined by the specification as follows:

$$F_s = \left\lfloor n \cdot \frac{BW}{8000} \right\rfloor \cdot 8000, \quad (1)$$

having denoted with $\lfloor x \rfloor$ the highest integer not greater than x .

Given the number N_{FFT} of OFDM subcarriers, the subcarriers spacing Δf and the useful OFDM symbol duration T_u can be immediately derived from the knowledge of F_s :

$$\Delta f = \frac{F_s}{N_{FFT}}, \quad T_u = \frac{1}{\Delta f}. \quad (2)$$

The guard time interval T_g and, finally, the OFDM symbol duration T_s follow:

$$T_g = G \cdot T_u, \quad T_s = T_u + T_g. \quad (3)$$

3.2 Number of OFDM symbols per frame

Once T_s has been obtained, the second step to derive the amount of OFDMA-slots available at the physical layer is to assess the number of useful OFDM symbols in a frame.

Since we are interested in the TDD version of IEEE802.16e WirelessMAN-OFDMA, we have to consider a frame structure consisting of two parts, that represent the downlink and uplink subframes, separated by the TTG and RTG time intervals (see figure 1).

In order to derive the number of useful OFDM symbols in a frame, let us recall that the first symbol of the frame is used to transmit the preamble (thus the number of preamble symbols is $N_{Pr} = 1$) and that both TTG and RTG cannot be smaller than $5 \mu s$ ($RTG_{min} = TTG_{min} = 5 \mu s$). Thus, once the frame duration T_F has been chosen (possible values admitted by the specification are 2, 2.5, 4, 5, 8, 10, 12.5 and 20 ms), and given the previously derived value of T_s , the maximum number N_{Sy}^{MAX} of OFDM symbols per frame (excluding the preamble) can be derived:

$$N_{Sy}^{MAX} = \left\lfloor \frac{T_F - TTG_{min} - RTG_{min}}{T_s} \right\rfloor - N_{Pr}, \quad (4)$$

as depicted in figure 2.

3.3 Number of OFDMA-slots per frame

Given the value of N_{Sy}^{MAX} , it is now possible to assess the number of OFDMA-slots that can be allocated in the downlink and uplink subframes. Since OFDMA-slots extend both in the time and in the frequency domains, the derivation of their amount requires considerations on both domains.

As for the time domain, let us recall that, depending on the adopted permutation scheme (DL-FUSC, DL-PUSC, UL-PUSC, ...), an OFDMA-slot is spread over $N_{GS} = 1, 2$ or 3 consecutive OFDM symbols, as reported in the last column of table 1, whereas, as far as the frequency domain is concerned, the amount of available subchannels N_{Ch} depends on the permutation scheme and the number N_{FFT} of OFDM subcarriers (as reported in the second column of table 1).

Please note that different permutation schemes are provided for the downlink and uplink subframes and that more than one scheme can be used in a single subframe. Each permutation

scheme (denoted in the following with the superscript m) requires the allocation of a multiple of an integer (1, 2 or 3, depending on the permutation scheme) number of OFDM symbols N_{Sy}^m ($N_{Sy_{dl}}^m$ in the downlink and $N_{Sy_{ul}}^m$ in the uplink, respectively) and the sum N_{Sy} of uplink and downlink symbols allocated for each permutation scheme is bounded by the above assessed N_{Sy}^{MAX} :

$$\sum_{m=0}^{M_{dl}-1} N_{Sy_{dl}}^m + \sum_{m=0}^{M_{ul}-1} N_{Sy_{ul}}^m = N_{Sy} \leq N_{Sy}^{MAX}, \quad (5)$$

where M_{dl} (M_{ul}) is the amount of permutation schemes adopted in the downlink (uplink) subframe.

As recalled in section 2, at least two OFDM symbols are allocated with DL-PUSC in the downlink subframe, in order to carry frame management messages, while three OFDM symbols are allocated with UL-PUSC in the uplink subframe, in order to carry signalling common channels; here we assume that the entire first two OFDM symbols in downlink (with DL-PUSC) and the entire first three OFDM symbols in uplink (with UL-PUSC) are used for this scope, denoting the related overhead with $N_{OSy_{dl}} = 2$ and $N_{OSy_{ul}} = 3$.

Moreover, the rest of the sub-frame is supposed to be transmitted adopting only one permutation scheme. Thus, the superscript correspondent to the adopted permutation scheme will be omitted in the following and (5) is rearranged as follows:

$$N_{OSy_{dl}} + N_{Sy_{dl}} + N_{OSy_{ul}} + N_{Sy_{ul}} = N_{Sy} \leq N_{Sy}^{MAX}, \quad (6)$$

where $N_{Sy_{dl}}$ and $N_{Sy_{ul}}$ now represent the amount of downlink/uplink symbols available for user data.

Let us observe that the choice of $N_{Sy_{dl}}$ and $N_{Sy_{ul}}$ is not only constrained to fulfill (6), but is also a consequence of the desired asymmetry between the downlink and uplink phases of the TDD frame, hereafter referred to as "desired asymmetry factor" and denoted as AF_{in} .

Let's keep in mind, in this regard, that the minimum resource that can be allocated is the OFDMA-slot and that the number of slots in a subframe is related not only to the number of OFDM symbols within the frame, but also to the adopted permutation scheme; as an example, with $N_{FFT} = 2048$ subcarriers two OFDM symbols adopting the DL-PUSC permutation scheme carry 60 slots while three OFDM symbols adopting UL-PUSC carry 70 slots (refer to (8) and table 1). Thus, defining AF_{Sl} as the asymmetry factor in terms of ratio between the amounts of downlink and uplink slots:

$$AF_{Sl} = \frac{N_{Sl_{dl}}}{N_{Sl_{ul}}}, \quad (7)$$

and deriving the amount of OFDMA-slots available for data transmission in the downlink/uplink subframe through the equation (where dl/ul denotes downlink or uplink as alternatives):

$$N_{Sl_{dl/ul}} = N_{Ch_{dl/ul}} \cdot \left\lfloor \frac{N_{Sy_{dl/ul}}}{N_{GS_{dl/ul}}} \right\rfloor, \quad (8)$$

the desired asymmetry AF_{in} can be approached finding the values $N_{Sy_{dl}}$ and $N_{Sy_{ul}}$ that make AF_{Sl} as near as possible to AF_{in} . In general, a perfect matching between AF_{in} and AF_{Sl} will be not possible, due to the system constraints.

The detection of $N_{Sy_{dl}}$ and $N_{Sy_{ul}}$ in such a way to minimize the resource wasting (that is, OFDM symbols within the frame that are unused because unable to accomodate entire

OFDMA-slots) for a given AF_{in} can be carried out by means of the algorithm provided in appendix I.

Equation (8) represent the final outcome of this section since, jointly with the constraints given by the desired asymmetry factor and system choices (bandwidth, guard interval, number of subcarriers, frame duration,) accounted for by the previous equations, allows to derive the amount of OFDMA-slots available in each subframe for data transmissions.

Please refer to figure 2 for a pictorial representation of the whole methodology described in this section.

4. From application layer packets to subcarriers allocation

In the previous section the amount $N_{S_{dl}}$ and $N_{S_{ul}}$ of OFDMA-slots available for data allocation have been derived, taking into account all the physical and MAC layers parameters. The next step is to understand how packets to be transmitted, coming from the higher protocol layers, are mapped onto these resources. A brief overview on the packet processing is given hereafter, followed by an analytical evaluation of the number of OFDMA-slots that are finally needed to allocate each packet.

4.1 Packet processing overview

The data mapping process, starting from the application layer data unit down to the physical layer, is illustrated step by step hereafter. Please refer to figure 1, where each step is depicted, to better understand the whole process.

Let us denote as *ASDU* the application level data fragment of S_{AS} bytes that is allocated into the payload of a TCP/IP packet. Each *ASDU* is firstly added with O_{HL} bytes, where O_{HL} represents the overhead added from the application to the network layer, and then mapped, at the MAC layer, onto a MAC service data unit (*MSDU*) of S_{MS} bytes. Each *MSDU* is then partitioned into fragments of S_{Fmax} bytes, whose value is negotiated during the connection setup phase; obviously the last fragment of each *MSDU* may be smaller (S_{Flast} bytes). If the ARQ mechanism is active fragments are also called ARQ blocks.

One or more fragments are then allocated into a MAC protocol data unit (MPDU), with some overhead: in particular, a MAC header will be added plus either (a) one fragmentation subheader if all fragments are contiguous and related to the same *MSDU* or (b) a packetization subheader per each group of contiguous fragments belonging to the same *MSDU*; a CRC (cyclic redundancy check) tail of 32 bits will be added at the end of the MPDU, in order to check its integrity at the receiver side.

At the physical layer, MPDUs are partitioned into groups of bytes that are subject to the forward error correction coding process, giving birth to a certain number of codewords. One or more OFDMA-slots can be combined in order to convey each codeword. Adjacent slots, both in the time and subchannels domain, are grouped into OFDMA data regions, which are two-dimensional (squared or rectangular) allocations of a group of contiguous subchannels in a group of contiguous OFDM symbols (see figure 1).

4.2 From application layer data to MPDUs

After the data processing overview provided above, in this subsection we analytically derive the amount of OFDMA-slots needed to deliver an *ASDU*. Having derived (section 3) the amount of OFDMA-slots available in the uplink/downlink subframes of a TDD Mobile-WiMAX system, this is the second step along the path that leads to the Mobile-WiMAX performance assessment.

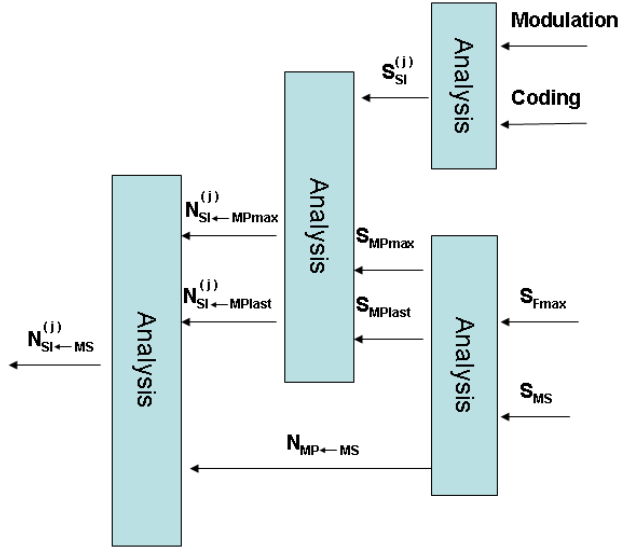


Fig. 3. Diagram of the calculation from the packet size to the number of OFDMA-slots that are needed to accomodate it.

Also in this case, in order to help the reader, the analytical framework outlined in the following has been summarized in a pictorial fashion, reported in figure 3.

Let us consider the aforementioned *ASDUs* of S_{AS} bytes; as represented in figure 1 they eventually arrive at the MAC layer with the addition of the higher layers overheads of O_{HL} bytes, thus originating MAC layer service data units (*MSDUs*) of S_{MS} bytes:

$$S_{MS} = S_{AS} + O_{HL}. \quad (9)$$

The *MSDUs* are then fragmented into a fixed number $N_{F \leftarrow MS}$ of fragments, each of them of size S_{Fmax} except, in case, the last one. This one has a size of $S_{Flast} \leq S_{Fmax}$ bytes; thus:

$$N_{F \leftarrow MS} = \left\lceil \frac{S_{MS}}{S_{Fmax}} \right\rceil, \quad (10)$$

$$S_{Flast} = S_{MS} - [(N_{F \leftarrow MS} - 1) \cdot S_{Fmax}] \quad (11)$$

where $\lceil x \rceil$ indicates the lowest integer not less than x .

It follows that each *MSDU* is carried at the MAC layer by:

- $(N_{F \leftarrow MS} - 1)$ fragments of size S_{Fmax} ,
- 1 fragment of size S_{Flast} .

Of course, if S_{MS} is a multiple of S_{Fmax} , then $S_{Flast} = S_{Fmax}$.

Let us assume now, for the sake of simplicity, that no packetization is performed at the MAC layer; each fragment is therefore mapped onto one *MPDU* with the addition of the MAC layer overhead of O_M bytes. It follows that the number $N_{MP \leftarrow MS}$ of *MPDUs* needed to carry a single *MSDU* is equal to $N_{F \leftarrow MS}$. Since all but (in case) the last *MPDU* have the same size, we have:

- $(N_{MP \leftarrow MS} - 1)$ *MPDUs* of size S_{MPmax} ,
- 1 *MPDU* of size S_{MPlast} ,

where:

$$\begin{aligned} N_{MP \leftarrow MS} &= N_{F \leftarrow MS}, \\ S_{MPmax} &= S_{Fmax} + O_M, \\ S_{MPlast} &= S_{Flast} + O_M. \end{aligned} \quad (12)$$

Of course, if $N_{MP \leftarrow MS} = 1$, each *MPDU* carries a complete *MSDU* and its size is S_{MPlast} .

4.3 MPDUs into OFDMA-slots

Starting from the results obtained in subsection 4.2, we can now derive the number of OFDMA-slots needed to carry any *MPDU* and, as a consequence, any *MSDU*.

Let us recall that every transmission mode j can convey a different amount $S_{Sl}^{(j)}$ of data bytes into a single OFDMA-slot (see table B in figure 1); since there are two possible sizes for *MPDUs* (S_{MPmax} and S_{MPlast}), we can derive, for every transmission mode j , the minimum number of slots needed to carry each of them:

$$\begin{aligned} N_{Sl \leftarrow MPmax}^{(j)} &= \left\lceil \frac{S_{MPmax}}{S_{Sl}^{(j)}} \right\rceil, \\ N_{Sl \leftarrow MPlast}^{(j)} &= \left\lceil \frac{S_{MPlast}}{S_{Sl}^{(j)}} \right\rceil. \end{aligned} \quad (13)$$

These equations show that when the *MPDU* size is not a multiple of the amount of bytes carried by a single OFDMA-slot, some padding bits have to be added in order to fill the last slot, wasting some resources.

Thus, a single *MSDU* is transmitted with the generic transmission mode j through:

- $(N_{MP \leftarrow MS} - 1)$ *MPDUs* of size S_{MPmax} , accommodated into $(N_{MP \leftarrow MS} - 1) \cdot N_{Sl \leftarrow MPmax}^{(j)}$ slots,
- 1 *MPDU* of size S_{MPlast} , accommodated into $N_{Sl \leftarrow MPlast}^{(j)}$ slots.

Assuming no resource wastage due to data regions' allocations ¹, the number $N_{Sl \leftarrow MS}^{(j)}$ of OFDMA-slots needed to carry a complete *MSDU* adopting transmission mode j is given by:

$$N_{Sl \leftarrow MS}^{(j)} = ((N_{MP \leftarrow MS} - 1) \cdot N_{Sl \leftarrow MPmax}^{(j)}) + N_{Sl \leftarrow MPlast}^{(j)}. \quad (14)$$

Recalling that the scope of the analysis reported in this section was to derive the amount of OFDMA-slots needed to accommodate an *ASDU*, we can state that (14) is the final outcome of this section since, jointly with the equations reported in subsection 4.2, it achieves our end.

5. System performance: throughput of a TCP connection

Having derived the resources (that is, OFDMA-slots) available for data allocation at the physical layer (in section 3) and the amount of resources needed to carry each *ASDU* (in section 4), we can now assess the performance level provided by IEEE802.16e for a given configuration.

¹ Data regions must be squared or rectangular.

In this section, in particular, a TCP connection is considered, requiring a best effort service. In order to evaluate the maximum end-to-end throughput at the application layer (hereafter simply denoted as throughput), a single connection is supposed to be active either in the downlink or in the uplink. Furthermore, the return link (the uplink when considering a downlink TCP flow and vice versa) is considered to be always sufficient for the transmission of TCP acknowledgments; this assumption implies that the analysis is focused on the direction of the data flow.

Since a single link (uplink or downlink) is here considered for throughput derivation, in the following the subscript *dl* or *ul* will be omitted. Similarly, since a single (generic) transmission mode is considered, the superscript *j* will also be omitted.

In section 3.3 we evaluated the number N_{Sl} of OFDMA-slots available in a given direction (uplink or downlink) and in section 4.3 we derived the number $N_{Sl \leftarrow MS}$ of OFDMA-slots needed to carry a complete *MSDU*, that is, a complete *ASDU*; it follows that the number N_{MS} of complete *MSDU*s that can be allocated in the considered subframe will be, therefore:

$$N_{MS} = \left\lfloor \frac{N_{Sl}}{N_{Sl \leftarrow MS}} \right\rfloor. \quad (15)$$

Since, in general, N_{Sl} is not a multiple of $N_{Sl \leftarrow MS}$, it follows that R_{Sl} slots will remain available for the allocation of *MPDUs* that do not form a complete *MSDU*, where $R_{Sl} = \text{mod}(N_{Sl}, N_{Sl \leftarrow MS})$, having denoted with $\text{mod}(A, B)$ the remainder of the division $\frac{A}{B}$.

In particular, the number of *MPDUs* of size S_{MPmax} (which occupy $N_{Sl \leftarrow MPmax}$ slots each) that can be accommodated in R_{Sl} free slots is given by:

$$N_{MPinR_{Sl}} = \left\lfloor \frac{R_{Sl}}{N_{Sl \leftarrow MPmax}} \right\rfloor. \quad (16)$$

In general, after the allocation of $N_{MPinR_{Sl}}$ *MPDUs* of size S_{MPmax} , a residual amount RR_{Sl} of slots will remain available, with $RR_{Sl} = \text{mod}(R_{Sl}, N_{Sl \leftarrow MPmax})$; of course, another *MPDU* of size S_{MPmax} cannot be allocated, but it could be possible to accommodate an *MPDU* of size S_{MPlast} occupying $N_{Sl \leftarrow MPlast}$ slots.

If we denote with $N_{MSextra}$ the fraction of *MSDU* that, in average, can be allocated in the residual region of size R_{Sl} , it results:

$$N_{MSextra} = \begin{cases} \frac{N_{MPinR_{Sl}}}{N_{MP \leftarrow MS}} & \text{if } RR_{Sl} < N_{Sl \leftarrow MPlast} \\ \frac{N_{MPinR_{Sl}}}{N_{MP \leftarrow MS} - 1} & \text{else.} \end{cases} \quad (17)$$

Obviously, if $N_{MPinR_{Sl}} = 0$ it results $N_{MSextra} = 0$.

Let us recall that the amount of application layer data conveyed by a single *MSDU* is given by S_{AS} bytes; it follows that the average amount of application layer data accommodated in a given subframe is:

$$S = (N_{MS} + N_{MSextra}) \cdot S_{AS}, \quad (18)$$

The application layer throughput can be thus immediately derived as:

$$Thr_A[\text{bit/s}] = \frac{8 \cdot S[\text{bytes}]}{T_F[\text{s}]}, \quad (19)$$

where T_F represents the frame duration (see section 3.2).

5.1 Numerical results

In this section some numerical results obtained through (19) are given. ASDUs of $S_{AS} = 1460$ bytes were chosen, since this is the payload size of a typical TCP/IP packet. Considering 20 bytes for the IP overhead, 20 bytes for the TCP overhead and neglecting the overhead introduced by the upper layers, we assumed that each MSDU has a size of $S_{MS} = S_{AS} + O_{HL} = 1500$ bytes.

A further overhead of $O_M = 10$ bytes is introduced by the MAC layer, following the assumption of no packetization.

The OFDM modulation parameters were set to $N_{FFT} = 2048$, $BW = 7$ MHz, $G = 1/32$; moreover a frame duration of $T_F = 10$ ms has been chosen and $RTG = TTG = 116$ μ s were considered; all other physical layer parameters are consequently derived (e.g. $T_S = 264$ μ s).

The impact of the remaining parameters affecting the throughput will be investigated in the following. In particular, different values of S_{Fmax} and all transmission modes and permutations schemes will be considered.

In figure 4, a comparison between physical layer and application layer throughput is given varying the fragments maximum size S_{Fmax} . The physical layer throughput Thr_P has been evaluated considering the total amount of bits carried over the medium by all available OFDMA-slots, as follows:

$$Thr_P[\text{bit/s}] = \frac{N_{SI} \cdot 8 \cdot S_{SI}[\text{bytes}]}{T_F[\text{s}]}, \quad (20)$$

where the same notation introduced in section 4.2 has been adopted (please note that the previous equation considers only those resources available for data transmission, thus excluding the preamble symbol and the subcarriers used for signalling and control messages).

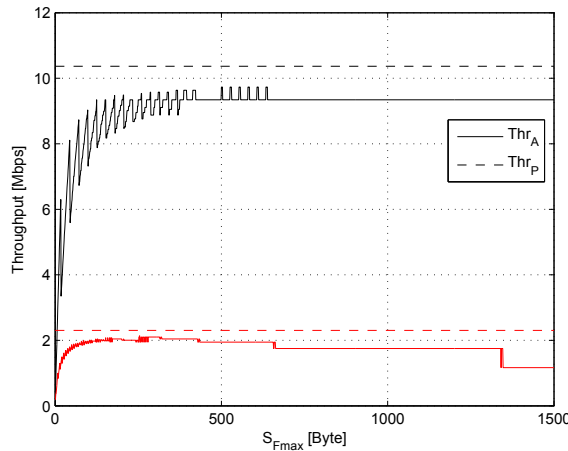


Fig. 4. Comparison between physical layer and application layer throughput (Thr_P and Thr_A) varying the fragment (ARQ block) maximum size S_{Fmax} . Transmission modes 0 and 6.

DL-PUSC has been considered in the downlink and UL-PUSC in the uplink, with $AF_{in} = 1$ and $AF_{SI} = 1.37$ (following the equations reported in section 3.3, we obtain $N_{Sy_{dl}} = 16$, $N_{Sy_{ul}} = 15$, $N_{SI_{dl}} = 480$ and $N_{SI_{ul}} = 350$). Transmission modes 0 and 6 are considered.

The comparison between the dashed and solid curves highlights the reduction of throughput due to both the allocation procedure of *ASDUs* and the overhead. As can be noted, small variations in the choice of S_{Fmax} may affect the system performance, due to the slot granularity in the physical resource allocation and the impossibility to further divide a fragment (i.e., an ARQ block).

These curves also show that a too small value of S_{Fmax} should not be chosen (this is mainly a consequence of the presence of the overheads O_{HL} and O_M). However, although considering error prone transmissions is out of the scope of the present work, it should be clear that a large value of S_{Fmax} should be avoided too, since each ARQ block must be entirely retransmitted if not correctly received.

Figure 5 deepens the previous results focusing the attention on the application layer throughput and considering all transmission modes.

A direct comparison of the throughput perceived adopting the different transmission modes as a function of S_{Fmax} shows that the choice of S_{Fmax} is a tricky task, since there is no optimal value providing the maximum throughput for all transmission modes.

As can be observed, a number of choices for S_{Fmax} are highlighted through vertical lines and the correspondent throughput values with circles. These values are somehow suboptimal and have been chosen according to the following steps:

1. for each value of S_{Fmax} in the range [1, 500 bytes] the throughput values achieved by each transmission mode normalized to the peak value for that mode were summed into $SUM_{Thr}(S_{Fmax})$;
2. the values of S_{Fmax} that brought to relative maximum of $SUM_{Thr}(S_{Fmax})$ were found, neglecting those values of S_{Fmax} that do not give an absolute value of $SUM_{Thr}(S_{Fmax})$ higher than the previous one.

The values of S_{Fmax} derived as previously described ($S_{Fmax} = 98, 125, 134, 152, 206, 254, 422$ bytes) allow to reduce resource wasting when a single fragment (ARQ block) is transmitted in a single *MPDU*.

In figure 6 the value of AF_{Sl} is compared to AF_{App} , which is defined as the ratio between the maximum application layer throughput in downlink and the one in uplink. The DL-PUSC and UL-PUSC permutation schemes have been considered in the downlink and in the uplink, respectively; $AF_{in} = 1$, $AF_{in} = 2$ and $AF_{in} = 3$ have been assumed as desired asymmetry factors. Transmission mode 6 only.

As can be noted, a good match between AF_{Sl} and AF_{App} is achieved for all the considered AF_{in} (avoiding to consider too large values for S_{Fmax}). On the contrary, it is quite hard to exactly respect the desired AF_{in} with no wasting: note, in fact, that the cases $AF_{in} = 1$ and $AF_{in} = 2$ bring to the same result (that is, the need to minimize the resource wasting brings, in both cases, to the same choice of $N_{Sy_{dl}}$ and $N_{Sy_{ul}}$).

In figure 7 the throughput is shown as a function of the number of OFDM symbols available for data transmission in the downlink subframe for all possible permutation schemes (refer to table 1). In this case, $N_{Sy_{dl}}$ is set and the correspondent AF_{Sl} follows as a consequence (see section 3.3). Transmission mode 6 and $S_{Fmax} = 206$ bytes have been considered. This figure also highlights that an increase (reduction) in $N_{Sy_{dl}}$ has an effect only if it involves at least N_{GS} symbols.

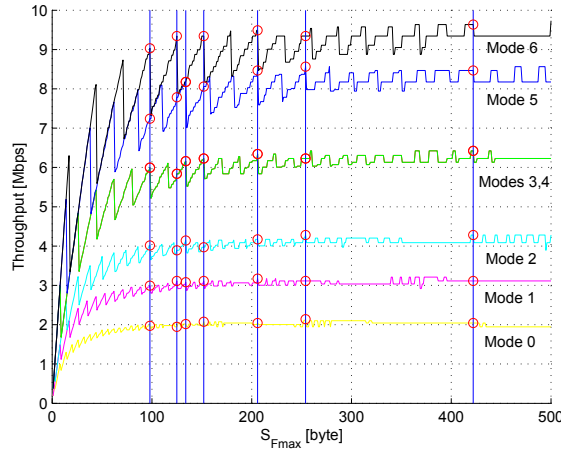


Fig. 5. Application layer throughput (Thr_A) varying the fragments (ARQ blocks) maximum size S_{Fmax} for all transmission modes. The values of S_{Fmax} that allow to have a good occupation adopting any possible transmission mode are marked with vertical lines and small circles (o); they correspond to $S_{Fmax} = 98, 125, 134, 152, 206, 254, 422$ bytes

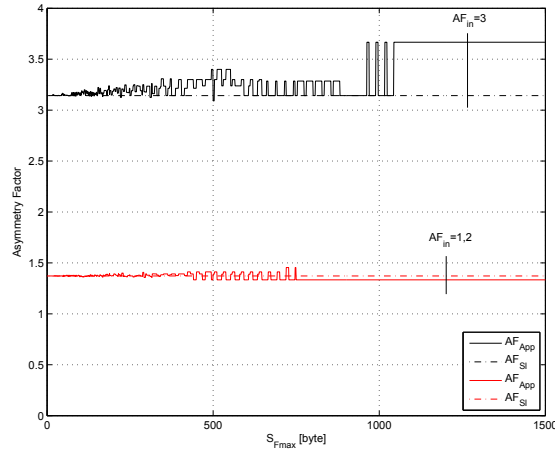


Fig. 6. Comparison between AF_{Sl} and AF_{App} , given $AF_{in} = 1$, $AF_{in} = 2$ and $AF_{in} = 3$. Transmission mode 6.

6. System performance: VoIP capacity on UGS or ertPS

In this section the maximum number of users performing a VoIP call that can be served by IEEE802.16e is evaluated, following the analysis described in sections 3 and 4.

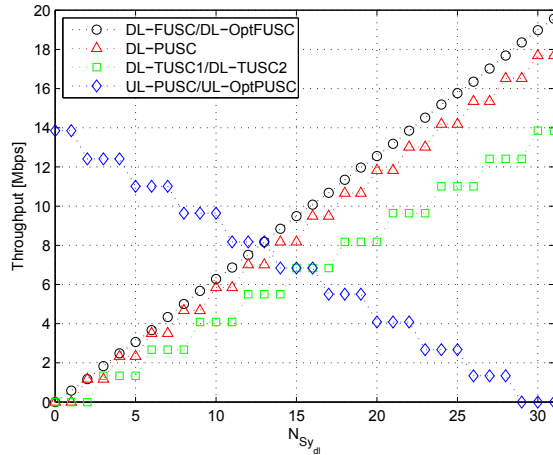


Fig. 7. Comparison of application layer throughput (Thr_A) adopting the various permutation schemes, varying the number of downlink useful symbols $N_{Sy_{dl}}$. Transmission mode 6. $S_{Fmax} = 206$ bytes.

A description of the considered VoIP codecs and scheduling services is given before entering into the details of the analytical model.

6.1 UGS and ertPS scheduling services

As already mentioned in section 2, five scheduling services are provided by the IEEE802.16e specification for traffic delivery. Since our attention is now focused on real-time VoIP traffic, UGS and ertPS are the only possible choices, due to latency constraints, and are therefore considered in the following:

- **Unsolicited grant service (UGS)** is designed to support real-time uplink service flows that generate transport fixed-size data packets on a periodic basis, such as T1/E1 and VoIP without silence suppression. The service offers fixed size grants on a real-time periodic basis, which eliminate the overhead and latency of user's requests and assure that grants are available to meet the flows' real-time needs.
- **Extended real-time polling service (ertPS)** (Lee et al., 2006) improves UGS when the application layer rate varies in time. The base station (BS) shall provide unicast grants in an unsolicited manner like in UGS, thus saving the latency of a bandwidth request. However, whereas UGS allocations are fixed in size, ertPS allocations are dynamic. The BS may provide periodic uplink allocations that may be used for requesting the bandwidth as well as for data transfer. By default, size of allocations corresponds to current value of maximum sustained traffic Rate at the connection. Users may request changing the size of the uplink allocation by either using an extended piggyback request field of the grant management subheader or using BR field of the MAC signaling headers, or sending a specific codeword over the signalling channel CQICH. The BS shall not change the size of uplink allocations until receiving another bandwidth change request from the user.

6.2 VoIP codecs

The most important and mainly adopted voice codecs have been considered:

1. **ITU G.711** (ITU-T Rec. G.711, 1988), the well known constant bit rate PCM at 64 kbps; this is the codec used in PSTN networks, with no compression, neither during the speech nor during silences of a conversation;
2. **ITU G.729** (ITU-T Rec. G.729, 1996a), the most used codec for VoIP, at 8 kbps; when an active speech period is detected, it produces one packet of 80 bits every 10 ms, but more than one packet may be concatenated in order to reduce protocols overheads (Goode, 2002); obviously, this process enlarges the average delivery delay of packets. Hereafter, we will consider the concatenation of a couple of packets (Goode, 2002) and we will denote this codec as *G.729*. 160 bits packets are thus generated every 20 ms.
3. **AMR** (adaptive multi rate (3GPP TS 26.071, 2008)), standardized by 3GPP and used for voice in second and third generation cellular radio access; it generates one packet every 20 ms with a variable data rate, going from a minimum of 4.75 kbps (95 data bits plus 18 overhead bits for each packet) to a maximum of 12.2 kbps (244 data bits plus 18 overhead bits for each packet). The variation of the data rate is given in order to better select the appropriate tradeoff between resource usage and speech quality (obviously, a data rate reduction leads to a quality degradation). In order to investigate the capacity of the IEEE802.16e WirelessMAN-OFDMA system with the AMR codec, we will consider both the minimum data rate (4.75 kbps, denoted as *AMR4.75*) and the maximum data rate (12.2 kbps, denoted as *AMR12.2*).

ITU G.729 and *AMR* codecs have been designed, in particular, with the specific goal to reduce the resources occupation: when no voice activity is detected the *silence suppression* procedure is activated. As a consequence small and less frequent packets are transmitted, which convey the information for a “comfortable noise” generation at the receiver side.

In particular, adopting the *AMR* codec, the detection of a silence period (3GPP TS 26.092, 2008) gives rise to the following steps:

- eight full voice packets are normally transmitted in the first interval, called *hangover period*;
- then, a SID (silence insert descriptor) packet (36 data bits plus 18 overhead bits) is transmitted after 60 ms;
- further SID packets are transmitted every 160 ms until a new speech activity is detected.

This process is depicted in the topmost part of figure 8.

As far as the *G.729* codec is concerned, the silence suppression procedure is defined in the Annex B (Benyassine et al., 1997; ITU-T Rec. G.729, 1996b). In this case, each SID packet has a length of 15 bits. However, the time interval between two successive SID packets is not fixed: in fact, for a good quality at the receiver, a lower or a greater transmission rate may be needed depending on the specific background noise observed in each environment.

In order to allow the derivation of meaningful results, SID packets are here supposed to be generated with a fixed rate during silences, as with *AMR*. In particular, the rate of *G.729* SID packets is assumed to be twice the one adopted by *AMR*, following the results provided in (Estepa et al., 2005).

In the following, the *activity factor* v is defined as the ratio between the time during which full packets are generated and the total duration of the conversation. Thus, in particular, we

Codec	B	ρ	B_{SID}	ρ_{SID}	ν	Maximum number of allowed users	
	[bits]	[pack/s]	[bits]	[pack/s]		Mode 0	Mode 6
G.711	1280	50	-	-	1	20 / __	87 / __
G.729	160	50	-	-	1	58 / __	233 / __
G.729 _{ss}	160	50	15	12.5	0.45	58 / 90	233 / 482
AMR4.75	113	50	-	-	1	63 / __	233 / __
AMR4.75 _{ss}	113	50	54	6.25	0.45	63 / 111	233 / 506
AMR12.2	262	50	-	-	1	50 / __	175 / __
AMR12.2 _{ss}	262	50	54	6.25	0.45	50 / 90	175 / 374

Table 2. Codec parameters and analytical calculation of maximum number of allowed users. Multiple values refer to *UGS/ertPS*.

will adopt $\nu = 1$ when no silence suppression is activated and $\nu < 1$ in the case of silence suppression.

In the latter case, in order to derive a realistic value of ν , we simulated the dynamic of a conversation according to a detailed model that takes into account also periods of simultaneous talks or silences of the two parties, and the short gaps through the speeches (Stern et al., 1996). This model gives a 33% of voice activity over the total conversation duration for each of the two parties. It must be noted, however, that since 8 full packets are still transmitted during the hangover period, the activity factor ν is approximately 0.45; please note that this value corresponds to both our simulations and the experimental results given in (Estepa et al., 2005). As for the less sophisticated *G.711* codec, here it was considered only as a reference, and no silence suppression is introduced.

The parameters of all considered codecs, with and without silence suppression, are given in the first four columns of table 2. In particular: the first column indicates the considered codecs (the subscript *ss* indicates that silence suppression is considered); the second column defines the number of bits B of full packets and the number of full packets per second ρ that are transmitted when the voice is detected; similarly, the third column defines the number of bits B_{SID} of SID packets and the number ρ_{SID} of SID packets per second that are transmitted when silences are detected; finally, the fourth column represents the activity factor ν ; the rest of the table will be illustrated in the following.

6.3 Amount of supported VoIP users

In this section, the number of VoIP users that can be served will be evaluated as a function of the adopted codec x , the scheduling service k and the transmission mode j (all users are supposed to be served adopting the same mode). Also in this case we need to consider the whole packet processing from the application layer down to the transmission over the medium. Before being transmitted, each packet generated by the codec must pass through the whole protocol pillar, thus increasing its size owing to the overheads introduced by each protocol layer. In particular, the RTP, UDP, IP and MAC layers overheads are added, which are, respectively, $O_{RTP} = 12$ bytes, $O_{UDP} = 8$ bytes, $O_{IP} = 20$ bytes and $O_M = 10$ bytes (thus, $O_{HL} = O_{RTP} + O_{UDP} + O_{IP} = 40$ bytes). Assuming that no fragmentation is carried out from the application to the MAC layer, the size (in bytes) of full packets and SID packets generated by the codec x is given by:

$$S_{MS}^{(x)} = B^{(x)} / 8 + O_{HL} + O_M, \quad (21)$$

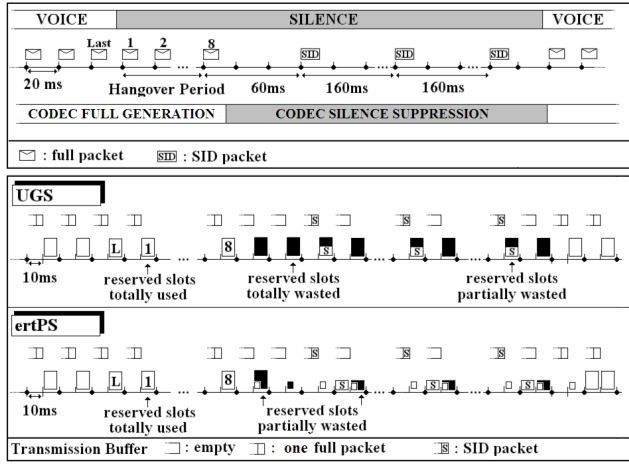


Fig. 8. Topmost part: AMR codec full packets and SID packets generation, with $\rho = 50 \text{ pack/s}$. Rest of the figure: buffer state and resource allocation of a generic uplink voice traffic flow with UGS and ertPS, following the packets generation depicted in the topmost part.

$$S_{MSID}^{(x)} = B_{SID}^{(x)} / 8 + O_{HL} + O_M. \quad (22)$$

Of course, the ARQ mechanism is assumed inactive, since we are considering a real time service.

Let us recall (from section 6.2), now, that for a given codec x :

$$\nu^{(x)} = \frac{\text{Time with full packets}}{\text{Total conversation duration}} \quad (23)$$

is the activity factor, while

$$\rho^{(x)} = \frac{\text{Number of full packets}}{\text{Time with full packets}} \quad (24)$$

is the rate of transmission of full packets (of $B^{(x)}$ bits) during voice activity and

$$\rho_{SID}^{(x)} = \frac{\text{Number of SID packets}}{\text{Time without full packets}} \quad (25)$$

is the rate of transmission of SID packets (of $B_{SID}^{(x)}$ bits) during silences, if silence suppression is considered.

As already recalled each packet to be transmitted is mapped onto OFDMA-slots, thus, in order to assess the maximum number of users that can be served, the number of slots that are needed for each packet must be firstly calculated, also considering that the number $S_{sl}^{(j)}$ of bytes carried by one slot depends on the adopted mode j (see table B of figure 1).

In particular, for a given codec x and a given transmission mode j , the number $N_{sl \leftarrow MS}^{(x,j)}$ of slots needed to carry a full packet and the number $N_{sl \leftarrow MSID}^{(x,j)}$ of slots needed to carry a SID

packet are:

$$N_{Sl \leftarrow MS}^{(x,j)} = \left\lceil \frac{S_{MS}^{(x)}}{S_{Sl}^{(j)}} \right\rceil, \quad (26)$$

$$N_{Sl \leftarrow MS_{SID}}^{(x,j)} = \left\lceil \frac{S_{MS_{SID}}^{(x)}}{S_{Sl}^{(j)}} \right\rceil. \quad (27)$$

Depending on the considered scheduling service k and the specific VoIP codec x , the average number of slots required in a given (DL/UL) subframe by a single user adopting mode j , is given by:

$$M_{Sl \leftarrow U}^{(x,k,j)} = f^{(x,k)} \left(N_{Sl \leftarrow MS}^{(x,j)}, N_{Sl \leftarrow MS_{SID}}^{(x,j)} \right), \quad (28)$$

where the analytical expression of $f^{(x,k)}(\cdot)$ will be provided in the following for both the considered scheduling services.

As a general consideration, please note that all difficulties in resource allocation are on the uplink, since the base station has a perfect knowledge of all buffers in the downlink and no resource requests are needed. For this reason, and assuming that $AF_{Sl} \geq 1$ (that is, we have more downlink slots than uplink slots), all evaluations will be done focusing on the uplink direction. Thus, given an amount $N_{Sl_{ul}}$ of available slots (in the uplink subframe), the maximum number $N_{U_{MAX}}$ of users can finally be evaluated:

$$N_{U_{MAX}}^{(x,k,j)} = \left\lfloor \frac{N_{Sl_{ul}}}{M_{Sl \leftarrow U}^{(x,k,j)}} \right\rfloor. \quad (29)$$

Performance on UGS. Since UGS resources are statically allocated, as negotiated during connection setup, the silence suppression procedure does not provide any benefit in the uplink direction.

Denoting with T_F the frame duration, (28) becomes:

$$M_{Sl \leftarrow U}^{(x,UGS,j)} = T_F \left(\rho^{(x)} \cdot N_{Sl \leftarrow MS}^{(x,j)} \right), \quad (30)$$

Please note that there is no dependence on the activity factor (some resources will be wasted if $\nu^{(x)} < 1$).

Combining (30) and (29) the maximum number $N_{U_{MAX}}^{(x,UGS,j)}$ of VoIP users supported by the UGS scheduling service can be easily derived.

Performance on ertPS. In the case of ertPS scheduling service, besides the adopted codec x and transmission mode j , also the activity factor $\nu^{(x)}$ must be considered in order to derive the amount of supported VoIP users. After the hangover period, in fact, the transmission rate is reduced and a single slot is allocated ($N_{Sl \leftarrow MH} = 1$) in order to allow the transmission of a stand-alone MAC signaling header for a quick modification of the resources request (see figure 8). Please note, by the way, that, although reduced, this allocation entails a resource wasting, since no transmission is performed in the most of cases. Resource wasting will occur also at any rate reduction (refer to figure 8): before a rate decreases, in fact, the request is sent in the uplink using an oversized resource.

Let us observe, furthermore, that following a rate increase request, the first (larger) resource is allocated in the subsequent frame, without respecting the normal rate of allocation, in order to reduce the latency.

In the case of ertPS scheduling service, therefore, in order to calculate $M_{Sl \leftarrow U}^{(x,ertPS,j)}$ we must take into account:

- the slots needed for full packets, that are generated with rate $\rho^{(x)}$ during active voice periods: $R \cdot N_{Sl \leftarrow MS}^{(x,j)}$ average slots per second, where $R = \nu^{(x)} \cdot \rho^{(x)}$ indicates the average number of full packets per second;
- the slots needed for SID packets, that are generated with rate $\rho_{SID}^{(x)}$ during silent periods: $R_{SID} \cdot N_{Sl \leftarrow MS_{SID}}^{(x,j)}$ average slots per second, where $R_{SID} = (1 - \nu^{(x)}) \cdot \rho_{SID}^{(x)}$ indicates the average number of SID packets per second;
- the slots needed for stand-alone MAC headers, allocated with rate $\rho^{(x)}$ when neither full packets nor SID packets are generated: $R_{MH} \cdot N_{Sl \leftarrow MH}$ average slots per second, where $R_{MH} = 1 - R - R_{SID}$; indicates the average number of slots left for MAC headers per second, transmitted at the same rate;
- the slots that are wasted during variations from full packets allocation to stand-alone MAC headers allocations: $R_{CFG} \cdot N_{Sl \leftarrow MS}^{(x,j)}$ average slots per second, where $R_{CFG} = \frac{1}{T_{CFG}^{av} + T_{CSS}^{av}}$ indicates the average number of uninterrupted periods of full packets per second, that depends on the average duration of a period of continuous full packets generation by the codec (T_{CFG}^{av}) and on the average duration of a period of silence suppression with SID packets generation (T_{CSS}^{av});
- the slots that are wasted during variations from SID packets to stand-alone MAC headers: $R_{SID} \cdot N_{Sl \leftarrow MS_{SID}}^{(x,j)}$ average slots per second.

Thus, in this case, $M_{Sl \leftarrow U}^{(x,ertPS,j)}$ is given by:

$$M_{Sl \leftarrow U}^{(x,ertPS,j)} = T_F \left((R + R_{CFG}) \cdot N_{Sl \leftarrow MS}^{(x,j)} + 2 \cdot R_{SID} \cdot N_{Sl \leftarrow MS_{SID}}^{(x,j)} + R_{MH} \cdot N_{Sl \leftarrow MH} \right), \quad (31)$$

Concerning the parameters T_{CFG}^{av} and T_{CSS}^{av} , please note that they are not only related to the adopted codec, but also to the characteristics of the specific conversation, also including, as an example, the language; in order to derive meaningful results, hereafter we adopted the values reported in (Stern et al., 1996) ($T_{CFG}^{av} = 0.17s$ and $T_{CSS}^{av} = 0.3428s$), although they are related to the voice-silence intervals rather than to codec full generation-codec silence suppression (they do not consider the hangover periods); for this reason, a slight underestimation of the maximum number of VoIP users is expected in the numerical results.

Finally, combining (31) and (29), the maximum number $N_{U,MAX}^{(x,ertPS,j)}$ of VoIP users supported by the ertPS scheduling service can be easily derived.

6.4 Numerical results

For the numerical results derivation, the amount of OFDMA-slots available for data transmission in the downlink subframe has been assumed equal to $N_{S_{dl}} = 350$; this value is a consequence of the same assumption reported in section 5.1: $BW = 7$ MHz, $N_{FFT}=2048$ OFDM subcarriers, normalized OFDM guard interval $G = 1/32$, frame duration $T_F = 10$ ms (which is the most suited value for VoIP traffic allocation); 15 of the 31 OFDM symbols of each frame

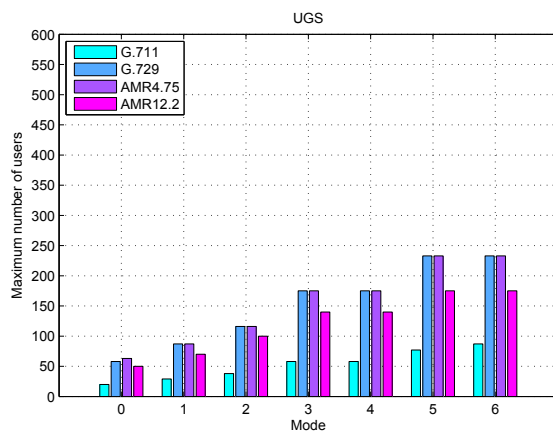


Fig. 9. Maximum number of users using UGS scheduling service. All modes. All VoIP codecs without silence suppression.

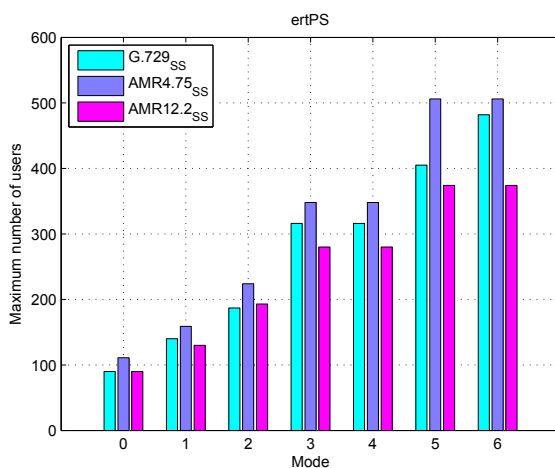


Fig. 10. Maximum number of users using ertPS scheduling services. All modes. All VoIP codecs with silence suppression.

are left for uplink data, adopting UL-PUSC (the rest of the symbols are used for the uplink common channels and the downlink subframe).

In the last column of table 2 the maximum number of VoIP users that can be served with *UGS* and *ertPS* (separated by the symbol "/") are reported for each codec; modes 0 and 6 are considered. Two underscores are typed when the adoption of that scheduling service makes no sense with that codec (i.e., *ertPS* with no silence suppression).

The maximum amount of VoIP users that can be supported is shown for all modes in figures 9 and 10 focusing on *UGS* and *ertPS*, respectively. Obviously, the VoIP capacity adopting *UGS*

with and without silence suppression is always the same; for this reason the results related to the case of silence suppression are not reported in figure 9.

It can be observed that, as expected, VoIP capacity is strictly related to the average data rate generated by the codec.

Comparing figures 9 and 10 we can also appreciate the significant benefit provided by the adoption of the *ertPS* instead of *UGS*.

7. Conclusions

The performance of IEEE802.16e WirelessMAN-OFDMA depends on a large number of system parameters and implementation choices, such as, among the others, the available bandwidth, the frame duration, the uplink/downlink traffic asymmetry and the fragmentation policy.

In the previous sections we provided an analytical framework that allows to evaluate the throughput achievable with TCP/IP connections as well as the amount of supported VoIP users as a function of the most significant parameters characterizing this technology. Beside the analytical model derivation, here we provided criteria, equations and algorithms to make the best choices from the viewpoint of system efficiency.

Furthermore, some numerical results were given, showing the impact of some specific parameters on the system performance. With reference to TCP/IP connections, for instance, the troublesome choice of the maximum size of ARQ blocks has been discussed as well as the potential resource wastage entailed by a wrong choice of the asymmetry factor between the downlink and uplink subframes. The main outcome of this analysis is given by a set of criteria to be followed in order to maximize the throughput provided to the final user.

As far as VoIP connections are concerned, here we assessed the maximum number of users that can be supported, carefully considering the voice codec characteristics and the adopted scheduling service. The outcomes of this investigation provide an indication about the capacity of this technology to be alternative to other technologies such as UMTS and LTE for the provision of the voice service.

As a final remark, let us observe that the analytical framework proposed in this chapter provides a tool to evaluate the upper limits of the throughput and the maximum amount of VoIP users that can be supported by IEEE802.16e WirelessMAN-OFDMA. However, in order to investigate the actual performance of such a complex technology in a given scenario considering the degradation due, for instance, to fading, shadowing, noise and interference, the only feasible way is to adopt a simulation tool able to carefully reproduce all aspects of communications, with particular reference to the physical layer behavior.

This kind of investigation has been carried out at WiLab (Italy) by means of the simulation platform SHINE, that has been developed in the last years to assess the performance of wireless networks in realistic scenarios. The interested reader may refer to (Andrisano et al., 2007; 2009; Bazzi et al., 2006).

Appendix I

Here we illustrate the algorithm that allows to maximize symbols usage starting from the number of useful symbols $N_{USy} = N_{Sy} - N_{OSy_{dl}} - N_{OSy_{ul}}$. The following steps must be followed:

1. derive $J = Res_{ul} = \text{mod}(N_{USy}, N_{GS_{ul}})$;
2. find, if possible:

$$a = \min_{a \in \left[0, \left\lfloor \frac{N_{USy}}{N_{GS_{ul}}} \right\rfloor\right]} \text{ so that } \text{mod}(J + a \cdot N_{GS_{ul}}, N_{GS_{dl}}) = 0. \quad (32)$$

3. if a was not found, then reduce J by one and return to step 2, else exit.

The obtained value a and the parameter J allow the derivation of the minimum value for $N_{Sy_{dl}}$ ($N_{Sy_{dl}}^{MIN}$) that maximally reduces the symbols wasting:

$$N_{Sy_{dl}}^{MIN} = J + a \cdot N_{GS_{ul}}, \quad (33)$$

All possible solutions will be:

$$N_{Sy_{dl}}^{OPT}(b) = N_{Sy_{dl}}^{MIN} + b \cdot \text{mcm}(N_{GS_{ul}}, N_{GS_{dl}}), \quad (34)$$

where $b \in \left[0, \left\lfloor \frac{N_{USy} - N_{Sy_{dl}}^{MIN}}{N_{GS_{dl}}} \right\rfloor\right]$ and $\text{mcm}(x, y)$ is the minimum common multiplier of x and y .

It follows:

$$N_{Sy_{ul}}^{OPT}(b) = N_{USy} - (Res_{ul} - J) - N_{Sy_{dl}}^{OPT}(b). \quad (35)$$

and

$$AF_{Sl}(b) = \frac{N_{Sl_{dl}}^{OPT}(b)}{N_{Sl_{ul}}^{OPT}(b)}. \quad (36)$$

Finally, we can choose the value of b that brings to the $AF_{Sl}(b)$ nearer to AF_{in} .

8. References

- 3GPP TS 26.071 (2008). Mandatory speech codec speech processing functions; amr speech codec; general description.
- 3GPP TS 26.092 (2008). Mandatory speech codec speech processing functions; adaptive multi-rate (amr) speech codec; comfort noise aspects.
- Ahmadi, S. (2009). An overview of next-generation mobile wimax technology, *IEEE Communications Magazine* **Vol.47**(n.6): pp.84–98.
- Andrisano, O., Bazzi, A., Leonardi, G. & Pasolini, G. (2007). Ieee802.16e best effort performance investigation, *Proceedings of IEEE International Conference on Communications, 2007 (ICC 2007)*, IEEE, Glasgow, Scotland, pp. 4837–4842.
- Andrisano, O., Bazzi, A., Leonardi, G. & Pasolini, G. (2009). Ieee802.16e simulation issues, *Proceedings of IEEE Mobile WiMAX Symposium 2009 (MWS 2009)*, IEEE, Napa Valley, California, pp. –.
- Bazzi, A., Gambetti, C. & Pasolini, G. (2006). Shine: Simulation platform for heterogeneous interworking networks, *Proceedings of IEEE International Conference on Communications, 2006 (ICC 2006)*, IEEE, Istanbul, Turkey, pp. 5534–5539.
- Benyassine, A., Shlomot, E. & Su, H. (1997). Itu-t recommendation g.729 annex b: a silence compression scheme for use with g.729 optimized for v.70 digital simultaneous voice and data applications., *IEEE Communications Magazine* **Vol.35**(Issue 9): pp.64–73.
- Cimini, J. (1985). Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing., *IEEE Trans. Comm.* **Vol.COM-33**(n.7): pp.665–675.
- Estepa, R., Vozmediano, J. & Estepa, A. (2005). Accurate prediction of voip traffic mean bit rate., *ELECTRONICS LETTERS* **Vol.41**(n.17): pp.985–987.
- Goode, B. (2002). Voice over internet protocol (voip), *Proceedings of the IEEE* **Vol.90**(n.9): pp.1495–1517.
- IEEE Std 802.16-2004 (2004). Ieee standard for local and metropolitan area networks part 16: Air interface for fixed broadband wireless access systems.
- IEEE Std 802.16e-2005 (2006). Ieee std 802.16e-2005 and ieee std 802.16-2004/cor1-2005 ieee standard for local and metropolitan area networks part 16: Air interface for fixed and mobile broadband wireless access systems amendment 2: Physical and medium access control layers for combined fixed and mobile operation in licensed bands and corrigendum 1.
- ITU-T Rec. G.711 (1988). Pulse code modulation (pcm) of voice frequencies.
- ITU-T Rec. G.729 (1996a). Coding of speech at 8kbit/s using conjugate-structure algebraic-code-excited linear-predictive (cs-acelp) coding.
- ITU-T Rec. G.729, A. (1996b). A silence compression scheme for g.729 optimized for terminals conforming to itu-t v.70.
- Lee, H., Kwon, T., Cho, D., Lim, G. & Chang, Y. (2006). Performance analysis of scheduling algorithms for voip services in ieee 802.16e systems, *Proceedings of IEEE Vehicular Technology Conference (VTC 2006-Spring)*., IEEE, Melbourne, Australia, pp. 1231–1235.
- Li, Q., Lin, X., Zhang, J. & Roh, W. (2009). Advancement of mimo technology in wimax: from ieee 802.16d/e/j to 802.16m, *IEEE Communications Magazine* **Vol.47**(n.6): pp.100–107.
- Stern, H., Mahmoud, S. & Wong, K. (1996). A comprehensive model for voice activity in conversational speech-development and application to performance analysis of new-generation wireless communication systems, *Wireless Networks* **Vol.2**(n.4): pp.359–367.
- Van Nee, R. & Prasad, R. (2000). *OFDM for Wireless Multimedia Communications*, Artech House.

Throughput-Enhanced Communication Approach for Subscriber Stations in IEEE 802.16 Point-to-Multipoint Networks

Chung-Hsien Hsu and Kai-Ten Feng

*Department of Electrical Engineering, National Chiao Tung University
Taiwan, R.O.C.*

1. Introduction

The IEEE 802.16 standard for wireless metropolitan area networks (WMANs) is designed to satisfy various demands for high capacity, high data rate, and advanced multimedia services (Abichar et al., 2006). The medium access control (MAC) layer of IEEE 802.16 networks supports both point-to-multipoint (PMP) and mesh modes for packet transmission (IEEE Std. 802.16-2004, 2004). Based on the application requirements, it is suggested in the standard that only one of the modes can be exploited by the network components within the considered time intervals, and the PMP mode is considered the well-adopted one. In the PMP mode, packet transmission is coordinated by a base station (BS) which is responsible for controlling the communication with multiple subscriber stations (SSs) in both downlink (DL) and up-link (UL) directions. All the traffic within an IEEE 802.16 PMP network can be categorized into two types, including inter-cell traffic and intra-cell traffic. For the inter-cell traffic, the source-destination pair of each traffic flow are located in different cells. On the other hand, the intra-cell traffic is defined if they are situated within the same cell. The inefficiency within the PMP mode occurs while two SSs are intended to conduct packet transmission, i.e., the intra-cell traffic between the SSs. It is required for the data packets between the SSs to be forwarded by the BS even though the SSs are adjacent with each others. Due to the packet rerouting process, the communication bandwidth is wasted which consequently increases the packet-rerouting delay.

In order to alleviate the drawbacks resulted from the indirect transmission, a directly communicable mechanism between SSs should be considered in IEEE 802.16 networks. Several direct communication approaches have been proposed for different types of networks. The direct-link setup (DLS) protocol is standardized in the IEEE 802.11z draft standard to support direct communication between two SSs in wireless local networks (IEEE P802.11zTM/D5.0, 2009). However, the DLS protocol is designed as a contention-based mechanism, which does not guarantee the access of direct link setup and data exchanges between two SSs. The dynamic slot assignment (DSA) scheme for Bluetooth networks is proposed in (Zhang et al., 2002) and (Cordeiro et al., 2003), which is primarily implemented based on the characteristics of the Bluetooth standard. Since frame structures and medium access mechanisms are different among these wireless communication technologies, both the DLS protocol and DSA scheme cannot be directly applied to IEEE 802.16 networks.

In this book chapter, a point-to-point direct communication (PDC) approach is proposed for achieving direct transmission between two SSs. The PDC approach is designed as a flexible and contention-free scheme especially for time division duplexing based IEEE 802.16 PMP networks. The BS is coordinating and arranging specific time intervals for the two SSs that are actively involved in packet transmission. Both the relative locations and channel conditions among the BS and SSs are utilized as constraints for determining if the direct communication should be adopted. The advantage of exploiting the PDC approach is that both the required bandwidth for packet transmission and packet-rerouting delay for intra-cell traffic can be significantly reduced. The effectiveness of the proposed PDC approach can be observed via the simulation results, which demonstrate that the PDC approach outperforms the conventional IEEE 802.16 transmission mechanism in terms of user throughput.

The remainder of this book chapter is organized as follows. Section 2 briefly reviews the MAC frame structure and packet transmission mechanism in IEEE 802.16 PMP networks. The proposed PDC approach, consisting of management structures, an admission control scheme, and direct communication procedures, is described in Section 3. The performance of the PDC approach is evaluated in Section 4. Section 5 draws the conclusions.

2. IEEE 802.16 PMP Networks

The PMP mode is considered the well-adopted network configuration in IEEE 802.16 networks wherein the BS is responsible for controlling all the communication among SSs. Two duplexing techniques are supported for the SSs to share common channels, i.e., time division duplexing (TDD) and frequency division duplexing. The MAC protocol is structured to support multiple physical (PHY) layer specifications in the IEEE 802.16 standard. In this book chapter, the WirelessMAN-OFDM PHY, utilizing the orthogonal frequency division multiplexing (OFDM), with TDD mode is exploited for the design of the proposed PDC approach. Both the frame structure and packet transmission mechanism of IEEE 802.16 PMP networks are described in the following subsections.

2.1 Frame Structure

Fig. 1 illustrates the schematic diagram of the IEEE 802.16 PMP OFDM frame structure with TDD mode. It can be observed that each frame consists of a DL subframe and a UL subframe. The DL subframe contains only one DL PHY protocol data unit (PDU), which starts with a long preamble for PHY synchronization. The preamble is followed by a frame control header (FCH) burst and several DL bursts. A DL frame prefix (DLFP), which is contained in the FCH, specifies the burst profile and length for the first DL burst (at most four) via the information element (IE). It is noted that each DL burst may contain an optional preamble and more than one MAC PDUs that are destined for the same or different SSs. The first MAC PDU followed by the FCH is the DL-MAP message, which employs DL-MAP IEs to describe the remaining DL bursts. The DL-MAP message can be excluded in the case that the DL subframe consists of less than five bursts; nevertheless, it must still be sent out periodically to maintain synchronization. A UL-MAP message immediately following the DL-MAP message denotes the usage of UL bursts via UL-MAP IEs. An interval usage code, corresponding to a burst profile, describes a set of transmission parameters, e.g., the modulation and coding type, and the forward error correction type. The DL interval usage code (DIUC) and UL interval usage code (UIUC) are specified in the DL channel descriptor (DCD) and UL channel descriptor (UCD) messages respectively. The BS broadcasts both the DCD and UCD messages periodically to define the characteristics of the DL and UL physical channels respectively.

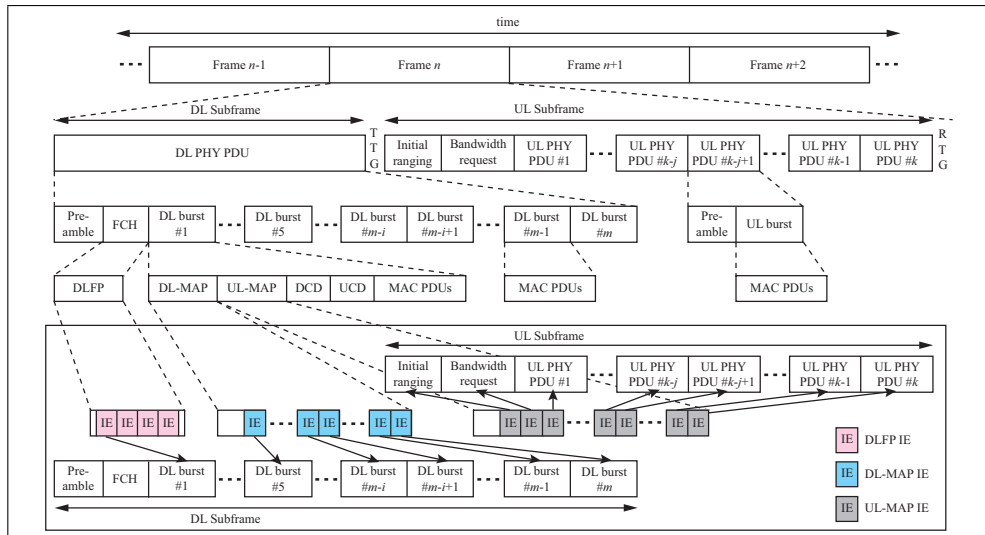


Fig. 1. Schematic diagram of IEEE 802.16 PMP OFDM frame structure with TDD mode.

On the other hand, as can be seen from Fig. 1, the UL subframe starts with the contention intervals that are specified for both initial ranging and bandwidth request. It is noted that more than one UL PHY PDU can be transmitted after the contention intervals. Each UL PHY PDU consists of a short preamble and a UL burst, where the UL burst transports the MAC PDUs for each specific SS. Moreover, a transmit-to-receive transition gap (TTG) and a receive-to-transmit transition gap (RTG) are inserted in between the DL and the UL subframes and at the end of each frame respectively. These two gaps provide the required time for the BS to switch from the transmit to receive mode and vice versa.

2.2 Packet Transmission Mechanism

A connection in IEEE 802.16 PMP networks is defined as a unidirectional mapping between the BS and an MS, which is identified by a 16-bit connection identifier (CID). Two kinds of connections, including management connections and transport connections, are defined in the IEEE 802.16 standard. The management connections are utilized for delivering MAC management messages; while the transport connections are employed to transmit user data. During the initial ranging of a SS, a pair of UL and DL basic connections are established, which belong to a type of the management connections. It is noted that a single Basic CID is assigned to a pair of UL and DL basic connections, which is served as the identification number for the corresponding SS. Thus the SS uses the individual transport CID to request bandwidth for each transport connection while the BS arranges the accumulated transmission opportunity by addressing the Basic CID of the SS.

An exemplified network topology that consists of one BS and two neighboring SSs is shown in Fig. 2. Two types of traffic exist in the network: inter-cell traffic and intra-cell traffic. For the inter-cell traffic, the source and the destination for each traffic flow are located in different cells, e.g., the traffic flow of SS_2 for accessing the Internet. On the other hand, the intra-cell traffic is defined while the source and destination are situated within the same cell network,

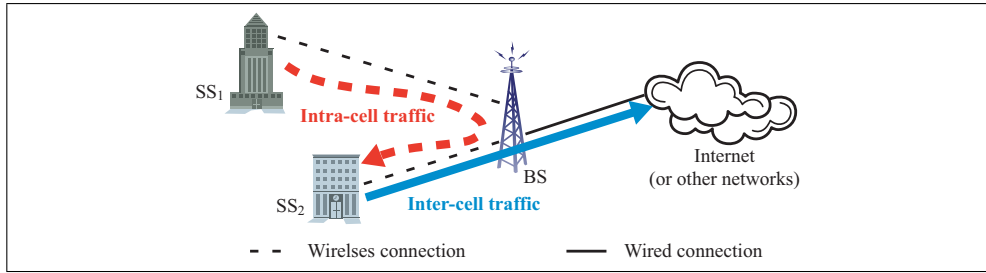


Fig. 2. Example of IEEE 802.16 PMP network topology.

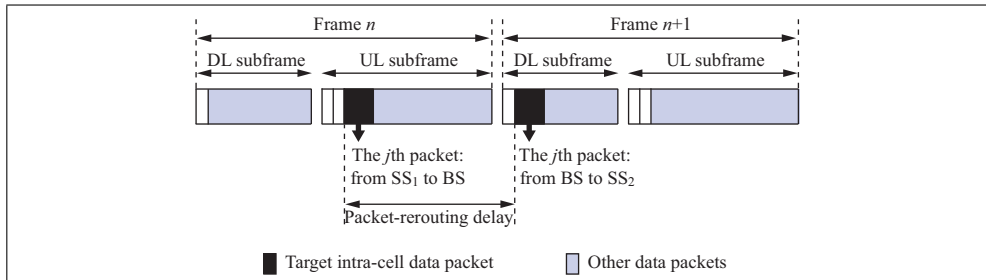


Fig. 3. Schematic diagram of IEEE 802.16 packet transmission mechanism in time sequence.

such as the traffic flow between SS_1 and SS_2 in Fig. 2. Considering the scenario that SS_1 intends to communicate with its neighboring station SS_2 , two transport connections are required to be established via the service flow management mechanism for the intra-cell traffic, i.e., the UL transport connection from SS_1 to the BS and the DL transport connection from the BS to SS_2 . Fig. 3 illustrates the conventional transmission mechanism of IEEE 802.16 PMP networks in time sequence. In the most ideal case, the j th intra-cell packet, transmitted from SS_1 to the BS in the n th frame, will be forwarded to SS_2 in the $(n+1)$ th frame by the BS. The rerouting process apparently requires twice of communication bandwidth for achieving the intra-cell packet transmission, which consequently increases control overhead by duplicating the corresponding data packet. Moreover, the delay time for packet-rerouting can be more than one half of a frame duration while the packet transmission from the BS to SS_2 is postponed to a latter DL subframe.

3. Point-to-point Direct Communication (PDC) Approach

The objective of the proposed PDC approach is to provide a directly communicable mechanism for SSs within IEEE 802.16 PMP networks such that both the communication bandwidth and packet-rerouting delay of intra-cell traffic are reduced. The PDC approach is designed as a flexible and contention-free scheme wherein the establishment of direct link is conducted along with packet transmission. Based on the channel conditions among the BS and SSs, the BS coordinates and arranges specific time intervals for the two SSs that are actively involved in packet transmission. It is worthwhile to mention that the PDC approach is carried out after the establishment of the original transmission path, which is compatible and can be directly integrated with the existing protocols defined in the IEEE 802.16 standard. In the following

subsections, the proposed architecture and management structures will be described in Subsection 3.1; while an admission control scheme for direct link establishment is explained in Subsection 3.2. The direct communication procedures of the PDC approach are given in Subsection 3.3.

3.1 Architecture and Management Structure

For the purpose of providing time intervals for direct transmission between SSs, a point-to-point direct link (PDL) subframe is proposed in the PDC approach. A PDL subframe that consists of one or more PDL PHY PDUs is designed as a subset of a DL or UL subframe. Each PDL PHY PDU starts with a short preamble followed by a PDL burst, which is designed to transport the MAC PDUs for each specific SS. Furthermore, in order to be compatible with the existing IEEE 802.16 standard, three categories of management structures are proposed, which are detailed as follows:

- **DL-PDL IE and UL-PDL IE.** The proposed DL-PDL IE and UL-PDL IE are designed to depict burst profiles and lengths of their corresponding PDLs in the DL and UL subframes respectively. The DL-PDL IE is a new type of the extended DIUC dependent IE within the OFDM DL-MAP IE; while the UL-PDL IE is a new type of the UL extended IE that is contained in the OFDM UL-MAP IE. It is noted that the formats of both the proposed DL-PDL IE and UL-PDL IE are designed to conform to the formats of the DL-MAP dummy IE and UL-MAP dummy IE, specified in the IEEE 802.16 standard, respectively.
- **PDL Subheader.** The PDL subheader is designed for implementing the request, response, announcement, and termination of the direct communication. It is a new type of per-PDU subheader, which can be inserted in the MAC PDUs immediately followed by the generic MAC header in both the DL and UL directions. For different purposes, the DL subheader carries various types of information, including MAC addresses, CIDs, and location information.
- **PBPC-REQ and PBPC-REP Messages.** In the IEEE 802.16 standard, the adaptive modulation and coding (AMC) is exploited as the link adaption technique to improve the network performance on time-varying channels. The BS selects an adequate modulation and coding scheme (MCS) for a SS based on the reported signal-to-interference and noise ratio (SINR) value. Moreover, the BS permits the changes in MCS that are suggested by the SS via the burst profile change request message. Similarly, both the proposed PDL burst profile change request (PBPC-REQ) and response (PBPC-REP) messages are designed to change the MCS applied in a direct link. The PBPC-REQ message is utilized to request the adjustment of assigned MCS for PDL burst. The BS will respond with the PBPC-REP message for either confirming or denying the alternation in the suggested MCS.

3.2 Admission Control Procedure

In the PDC approach, some criteria should be exploited to determine the execution of direct communication between two SSs. A two-tiered admission control scheme for a BS and two attached SSs is presented in this subsection. In wireless communication system, the data transmission range for each station is proportional to its corresponding transmission power. In order to avoid potential interference introduced by adopting the PDC approach, the distance

factor is considered as the first-tiered constraint (\mathcal{C}_1), which is defined as

$$\mathcal{C}_1 : D(SS_s, SS_d) \leq D(SS_s, BS),$$

where $D(x, y)$ denotes the relative distance between x and y ; while the source SS and destination SS of a intra-cell traffic is represented as SS_s and SS_d respectively. In other words, the transmission power utilized by SS_s for achieving direct transmission is adjusted to be equal to or less than that as specified in the conventional IEEE 802.16 mechanism.

On the other hand, for the purpose of enhancing the efficiency for data transmission, channel conditions among the BS and (SS_s, SS_d) pair should be taken into account. Different MCSs associated with various number of data bits are adopted for data transmission under different channel conditions. Based on channel states and the corresponding MCSs, the second-tiered constraint (\mathcal{C}_2) is defined as

$$\mathcal{C}_2 : T_{PDC}(SS_s, SS_d) \geq T_{Conv}(SS_s, SS_d),$$

where $T(SS_s, SS_d)$ represents the raw user throughput defined as "number of bits per second that is received by the destination SS_d while the source is SS_s ". In other words, the raw user throughput resulted from the PDC approach (T_{PDC}) should be at least equal to or higher than that from the conventional IEEE 802.16 mechanism (T_{Conv}). The values of both T_{PDC} and T_{Conv} are derived as the description in the following paragraph.

MCS index	Modulation	Coding rate	Coded block size (byte)	Receiver SNR (dB)
0	BPSK	1/2	24	3.0
1	QPSK	1/2	48	6.0
2	QPSK	3/4	48	8.5
3	16-QAM	1/2	96	11.5
4	16-QAM	3/4	96	15.0
5	64-QAM	2/3	144	19.0
6	64-QAM	3/4	144	21.0

Table 1. OFDM Modulation and Coding Schemes

Table 1 shows the supported MCSs that are specified within the IEEE 802.16 standard. In the considered OFDM system, the raw data rate R_d of a MCS with index ξ is represented as

$$R_d[\xi] = \frac{B_u[\xi]}{T_s}, \quad (1)$$

where T_s is the OFDM symbol duration. The notation $B_u[\xi]$ indicates the number of uncoded bits per OFDM symbol of a MCS with index ξ , which is obtained as

$$B_u[\xi] = N_d \cdot \log_2 M \cdot R_c[\xi], \quad (2)$$

where N_d denotes the number of data subcarriers and $R_c[\xi]$ is the coding rate of a MCS with index ξ . The value of the parameter M depends on the adopted MCS, i.e., $M = 2$ for BPSK, $M = 4$ for QPSK, $M = 16$ for 16-QAM, and $M = 64$ for 64-QAM. Moreover, the OFDM symbol duration T_s can be acquired as

$$T_s = T_b + T_g = T_b + G \cdot T_b = \frac{1 + G}{\Delta f}, \quad (3)$$

where T_b and T_g represent the useful symbol time and the cyclic prefix (CP) time respectively. The notation G denotes the ratio of T_g to T_b . The subcarrier spacing Δf is obtained as

$$\Delta f = \frac{F_s}{N_s} = \frac{8000}{N_s} \cdot \left\lfloor \frac{n \cdot BW}{8000} \right\rfloor, \quad (4)$$

where N_s indicates the number of total subcarriers. The notation F_s represents the sampling frequency with its value specified by the IEEE 802.16 standard as in (4), where n is the sampling factor and BW is the channel bandwidth. By substituting (4) into (3), the OFDM symbol time can be approximated as

$$T_s = \frac{N_s}{F_s} \cdot (1 + G) \approx \frac{N_s}{n \cdot BW} \cdot (1 + G). \quad (5)$$

With (2) and (5), the raw data rate R_d of a MCS with index ξ in (1) becomes

$$R_d[\xi] \approx \frac{N_d \cdot \log_2 M \cdot R_c[\xi] \cdot n \cdot BW}{N_s \cdot (1 + G)}. \quad (6)$$

Based on (6), the raw user throughput by adopting the PDC approach is acquired as

$$T_{PDC}(SS_s, SS_d) = R_d[\xi_{(s,d)}], \quad (7)$$

where $\xi_{(s,d)}$ represents the index of a MCS that will be assigned to the direct link of the (SS_s, SS_d) pair. On the other hand, the raw user throughput in the conventional IEEE 802.16 mechanism is constrained by the two-hop transmission, i.e., from SS_s to BS and from BS to SS_d . Thus the T_{Conv} can be obtained as

$$T_{Conv}(SS_s, SS_d) = \frac{1}{2} R_d[\phi_{(s,d)}], \quad (8)$$

where

$$\phi_{(s,d)} = \min [\xi_{(s,BS)}, \xi_{(BS,d)}]. \quad (9)$$

The notation $\xi_{(s,BS)}$ denotes the index of the MSC utilized in the link between the SS_s and BS; while that is assigned to the link between the BS and SS_d is represented as $\xi_{(BS,d)}$.

3.3 Direct Communication Procedures

Based on the aforementioned management structures and admission control scheme, the direct communication procedures of the PDC approach are explained in this subsection. Considering a basic IEEE 802.16 PMP network that consists of a BS and two SSs, an intra-cell traffic flow is existed between the SSs. Two transport connections are established for packet transmission, i.e., a UL transport connection from the source station SS_s to the BS and a DL transport connection from the BS to the destination station SS_d . The initialization of direct communication is achieved by conducting the link request and information collection. The source-destination pair (SS_s, SS_d) anticipating to establish the direct link are required to provide their location information and channel conditions to the BS. The collected information is utilized in the admission control scheme mentioned above.

Fig. 4 illustrates an exemplified message flows of the SS-initiated procedure for direct communication. In the case that SS_s intends to conduct direct communication with SS_d , it attaches a PDL subheader to a data packet that will be delivered to the BS; meanwhile, the location

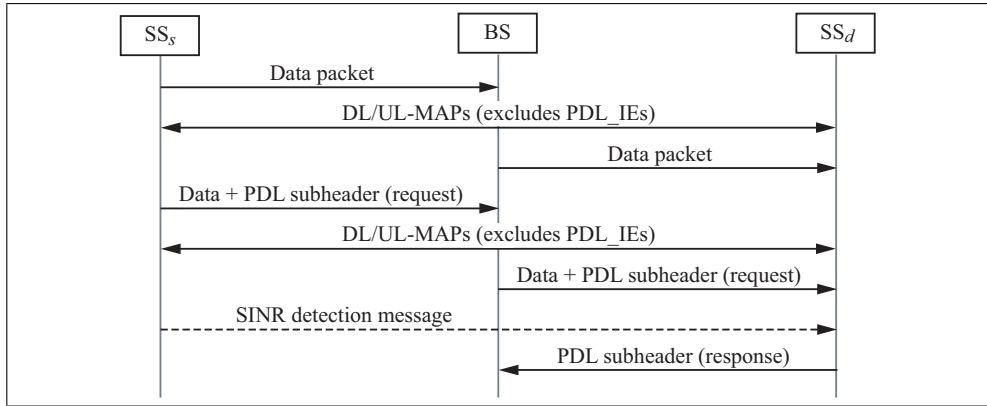


Fig. 4. Schematic diagram of SS-initiated procedure for direct communication.

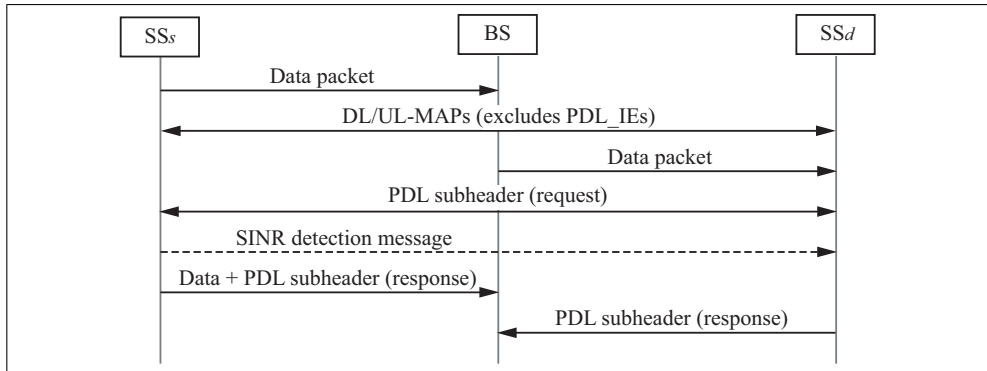


Fig. 5. Schematic diagram of BS-initiated procedure for direct communication.

information of SS_s will be filled into the PDL subheader. As the BS receives the request PDL subheader from the SS_s , the BS will attach a PDL subheader to the data packet and conduct the transmission to SS_d . Moreover, the BS will arrange a DL burst for SS_s with the assignment in the corresponding DL-MAP message. SS_s will transmit an SINR detection message to SS_d with the BPSK-1/2 MCS for estimating the channel state of the direct link. After receiving the PDL subheader and the SINR detection message from the BS and SS_s respectively, SS_d will transmit a response PDL subheader associated with the calculated SINR value. It is noted that the location information of SS_d is carried in the response PDL subheader if it is required by the BS. On the other hand, the BS-initiated direct communication procedure is shown in Fig. 5. Contrary to the SS-initiated procedure, the BS actively announces the link request along with the PDL subheader to the specific SS_s , i.e., SS_s and SS_d . As SS_s receives the requesting PDL subheader from the BS, it will utilize the response PDL subheader to provide the location information that is requested by the BS. The remaining steps of the BS-initiated procedure are similar to that of the SS-initiated case, such as the SINR detection and SS_d response. The BS executes the admission control procedure after it received the response PDL subheader transmitted from SS_d . Based on the collected information, the aforementioned two-tiered con-

trol scheme is exploited by the BS to either confirm or deny the direct communication request between SS_s and SS_d . If the request is rejected, the BS will broadcast a denying announcement along with the PDL subheader. On the other hand, a confirming announcement will be transmitted if the request is granted. Consequently, the BS will arrange the PDL bursts for the direct link in the subsequent frames.

After receiving the confirmation announcement, the considered SSs will activate the procedure of direct communication. According to the received MAPs associated with the PDL IEs, SS_s will conduct packet transmission directly to SS_d within the PDL bursts. Moreover, SS_d will continuously observe and evaluate the channel condition for the direct link with the adaptation to an appropriate MCS. The calculated SINR is compared with the receiver SNR range of the current MCS (as listed in Table 1) by SS_d . If the existing MCS is observed to be improper for the current channel condition, SS_d will initiate a PBPC-REQ message to the BS for suggesting an appropriate MCS. Consequently, the BS will respond a PBPC-REP message with a recommended MCS.

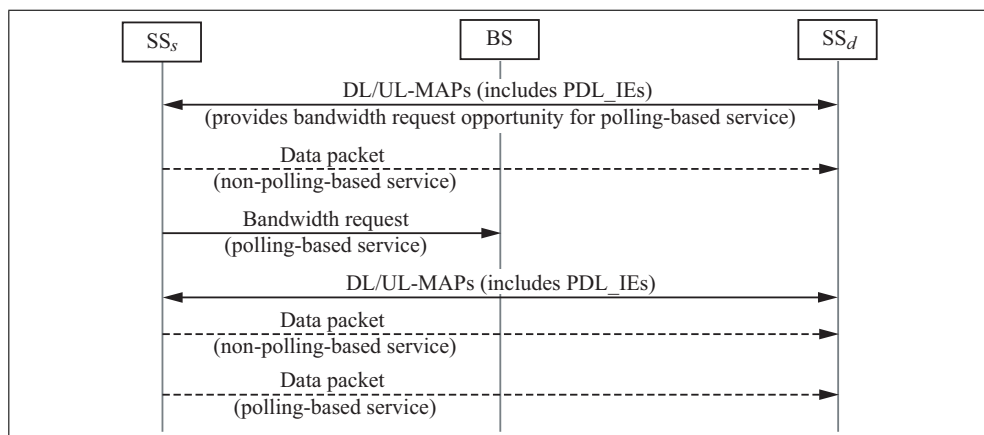


Fig. 6. Schematic diagram of bandwidth request procedure in PDC approach.

It is worthwhile to mention that bandwidth requests are conducted by an SS based on individual transport connection; while bandwidth grants from the BS is executed according to the accumulated requests from the SS. In other words, the bandwidth grant is addressed to the Basic CID of the corresponding SS, not to the individual transport CIDs. As a result, the CID specified for the PDL burst becomes the Basic CID of SS_s . Furthermore, in order to integrate with the existing specification, the procedures of bandwidth requests and allocations specified in the IEEE 802.16 standard are implemented within the proposed PDC approach. Fig. 6 illustrates the bandwidth request procedure while the PDC approach is adopted. It can be observed that the BS preserves the PDL burst for non-polling based service periodically. Furthermore, the BS will continue to provide unicast bandwidth request opportunity for the polling-based services based on the original transport CIDs of SS_s . The unicast bandwidth grant of those services will consequently be assigned to the PDL burst based on the Basic CID of SS_s .

The procedure for the link termination occurs as one of the following two conditions is satisfied: (i) the channel condition of the direct link is becoming worse than that from the indirect channels (i.e., via the BS); (ii) the direct communication is determined to be ceased. It is noted

that the link termination can be initiated by either the BS or SS. In the SS-initiated termination procedure, the SS will transmit a termination PDL subheader to the BS. As the message is received by the BS, it will broadcast an announcement along with a PDL subheader to both SS_s and SS_d regarding the termination of the direct link. On the other hand, for the BS-initiated termination procedure, the termination information is actively announced by the BS. As a result, the BS and the associated SSs will return to adopt the original packet transmission mechanism as defined in the IEEE 802.16 standard.

4. Performance Evaluation

The performance of the proposed PDC approach is evaluated and compared with the conventional packet transmission mechanism in IEEE 802.16 PMP networks via simulations. A single BS with 12 SSs uniformly distributed within the BS's coverage are considered as the simulation layout. The OFDM modulation and coding schemes listed in Table 1 are adopted in the simulation. The occurring frequencies for both inter-cell traffic and intra-cell traffic are considered uniformly distributed. The packet lengths are selected to follow the exponential distribution; while the Poisson distribution is adopted for packets arrival time. Since scheduling algorithm is not specified in the IEEE 802.16 standard, the direct round robin (DRR) (Shreedhar & Varghese, 1996) and weighted round robin (WRR) (Katevenis et al., 1991) algorithms are selected as the BS's DL and UL schedulers respectively. The DRR algorithm is also utilized by the SS to share the UL grants that are provided by the BS among their connections. The parameters adopted in the simulations are listed in Table 2.

Parameter	Value
Channel bandwidth (BW)	7 MHz
Number of total subcarriers (N_s)	256
Number of data subcarriers (N_d)	192
Sampling factor (n)	8/7
Sampling frequency (F_s)	8 MHz
Useful symbol time (T_b)	32 μ s
CP time (T_g)	2 μ s
The ratio of CP time and useful time (G)	1/16
OFDM symbol duration (T_s)	34 μ s
Maps modulation	BPSK
Data modulation	QPSK, 16-QAM, 64-QAM
Frame duration	5 ms, 10 ms
SSTG/SSRTG	35 μ s
Initial ranging interval	5 OFDM symbols
Bandwidth request interval	5 OFDM symbols
Average packet size	200 bytes
Simulation time	1 sec

Table 2. Simulation Parameters

Fig. 7 shows the comparison of the user throughput with an increasing number of intra-cell traffic flows ranging from 10 to 100 (frame duration = 5 and 10 ms). As can be expected that the user throughput increases as the number of intra-cell traffic flows is augmented. It can be observed that the proposed PDC approach outperforms the conventional IEEE 802.16 scheme

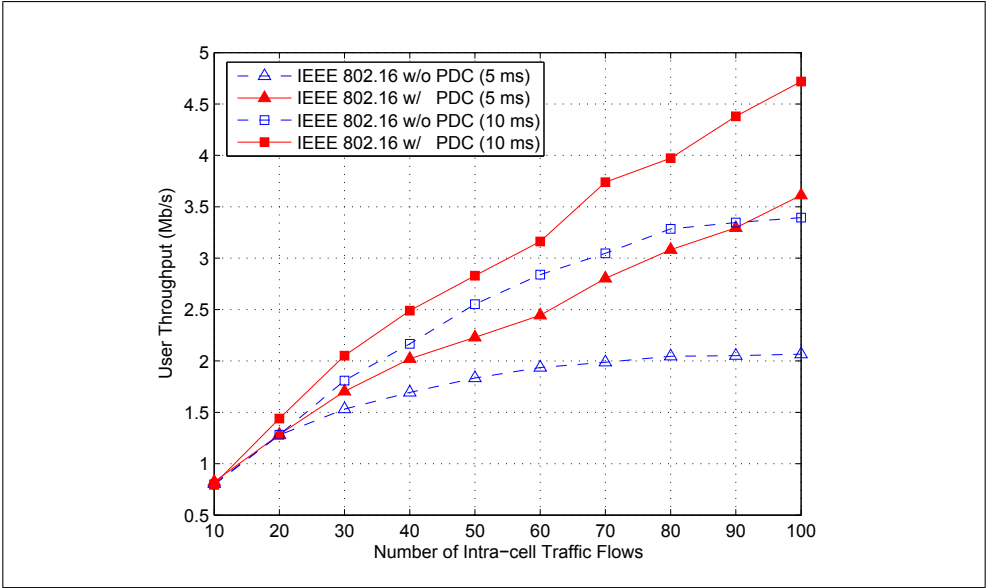


Fig. 7. Performance comparison: user throughput versus number of intra-cell traffic flows.

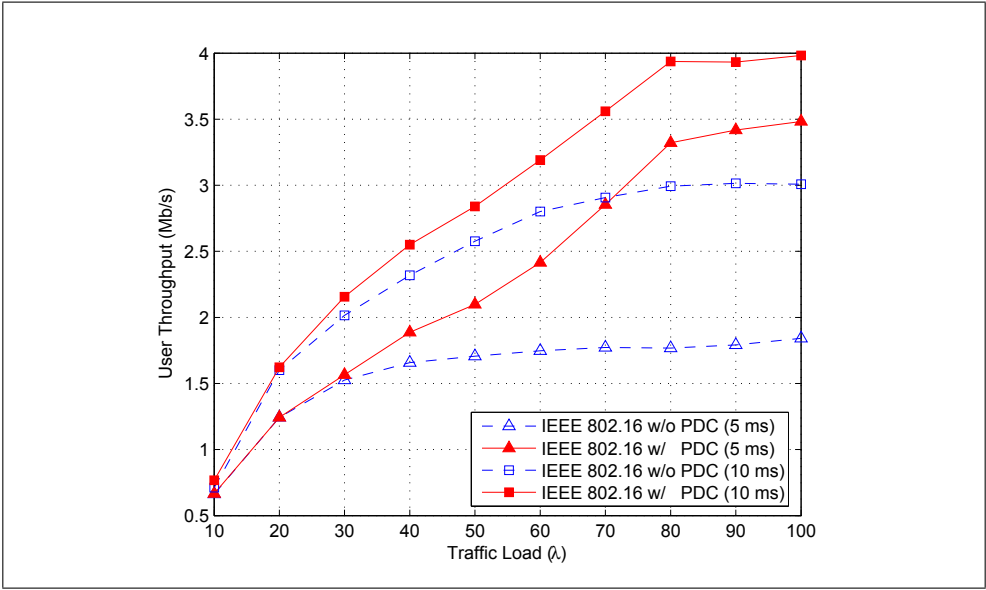


Fig. 8. Performance comparison: user throughput versus traffic load.

with higher user throughput under different frame durations. In the conventional mechanism, it is required for the intra-cell traffic to be forwarded by the BS. Consequently, more than twice

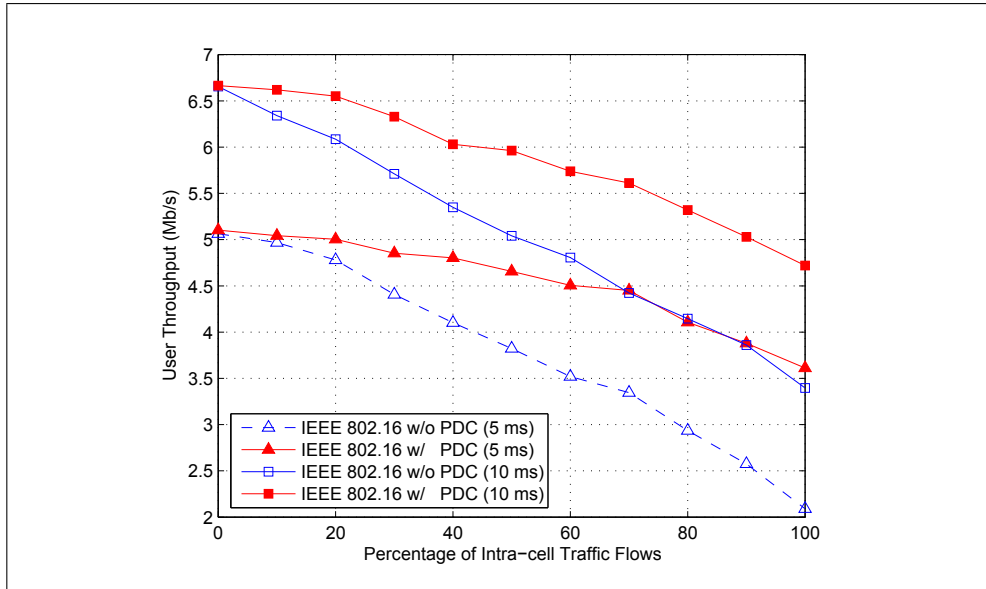


Fig. 9. Performance comparison: user throughput versus percentage of intra-cell traffic flows.

of the communication bandwidth is necessitate for the packet transmission. By adopting the proposed PDC approach, the intra-cell traffic can be directly transmitted from the source station to the destination station, which resulted in saved bandwidth. Moreover, the longer frame duration can achieve higher user throughput owing to the reason that less control overheads are required within the transmission. The comparison of the user throughput under different traffic load (λ) is illustrated in Fig. 8, wherein there are 50 intra-cell traffic flows. Similar performance benefits can be observed by adopting the proposed PDC approach.

In order to evaluate the influence from the inter-cell traffic, the user throughput with an increasing number of inter-cell traffic flows ranging from 10 to 100 is shown in Fig. 9 (with the number of total traffic flows is equal to 100). It is noticed that the inter-cell traffic can be considered as a particular type of direct communication within the cell since the packets are passed from the BS to SS directly. Consequently, the user throughput is decreased as the percentage of the intra-cell traffic is augmented since there are increasing amounts of indirect links within the network. Nevertheless, the PDC approach can still provide comparably higher user throughput under different percentages of intra-cell traffic flows. The merits of the proposed PDC scheme can be observed.

5. Conclusions

In this book chapter, a flexible and contention-free point-to-point direct communication (PDC) approach is proposed to achieve direct transmission between SSs within IEEE 802.16 PMP networks. With the considerations of both relative locations and channel conditions among the BS and SSs, a two-tiered admission control scheme is proposed to determine the establishment of direct link between the SSs in the PDC approach. While adapting the PDC approach, the

BS arranges specific time intervals for the two SSs that are actively involved in direct transmission. The advantage of exploiting the PDC approach is that both the required bandwidth for packet transmission and packet-rerouting delay for intra-cell traffic can be significantly reduced. Furthermore, the design of the PDC approach is compatible and can be directly integrated with the existing protocols defined in the IEEE 802.16 standard. The effectiveness of the proposed PDC approach can be observed via the simulation results, which demonstrate that the PDC approach outperforms the conventional IEEE 802.16 transmission mechanism in terms of user throughput.

6. Acknowledgments

This work was in part funded by the Aiming for the Top University and Elite Research Center Development Plan, NSC 96-2221-E-009-016, NSC 98-2221-E-009-065, the MediaTek research center at National Chiao Tung University, the Universal Scientific Industrial (USI) Co., and the Telecommunication Laboratories at Chunghwa Telecom Co. Ltd, Taiwan.

7. References

- Abichar, Z., Peng, Y. & Chang, J. M. (2006). WiMAX: The emergence of wireless broadband, *IEEE IT Prof.* 8(4): 44–48.
- Cordeiro, C., Abhyankar, S. & Agrawal, D. P. (2003). A dynamic slot assignment scheme for slave-to-slave and multicast-like communication in Bluetooth personal area networks, *Proc. IEEE Global Telecommunications Conf. (GLOBECOM)*, San Francisco, CA, pp. 4127–4132.
- IEEE P802.11zTM/D5.0 (2009). *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 6: Extensions to Direct Link Setup (DLS)*, IEEE, 3 Park Avenue, New York, NY 10016-5997, USA.
- IEEE Std. 802.16-2004 (2004). *IEEE Standard for Local and Metropolitan Area Networks- Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, IEEE, 3 Park Avenue, New York, NY 10016-5997, USA.
- Katevenis, M., Sidiropoulos, S. & Courcoubetis, C. (1991). Weighted round-robin cell multiplexing in a general-purpose atm switch chip, *IEEE J. Sel. Areas Commun.* 9(8): 1265–1279.
- Shreedhar, M. & Varghese, G. (1996). Efficient fair queueing using deficit round robin, *IEEE/ACM Trans. Netw.* 4(3): 375–385.
- Zhang, W., Zhu, H. & Cao, G. (2002). Improving Bluetooth network performance through a time-slot leasing approach, *Proc. IEEE Wireless Communications and Networking Conf. (WCNC)*, Orlando, FL, pp. 592–596.

Holdoff Algorithms for IEEE 802.16 Mesh Mode in Multi-hop Wireless Mesh Networks

Bong Chan Kim¹ and Hwang Soo Lee²

¹*Samsung Electronics*

²*KAIST*

Republic of Korea

1. Introduction

Multi-hop wireless mesh networks (M-WMNs) [Akyildiz, I. F., 2005] are one of the key features of beyond 3G systems because of their flexibility and low-cost deployment. So far, most of existing studies on multi-hop wireless mesh networks have been accomplished based on the IEEE 802.11 ad hoc mode. The IEEE 802.16 working group (WG) specified the IEEE 802.16-2004 standard [IEEE Std. 802.16-2004, 2004] in October 2004 and the standard defined two modes: the point-to-multi-point (PMP) mode and the mesh mode. The IEEE 802.16 mesh standard defines three mechanisms to schedule the data transmission: centralized scheduling (CSCH) [Morge, P. S., 2007], [Han, B., 2007], coordinated distributed scheduling (C-DSCH) [Morge, P. S., 2007], and uncoordinated distributed scheduling (Un-DSCH). In the IEEE 802.16 mesh mode with the CSCH, C-DSCH, and Un-DSCH, multi-hop communication is possible between nodes such as mesh base stations (MeshBSs) and mesh subscriber stations (MeshSSs) because all nodes are peers and each node can act as routers to support multi-hop packet forwarding. In particular, in the IEEE 802.16 mesh mode with the C-DSCH, every node competes for channel access using a distributed election algorithm (DEA) based on the scheduling information of the extended neighborhoods (one-hop and two-hop neighbors) in a completely distributed manner and reserves radio resource by a three-way handshaking mechanism in which nodes request, grant, and confirm available radio resource using mesh distributed scheduling (MSH-DSCH) message. Like this, because the IEEE 802.16 mesh mode with the C-DSCH has good flexibility and scalability, it is suitable as an alternative medium access control (MAC) protocol for establishing M-WMNs. For M-WMNs to serve as a wireless network infrastructure, the protocol design for M-WMN should target a high network throughput. In the IEEE 802.16 mesh mode with the C-DSCH, after occupying radio resource, a node cannot transmit any MSH-DSCH message for a holdoff time in order to share radio resource with other nodes in M-WMN. If nodes get a short holdoff time in a heavily loaded network situation, the competition between nodes will happens severe and thus they will experience long contention times before reserving radio resource. On the other hand, if nodes get a long holdoff time in a lightly loaded

¹ This work was performed while the first author was a Ph.D. student.

2.1 Frame structure of IEEE 802.16 mesh mode

Figure 1 shows the frame structure of the IEEE 802.16 mesh mode. As shown in the figure, the frame of the IEEE 802.16 mesh mode is based on the time division multiple access (TDMA) frame structure and consists of the control and data subframes.

The data subframe is used to transmit data packets and the control subframe is used to transmit MAC management messages. The control and data subframes are composed of multiple transmission opportunities (TOs) and minislots, respectively. As one of network parameters, the number of TOs consisting of the control subframe (MSH_CTRL_LEN) is set to one value between 0 and 15. Each TO in the control subframe consists of seven OFDM symbols and can carry only one MAC management message.

The control subframe consists of two types of subframes: the network control subframe and the schedule control subframe. The network control subframe enables new nodes to join mesh network. In a mesh network, nodes broadcast network entry and network configuration information and this enables a new node to get synchronization and initial network entry into the mesh network. The schedule control subframe is used to transmit MAC management messages in order to reserve minislots in the data subframe and it is divided into two parts, as shown in Fig. 1. The first part is used for the MSH-CSCH and MSH-CCFG messages transmissions in the CSCH mechanism and the second part is used for the MSH-DSCH message transmission in the C-DSCH mechanism.

The number of TOs (MSH_DSCH_NUM) consisting of the C-DSCH part is selected in a range between 0 and 15 and thus, the length of the CSCH part is equal to MSH_CTRL_LEN - MSH_DSCH_NUM.

2.2 MSH-DSCH message format

In this chapter, because we focus on holdoff algorithms for the IEEE 802.16 mesh mode with the C-DSCH, only the MSH-DSCH message format related with the C-DSCH is given in detail. Every node sends its available resource information to neighbor nodes via MSH-DSCH messages. The request, grant, and confirmation of resource are also accomplished by exchanging the MSH-DSCH messages between a pair of nodes.

As shown in Fig. 2, in the C-DSCH, the MSH-DSCH message contains the following information elements (IE): [Morge, P. S., 2007]

- *MSH-DSCH_Scheduling_IE* includes the next MSH-DSCH transmission times and holdoff exponents of a node and its neighbor nodes.
- *MSH-DSCH_Request_IE* is used by a node to specify its bandwidth demand for a specific link.
- *MSH-DSCH_Availability_IE* is used by a node to convey its own status for individual minislots to its neighbors.
- *MSH-DSCH_Grant_IE* is used by a node to send bandwidth grant in response to a bandwidth request as well as to send a grant confirmation for a received bandwidth grant.

Syntax	Size
MSH-DSCH_Message_Format() {	
Management_Message_Type = 41	8 bits
Coordination Flag	1 bit
Grant/Request Flag	1 bit
Sequence Counter	6 bits
No. Requests	4 bits
No. Availabilities	4 bits
No. Grants	6 bits
reserved	2 bits
if(Coordination_Flag == 0)	
MSH-DSCH_Scheduling_IE()	variable
for (i = 0; I < No_Requests; i ++)	
MSH-DSCH_Request_IE()	16 bits
for (i = 0; I < No_Availabilities; i ++)	
MSH-DSCH_Availability_IE()	32 bits
for (i = 0; I < No_Grants; i ++)	
MSH-DSCH_Grant_IE()	40 bits
}	

Fig. 2. MSH-DSHC message format

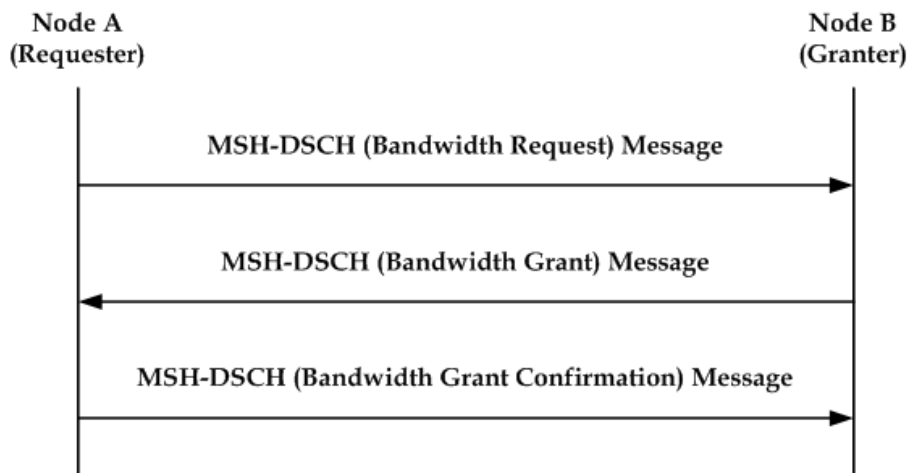


Fig. 3. Three-way handshaking

2.3 Bandwidth reservation by three-way handshaking

Based on the four IEs above, in the C-DSCH, nodes perform the bandwidth reservation process by three-way handshaking: bandwidth request, bandwidth grant, and bandwidth grant confirmation, as shown in Fig. 3.

In the bandwidth reservation procedure, node A (requester) uses the Link ID to uniquely identify a link for which node A needs bandwidth, and it sends an MSH-DSCH message that contains a set of MSH-DSCH_Request_IEs and a set of MSH-DSCH_Availability_IEs. The MSH-DSCH message with an MSH-DSCH_Request_IE is received by all the neighbors around node A. After receiving the bandwidth request message, node B (granter) looks up the set of available minislots in order to select a subset of minislots that is available for data receptions from node A. Node B chooses an available range of minislots for a bandwidth grant and sends the MSH-DSCH message with the MSH-DSCH_Grant_IE, which contains the set of minislots for the bandwidth grant. All the neighbors around node B receive the bandwidth grant message and they update their availability status by reflecting the scheduled data reception specified in the bandwidth grant message.

In the bandwidth grant confirmation, node A transmits an MSH-DSCH message with the MSH-DSCH_Grant_IE in order to inform all the neighbors around node A of the scheduled data transmission information. The neighbors update their availabilities by reflecting the newly scheduled data transmission. Data transmission is accomplished only over the reserved minislots, after the successful transmission of the bandwidth grant confirmation message.

2.4 Distributed election algorithm

In the three-way handshaking process, nodes independently select their own transmission times of MSH-DSCH messages using the distributed election algorithm. The DEA enables nodes to forward MSH-DSCH messages in an M-WMN in a completely distributed manner. A node can transmit an MSH-DSCH message without message collision at a selected

transmission time by the DEA. In the C-DSCH, the transmission time corresponds to a specific TO in the schedule control subframe shown in Fig. 1. In the DEA, a node performs two functions: collecting the next transmission times (Next_Xmt_Times) of all nodes within its extended neighborhood (one-hop and two-hop neighbors) and selecting its Next_Xmt_Time.

1) *Collecting the Next_Xmt_Times of all nodes within extended neighborhood*: to select a collision-free TO for the MSH-DSCH transmission, every node should collect the Next_Xmt_Times of all nodes within its extended neighborhood. So, every node must inform its neighbors of the next MSH-DSCH transmission time of its neighbors as well as itself. In the DEA, nodes broadcast not an exact Next_Xmt_Time but the next transmission time interval (Next_Xmt_Time_Interval) information, which is expressed by two parameters of next transmission maximum (Next_Xmt_Mx) and transmission holdoff exponent (Xmt_Holdoff_Exp) as follows:

$$2^{Xmt_Holdoff_Exp} * Next_Xmt_Mx < Next_Xmt_Time$$

$$< 2^{Xmt_Holdoff_Exp} * (Next_Xmt_Mx + 1) \quad (1)$$

The Next_Xmt_Time_Interval is a series of one or more C-DSCH TOs. Every node broadcasts MSH-DSCH message containing the next transmission time interval information (Next_Xmt_Mx and Xmt_Holdoff_Exp) of its one-hop neighbors as well as itself. Using the next transmission interval information received from neighbors via MSH-DSCH messages, every node can calculate the Next_Xmt_Time_Intervals of all nodes within its extended neighborhood. In the IEEE 802.16 mesh standard, a node first selects its own Next_Xmt_Time at a current MSH-DSCH transmission time (Current_Xmt_Time) and then, it calculates a Next_Xmt_Mx value using equation (1) and a current Xmt_Holdoff_Exp value. Next, the node broadcasts an MSH-DSCH message containing the Next_Xmt_Mx and Xmt_Holdoff_Exp, which represents its Next_Xmt_Time_Interval, in order to inform neighbors of its next transmission time information.

2) *Selecting a Next_Xmt_Time*: in the IEEE 802.16 mesh standard, after an MSH-DSCH message transmission, a node is not eligible to transmit an MSH-DSCH message for transmission holdoff time (Xmt_Holdoff_Time) in order to share radio resource with other nodes in an M-WMN. Therefore, when a node selects a Next_Xmt_Time in the DEA, it first sets the temporary transmission time (Temp_Xmt_Time) as follows:

$$Temp_Xmt_Time = Current_Xmt_Time + Xmt_Holdoff_Time + 1 \quad (2)$$

Next, the node should determine the set of eligible competing nodes related to the Temp_Xmt_Time from its neighbor table. This set will include neighbors that meet at least one of the following conditions: [Bayer, N., 2006]

- The Next_Xmt_Time_Interval of neighbor includes the Temp_Xmt_Time,
- The earliest subsequent transmission time (Earliest_Subsequent_Xmt_Time) of neighbor is equal to or smaller than Temp_Xmt_Time, or
- The Next_Xmt_Time of neighbor is not known.

In the conditions above, the *Earliest_Subsequent_Xmt_Time* is the earliest time that the node is eligible to transmit a MSH-DSCH message after the *Next_Xmt_Time*, as shown in equation (3):

$$\begin{aligned} \text{Earliest_Subsequent_Xmt_Time} = & \text{Next_Xmt_Time} + \text{Xmt_Holdoff_Exp} \\ & + 2^{\text{Xmt_Holdoff_Exp}} * \text{Next_Xmt_Mx} \end{aligned} \quad (3)$$

The DEA is performed based on this set of eligible competing nodes. With this set, a node checks whether any competing nodes do not use a specific TO corresponding to the *Temp_Xmt_Time*, or not. If any competing nodes of the node do not use the specific TO, the node is the winner of the distributed election and the *Next_Xmt_Time* is set equal to the *Temp_Xmt_Time*. If a node does not win the distributed election, the *Temp_Xmt_Time* is set to *Temp_Xmt_Time* + 1 and the above process is performed again in order to select a collision-free *Next_Xmt_Time*.

3. Holdoff algorithms for IEEE 802.16 mesh mode

In the IEEE 802.16 mesh standard with the C-DSCH, a node is not eligible to transmit any MSH-DSCH messages for at least holdoff time after an MSH-DSCH message transmission. The reason to stop an MSH-DSCH message transmission for a holdoff time is for nodes to share radio resource with other nodes in M-WMN. However, a holdoff mechanism can result in resource waste by holding MSH-DSCH message transmission even in a lightly loaded network situation. Some holdoff algorithms have been proposed for the IEEE 802.16 mesh mode with the C-DSCH in order to improve the network throughput and to share the radio resource.

3.1 Static holdoff algorithm

In the IEEE 802.16 mesh standard, the *Xmt_Holdoff_Time* is calculated based on the static *Xmt_Holdoff_Exp* and *Xmt_Holdoff_Exp_Base* values, as shown in equation (4):

$$\text{Xmt_Holdoff_Time} = 2^{(\text{Xmt_Holdoff_Exp} + \text{Xmt_Holdoff_Exp_Base})} \quad (4)$$

The IEEE 802.16 mesh standard defines static holdoff algorithm. In the static holdoff algorithm, the *Xmt_Holdoff_Exp_Base* is fixed to 4 for resource sharing between nodes in an M-WMN. The *Xmt_Holdoff_Exp* is also set to a specific value between 0 and 7 in initial node configuration procedure. That is, all nodes in M-WMN have an identical transmission holdoff time regardless of current network situation. Thus, when *Xmt_Holdoff_Exp* = 0, the competition between nodes happen severe and a node experiences long contention times before reserving resource. On the other hand, as the *Xmt_Holdoff_Exp* value increases, the contention between nodes becomes less competitive; however, nodes get a longer holdoff time and thus transmission interval becomes longer [Cao, M., 2005]. In particular, if the *Xmt_Holdoff_Exp* is set to an unnecessarily large value in a lightly loaded network situation, the waste of resource happens. Hence, the static holdoff algorithm can result in network throughput degradation regardless of an assigned *Xmt_Holdoff_Exp* value.

3.2 Dynamic holdoff algorithm

In [Cao, M., 2005], the authors performed the modelling and performance analysis of distributed scheduler in the IEEE 802.16 mesh mode, and they concluded that the capacity of IEEE 802.16 mesh network can be optimized by assigning appropriate $Xmt_Holdoff_Exp$ values to nodes in the network.

In [Bayer, N., 2007], the dynamic exponent (DynExp) algorithm was proposed. In the DynExp algorithm, nodes that are currently not sending, receiving, or forwarding data packets use large $Xmt_Holdoff_Exp$ values and thus get large $Xmt_Holdoff_Time$ values. Nodes that transmit, receive, and forward data packets or nodes that have been selected by the routing protocol as potential forwarding nodes use small $Xmt_Holdoff_Exp$ values and thus get small $Xmt_Holdoff_Time$ values. For this purpose, nodes are classified as follows:

- Mesh base station (M-BS) is a normal mesh base station.
- Active node (ACT) is a node that is part of an active route and sends, receives, or forwards data packets.
- Sponsor node (SN) is a node that is not part of an active route but has been selected as a potential forwarding node by at least one of its neighbors.
- Inactive node (IN-ACT) is a node that is not part of an active route and does not send, receive or forward data packets.

According to the node types above, different $Xmt_Holdoff_Exp$ values are defined as follows:

$$\begin{aligned}
 0 < Xmt_Holdoff_Exp_{M-BS} &< Xmt_Holdoff_Exp_{ACT} \\
 &< Xmt_Holdoff_Exp_{SN} \\
 &< Xmt_Holdoff_Exp_{IN-ACT}'
 \end{aligned}$$

By using different $Xmt_Holdoff_Exp$ values according to current node types, the DyExp algorithm improves network capacity. However, because the operation of DynExp algorithm depends on information from the routing layer, the DynExp algorithm cannot be operated independently without cooperation with routing layer.

3.3 Adaptive holdoff algorithm

In this section, we propose an adaptive holdoff algorithm. In the adaptive holdoff algorithm, nodes adaptively adjust the $Xmt_Holdoff_Exp$ value according to current node state.

3.3.1 Definitions

Before proposing the adaptive holdoff algorithm, the neighbors of a node are classified into four types: tx-neighbor, rx-neighbor, tx/rx-neighbor, and null-neighbor. Each classification is defined as follows:

- A tx-neighbor of a node is a neighbor that has a data packet to send to the node.
- An rx-neighbor of a node is a neighbor to which the node has a data packet to send.
- A tx/rx-neighbor of a node is a neighbor that is both tx-neighbor and rx-neighbor of the node.

- A null-neighbor of a node is a neighbor that has no data packet to send to or receive from the node.

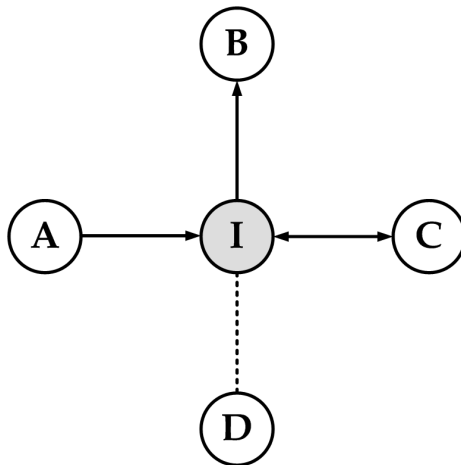


Fig. 4. Classification of neighbor nodes: A solid line between a pair of nodes indicates that there is at least one data packet to transmit between the two nodes, with an arrow showing the direction of the pending data transmission. A dotted line indicates that there is no data waiting for transmission between the pair of nodes

In Fig. 4, node I has four types of neighbors. Nodes A, B, C, and D correspond to the tx-neighbor, rx-neighbor, tx/rx-neighbor, and null-neighbor of node I, respectively.

3.3.2 Adaptive holdoff algorithms based on node state

In this section, an adaptive holdoff algorithm based on node state is presented. The adaptive holdoff algorithm does not depend on information from higher layer such as routing layer and it operates independently. In addition, it maintains backward compatibility with the IEEE 802.16 mesh standard.

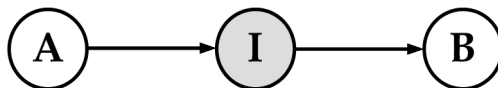


Fig. 5. Example topology: A solid line between a pair of nodes indicates that there is at least one data packet to transmit between the two nodes, with an arrow showing the direction of the pending data transmission

In Fig. 5, node I has two links (Links 1 and 2) for communication with nodes A and B, respectively. As shown in Fig. 5, node I should reserve each minislots for data reception from node A and data transmission to node B. On the other hand, node A needs to reserve only minislots for data transmission to node I, and node B needs to reserve only minislots for data reception from node I. Therefore, node I should be able to access schedule control subframe at a higher rate than nodes A and B in order to prevent node I from being a bottleneck.

Node_ID	Type_Tx	Type_Rx	Expire_Tx	Expire_Rx	Expire
A	1	0	nonzero	0	nonzero
B	1	0	nonzero	0	nonzero
C	1	1	nonzero	nonzero	nonzero
D	0	0	0	0	nonzero

Fig. 6. Neighbor table of node I in Fig. 4

Based on the features above, the *Weight* of a node is first defined for the adaptive holdoff algorithm as follows:

$$Weight = tx_exist * 1 + rx_exist * 2 \quad (5)$$

In equation (5), the *tx_exist* represents whether a node has at least one tx-neighbor or not. And, the *rx_exist* indicates whether a node has at least one rx-neighbor or not. Because a node transmits only one MSH-DSCH message (bandwidth grant message) to tx-neighbor over the TO of the schedule control subframe during the three-way handshaking, the *tx_exist* is multiplied by 1 in equation (5). On the other hand, because a node transmits two MSH-DSCH messages (bandwidth request and bandwidth grant confirmation messages) to rx-neighbor during the three-way handshaking, the *rx_exist* is multiplied by 2. In addition, as shown in Fig. 2, the MSH-DSCH message can contains multiple MSH-DSCH_Request_IE(s), MSH-DSCH-Availability_IE(s), and MSH-DSCH_Grants_IE(s) at one time. Namely, a node can simultaneously request, grant, and confirm resources for multiple links via only one MSH-DSCH message. Therefore, the *Weight* depends on only the *tx_exist* and *rx_exist* in equation (5).

Next, a neighbor table is presented in order to explain how to obtain the *Weight* of a node. In the adaptive holdoff algorithm, nodes store neighbor information in their neighbor table, as shown in Fig. 6. Each entry in the neighbor table contains Node_ID, Type_Tx/Rx, Expire_Tx/Rx, and Expire. Node_ID denotes the identifier of a neighbor. Type_Tx/Rx is used to distinguish the neighbor types: tx-neighbor, rx-neighbor, tx/rx-neighbor, or null-neighbor. Expire_Tx/Rx indicates the expiration time of a neighbor as a tx-neighbor or as an rx-neighbor. Expire represents the expiration time of a neighbor table entry.

In the adaptive holdoff algorithm, the neighbor table is managed in the following way:

1) *Setting Type_Tx/Rx, Expire_Tx/Rx, and Expire*: Whenever a node sends or receives a data packet to or from a neighbor over minislots of data subframe, it sets the values of Type_Tx/Rx, Expire_Tx/Rx, and Expire in the neighbor table entry for the neighbor. When a node receives a data packet from a neighbor, it finds the neighbor table entry for that neighbor in its neighbor table and sets the values of Type_Tx, Expire_Tx, and Expire as follows: the Type_Tx is set to 1. The Expire_Tx and Expire are set to EXPIRE_TIME, which is

a pre-defined constant value. When a node sends a data packet to a neighbor, it finds the neighbor table entry for that neighbor in its neighbor table and sets Type_Rx, Expire_Rx, and Expire as follows: the Type_Rx is set to 1. The Expire_Rx and Expire are set to EXPIRE_TIME.

2) *Maintaining/removing neighbor table entries*: Maintenance and removal of a neighbor table entry is performed using the neighbor timer. The neighbor timer expires periodically; when this occurs, a node compares the current time with each values of Expire_Tx/Rx and Expire in a neighbor table entry and updates Type_Tx/Rx, Expire_Tx/Rx, and Expire in the following way: if the values of Expire_Tx/Rx are higher than the current time, the values of Type_Tx/Rx and Expire_Tx/Rx are not changed; otherwise, they are set to 0. If the value of Expire is less than the current time, the neighbor table entry is removed; otherwise, the value of Expire is not changed. This procedure is repeated for all the entries in the neighbor table.

According to the neighbor table management above, node I in Fig. 4 has the neighbor table as shown in Fig. 6. In the adaptive holdoff algorithm, a node can calculate its *Weight* using its neighbor table. For example, node I can calculate the *Weight* by checking the values of Type_Tx/Rx in its neighbor table entries as follows: if a node has at least one neighbor entry with the Type_Tx value of 1 in its neighbor table, the tx_exist is set to 1; otherwise, the tx_exist is set to 0. In a similar way, if a node has at least one neighbor entry with the Type_Rx value of 1 in its neighbor table, the rx_exist is set to 1; otherwise, the rx_exist is set to 0. Therefore, node I in Fig. 4 has the *Weight* value of 3 ($= 1 * 1 + 1 * 2$).

In the adaptive holdoff algorithm, the *Weight* of a node represents its access rate required to reserve a shared resource for data transmission or reception. Therefore, a node with a large *Weight* value should be able to access the TOs of the schedule control subframe at a high rate, and a node with a small *Weight* value should access the TOs of the schedule control subframe at a low rate.

To achieve this goal, in the adaptive holdoff algorithm, current transmission holdoff exponent (Cur_Xmt_Holdoff_Exp) is defined as follows:

$$Cur_Xmt_Holdoff_Exp = Max_Xmt_Holdoff_Exp * (1 - Weight / 3) \quad (6)$$

In equation (6), the Max_Xmt_Holdoff_Exp is the maximum value of transmission exponent. The Max_Xmt_Holdoff_Exp of a node is set to a constant in initial node configuration process.

And then, a node selects its Xmt_Holdoff_Time as follows:

$$Xmt_Holdoff_Time = 2^{(Xmt_Holdoff_Base + Cur_Xmt_Holdoff_Exp)} \quad (7)$$

Algorithm 1 shows the operation of the adaptive holdoff algorithm. In Algorithm 1, it is assumed that the Xmt_Holdoff_Exp_Base is set to 4 and the Max_Xmt_Holdoff_Exp to 3.

As shown in Algorithm 1, if a node has a large *Weight* value, the small Xmt_Holdoff_Time value is obtained by setting the Cur_Xmt_Holdoff_Exp to a small value. Thus, the node can access the TOs of the schedule control subframe for resource reservation at a high rate. On the other hand, if a node has a small *Weight* value, the large Xmt_Holdoff_Time value is obtained by setting the Cur_Xmt_Holdoff_Exp to a large value. Thus, the node can access the TOs of the schedule control subframe for resource reservation at a low rate. In this way, nodes can adjust their access rate to the TOs of the schedule control subframe according to their current state.

Algorithm 1 *Adaptive Holdoff Algorithm based on Node State*

Notations Used:

tx_exist = indicate whether a node currently has at least one tx-neighbor

rx_exist = indicate whether a node currently has at least one rx-neighbor

$Weight$ = access rate of a node

$Current_Xmt_Holdoff_Exp$ = current transmission holdoff exponent

$Xmt_Holdoff_Exp_Base$ = transmission holdoff exponent base (= 4)

$Max_Xmt_Holdoff_Exp$ = maximum value of transmission holdoff exponent (= 3)

$Xmt_Holdoff_Time$ = transmission holdoff time

01: Set the tx_exist by cheking the Type_Txs of neighbor table entries

02: Set the rx_exist by cheking the Type_Rxs of neighbor table entries

03: Compute $Weight = tx_exist * 1 + rx_exist * 2$

04: **if** ($Weight == 0$)

05: $Current_Xmt_Holdoff_Exp = Max_Xmt_Holdoff_Exp$;

06: **end if**

07: **else**

08: $Current_Xmt_Holdoff_Exp = Max_Xmt_Holdoff_Exp * (1 - Weight / 3)$;

09: **end else**

10: Compute $Xmt_Holdoff_Time = 2^{(Xmt_Holdoff_Exp_Base + Current_Xmt_Holdoff_Exp)}$

4. Simulation Evaluation

Computer simulations were performed to evaluate the performance of the holdoff algorithms. To show the importance of adjusting holdoff time according to current network situation from the perspective of performance improvement, the performance of the adaptive holdoff algorithm based on node state (AHA) is compared with that of the static holdoff algorithm (SHA), which is described in the IEEE 802.16 mesh standard.

4.1 Simulation environments

NCTUns-4.0 [Wang, S. Y., 2007-(a)], [Wang, S. Y., 2007-(b)], [Wang, S. Y., 2007-(c)] is used to evaluate the performance of the holdoff algorithms for the IEEE 802.16 mesh mode with the C-DSCH. Table 1 shows the parameters used in the simulation. For all other parameters, the default values provided in NCTUns-4.0 are used. Figure 7 shows the network topology used to evaluate the performance of the holdoff algorithms. The network consists of 16 nodes. Among 16 nodes, one node acts as MeshBS and the others as MeshSSs. Node 11 corresponds to the MeshBS and other nodes to MeshSSs. All nodes remain stationary for a simulation time of 300s. The number of traffic flows is varied from 1 to 6 to investigate the performance variation in different offered loads. Each traffic source generates user datagram protocol (UDP)-based data packets with the size of 512 bytes at a rate of 2Mbits/s.

Parameters	Value
Max. transmission range	MeshSS: 400m, MeshBS: 400m
Bytes per OFDM symbol	108 (64-QAM_3/4)
Xmt_Holdoff_Exp_Base	4
Xmt_Holdoff_Exp	3 (only static holdoff algorithm)
Distance between MeshSSs	300m

Table 1. Simulation parameters for holdoff algorithms

Simulations are performed in two scenarios: multi-hop peer-to-peer and multi-hop Internet access scenarios. Figure 8 shows traffic flows in the multi-hop peer-to-peer scenario. The multi-hop Internet access scenarios are classified into upload and download patterns. The traffic flows for upload and download patterns are shown in Figs. 9 and 10, respectively. To evaluate the performances of two holdoff algorithms (SHA and AHA), the following performance metrics are used:

- The average per-flow throughput is the average packet throughput at a receiver.
- The total network throughput is the sum of the average packet throughputs at all receivers.

4.2 Simulation results

In the multi-hop peer-to-peer scenario where traffic is distributed throughout entire network, the performance of AHA is compared with that of SHA. The average per-flow throughput and total network throughput are shown in Figs. 11 and 12, respectively. In the SHA, because all nodes have an identical Xmt_Holdoff_Exp value (= 3), they have an identical

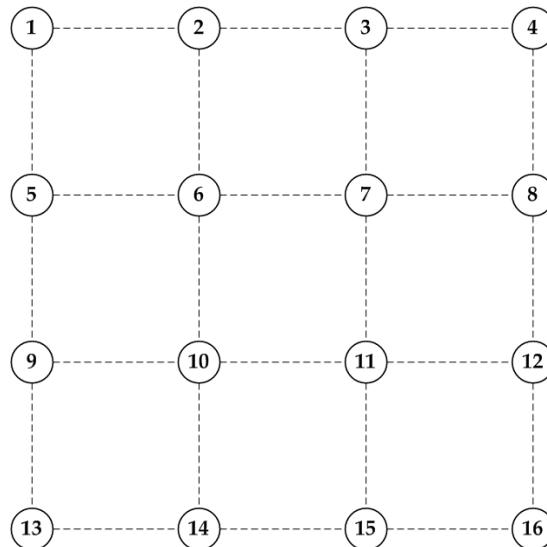


Fig. 7. Simulation network topology

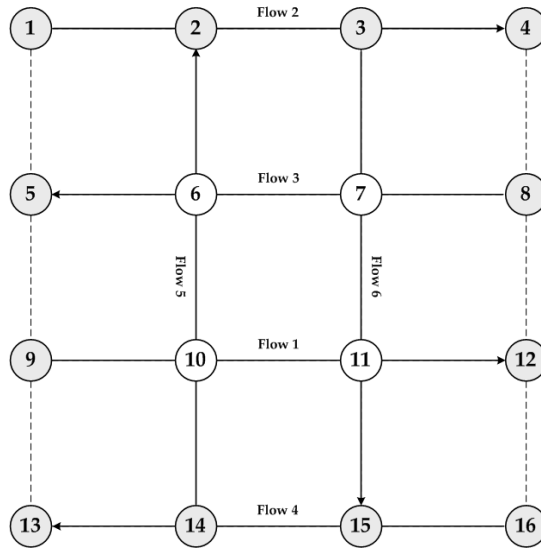


Fig. 8. Multi-hop peer-to-peer

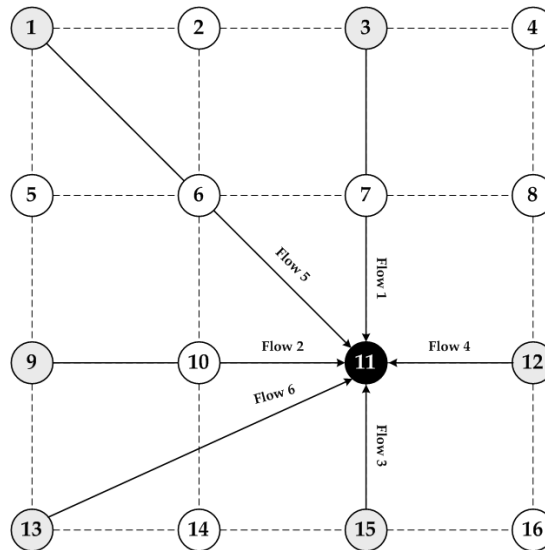


Fig. 9. Multi-hop Internet access with upload pattern

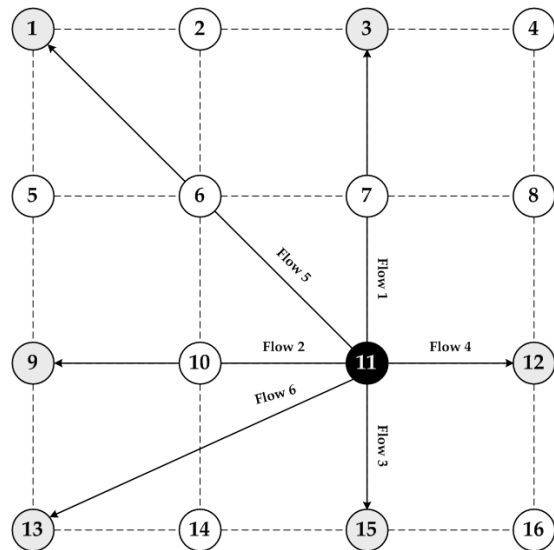


Fig. 10. Multi-hop Internet access with download pattern

holdoff time regardless of their current state. However, in the AHA, because nodes adjust their $Xmt_Holdoff_Exp$ in a range between 0 and 3 according to their current state, the radio resource can be used efficiently.

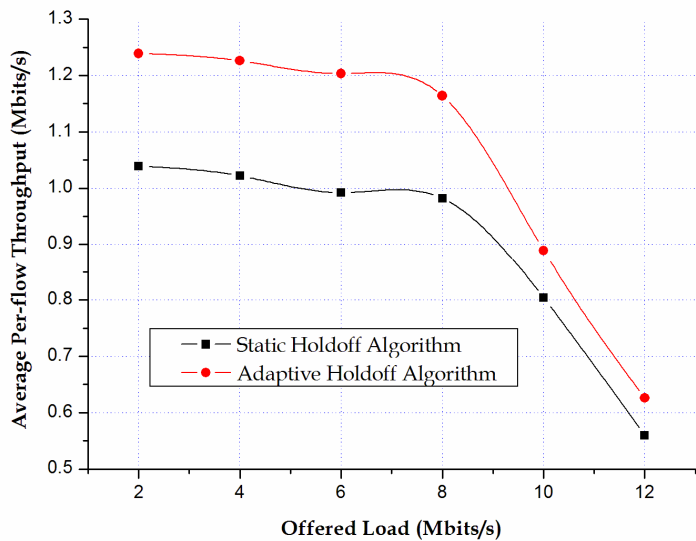


Fig. 11. Average per-flow throughput in multih-hop peer-to-peer

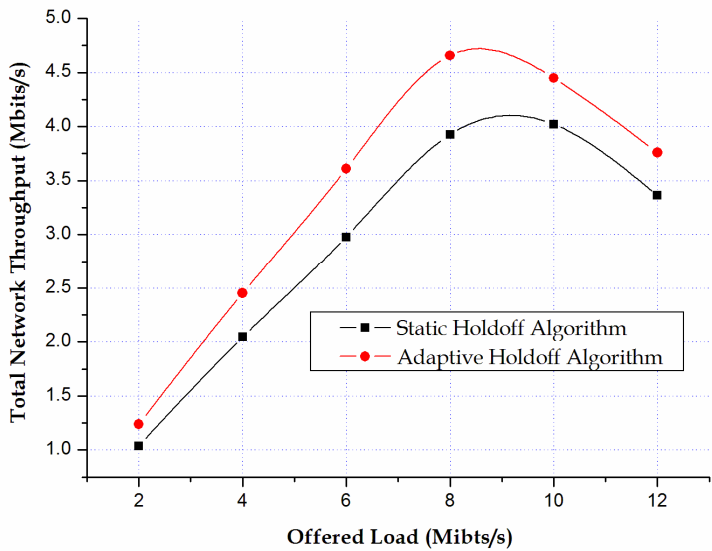


Fig. 12. Total network throughput in multi-hop peer-to-peer

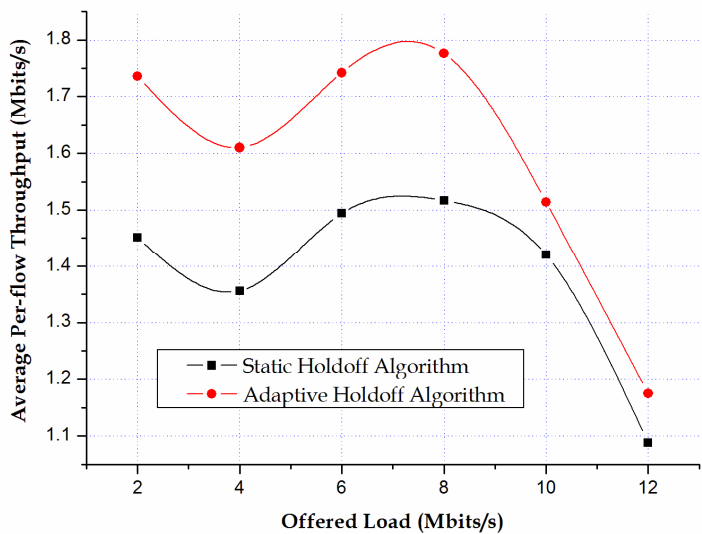


Fig. 13. Average per-flow throughput in multi-hop Internet access - upload

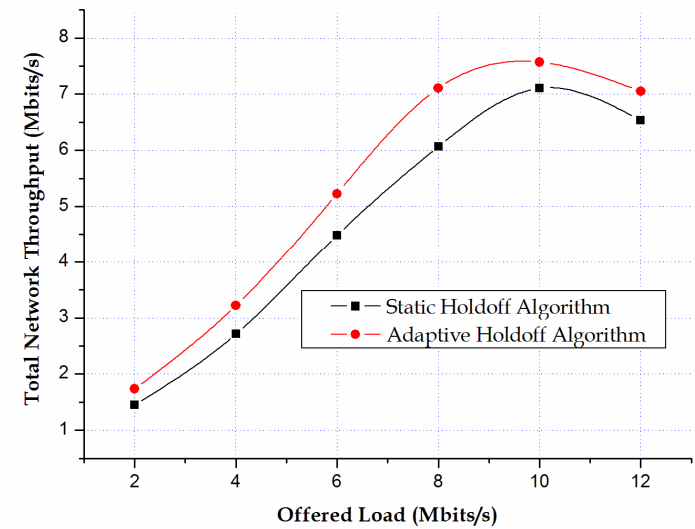


Fig. 14. Total network throughput in multi-hop Internet access - upload

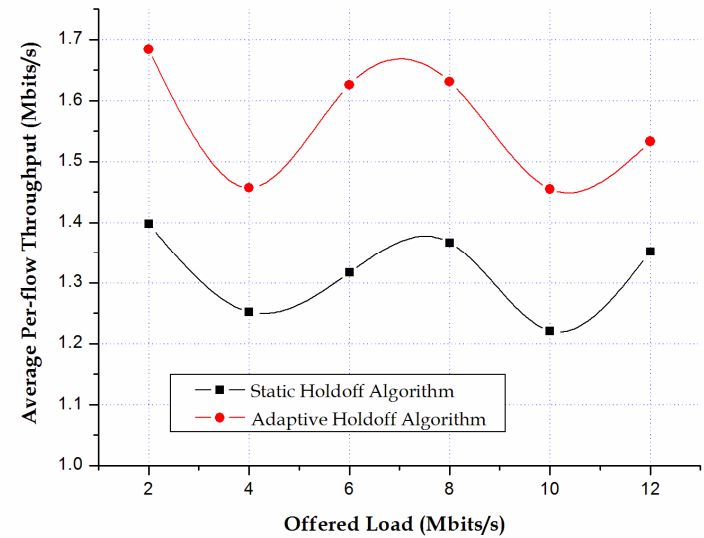


Fig. 15. Average per-flow throughput in multi-hop Internet access - download

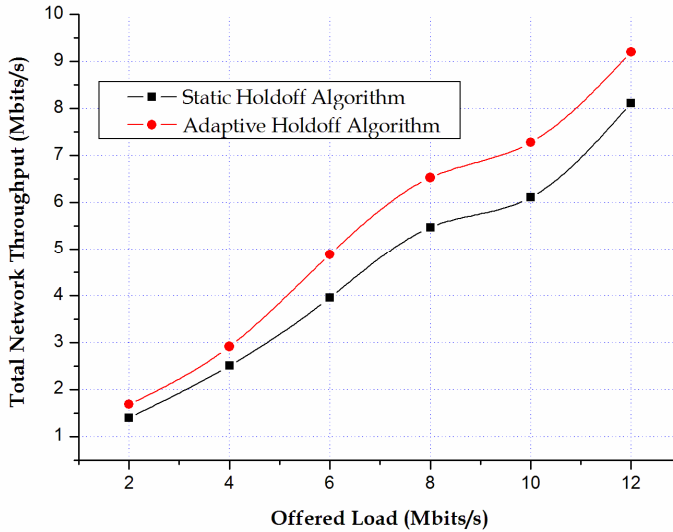


Fig. 16. Total Total network throughput in multi-hop Internet access - download

For example, because a node with the *Weight* = 0 currently has no data communication with neighbors, even though the node uses the *Xmt_Holdoff_Exp* of 3 in the holdoff process, the communication problem does not happen. Furthermore, that a node with a small *Weight* value reduces the access rate to the schedule control subframe enables other nodes with a large *Weight* value to access the schedule control subframe at a high rate.

A node with the *Weight* = 3 can access the schedule control subframe at a high rate by setting the *Xmt_Holdoff_Exp* to 0. By adaptively adjusting the *Xmt_Holdoff_Exp* value according to the current node state, the AHA performs better than the SHA in terms of average per-flow throughput and total network throughput, as shown in Figs. 11 and 12.

In general, access to the Internet through MeshBS is desirable for MeshSSs to obtain necessary service. In the multi-hop Internet access scenario, it frequently happens that MeshSSs upload files to Internet through MeshBS and MeshSSs download files provided in the Internet through MeshBS. As illustrated in Figs. 9 and 10, computer simulations are also performed in the multi-hop Internet access scenario with upload and download patterns. Figures 13 and 14 show the average per-flow throughput and total network throughput in the multi-hop Internet access scenario with upload pattern, respectively. In addition, Figs. 15 and 16 show the average per-flow throughput and total network throughput in the multi-hop Internet access scenario with download pattern, respectively. In the multi-hop Internet access scenario, it frequently happens that nodes act as one of sender or receiver. For example, node 12 serves only as sender in Fig. 9, and only as receiver in Fig. 10. In the AHA, because nodes set their *Xmt_Holdoff_Exp* to an appropriate value according to their current role, the AHA outperforms the SHA, as shown in Figs. 13, 14, 15, and 16.

5. Conclusion

Multi-hop wireless mesh networks are one of the key features of beyond 3G system because of their flexibility and low-cost deployment. The IEEE 802.16 mesh mode with the C-DSCH has recently emerged as an alternative MAC protocol for establishing M-WMNs. For the IEEE 802.16 mesh mode to serve as a MAC protocol for M-WMN, it should get a high network throughput. However, the holdoff algorithm of the IEEE 802.16 mesh standard has a limitation to the performance improvement of M-WMN. In this chapter, we dealt with existing holdoff algorithms such as static holdoff algorithm and dynamic holdoff algorithm and introduced their limitations. In addition, we proposed an adaptive holdoff algorithm based on node state for the IEEE 802.16 mesh mode with the C-DSCH. The adaptive holdoff algorithm assigns an appropriate Xmt_Holdoff_Exp value according to current network situation and it maintains the backward compatibility with the IEEE 802.16 mesh standard. Simulation results show that it is required for nodes to adaptively adjust their holdoff time according to current network situations in order to improve the performance of an IEEE 802.16 mesh system.

6. Acknowledgments

This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) and MMPC (Mobile Media Platform Center) support programs supervised by the IITA (Institute of Information Technology Advancement)

7. References

- Han, B., Jia, W., and Lin, L. (2007). Performace Evaluation of Scheduling in IEEE 802.16 based Wireless Mesh Networks, Elsevier Computer Communications, Vol. 30, No. 4, Feb. 2007, pp. 782-792, ISSN 0140-3664
- Akyildiz, I. F., Wang, X., and Wang W. (2005). Wireless Mesh Networks : A Suurvey, Elsevier Computer Networks, Vol. 47, No. 4, Mar. 2005, pp. 445-487, ISSN 1389-1286
- IEEE Std. 802.16-2004 (2004). IEEE Standard for Local and Metropolitan Area Networks: Part 16: Air Interface for Fixed Broadband Wireless Access Systems, Oct. 2004
- Cao, M.; Ma, W.; Zhang, Q.; Wang, X. & Zhu, W. (2005). Modeling and Performance Analysis of the Distributed Scheduler in IEEE 802.16 Mesh Mode, *Proceedings of ACM MobiHoc*, pp. 78-89, ISBN 1-59593-004-3, Urbana-Champaign, May 2005, ACM, Illinois, USA
- Bayer, N.; Xu, B. ; Rakocevic, V. & Habermann, J. (2007). Improving the Performance of the Distributed Scheduler in IEEE 802.16 Mesh Networks, *Proceedings of IEEE VTC Spring*, pp. 1193-1197, ISBN 1-4244-0266-2, Burlington Hotel, Apr. 2007, IEEE, Dublin, Ireland
- Bayer, N. ; Sivchenko, D.; Xu, B.; Rakocevic, V. & Habermann, J. (2006). Transmission Timing of Signaling Messages in IEEE 802.16 based Mesh Networks, *Proceedings of European Wireless*, ISBN 978-3-8007-2961-6, National Technical University of Athens, Apr. 2006, Athens, Greece

- Morge, P. S.; Hollick, M. & Steinmetz, R. (2007). The IEEE 802.16-2004 MeSH Mode Explained Technical Report, KOM-TR-2006-08, Feb. 2007
- Wang, S. Y.; Huang, C. H.; Lin, C. C.; Chou, C. L. & Liao, K. C. (2007-(a)). The Protocol Developer Manual for the NCTUns 4.0 Network Simulator and Emulator, July 2007.
- Wang, S. Y.; Chou, C. L. & Lin, C. C. (2007-(b)). NCTUns Homepage. [Online]. Available: <http://nsl10.csie.nctu.edu.tw/>
- Wang, S. Y.; Chou, C. L. & Lin, C. C. (2007-(c)). The GUI User Manual for the NCTUns 4.0 Network Simulator and Emulator, July 2007

Call Admission Control Algorithms based on Random Waypoint Mobility for IEEE802.16e Networks¹

Khalil Ibrahimia^{a,b}, Rachid El-Azouzi^a, Thierry Peyrea
and El Houssine Bouyakhf^b

^aLIA/CERI, University of Avignon

339 chemin des Meinajariès B.P. 1228, 84911 cedex 9 - Avignon - France

^bLIMIARF/FSR, Mohammed V-Agdal University

4 Avenue Ibn Battouta B.P. 1014 Agdal - Rabat - Morocco

1. Introduction

The next generation network WiMAX (Worldwide Interoperability for Microwave Access), has become synonymous with the IEEE802.16 Wireless Metropolitan Area Network (MAN) air interface standard. In its original release the 802.16 standard addressed applications in licensed bands in the 10 to 66 GHz frequency range. Subsequent amendments have extended the 802.16 air interface standard to cover non-line of sight (NLOS) applications in licensed and unlicensed bands from 2 to 11 GHz bands. These 802.16 networks are able to provide high data rates and are preferably based, for NLOS applications, on Orthogonal Frequency Division Multiple Access (OFDMA) (Piggin, 2004). In OFDMA, modulation and/or coding can be chosen differently for each sub-carrier, and can also change with time. Indeed, in the IEEE802.16 standard, coherent modulation schemes are used starting from low efficiency modulations (BPSK with coding rate 1/2) to very high efficiency ones (64-QAM with coding rate 3/4) depending on the SNR (Signal-to-Noise Ratio). It has been shown that systems using adaptive modulation perform better than systems whose modulation and coding are fixed (Yaghoobi, 2004). Adaptive modulation increases data transmission throughput and the system reliability by using different constellation sizes on different sub-carriers.

The authors in (Peyre et al., 2008) introduce a new Quality of Service (QoS) for real-time calls in IEEE802.16e Multi-class Capacity including AMC scheme and QoS Differentiation for Initial and Bandwidth Request Ranging (Seo et al., 2004). The QoS defined in this work is to maintain a same bit rate for RT calls independently of user position in the cell. The authors use a Discrete Time Markov Chain (DTMC) to model the system over a decomposition of IEEE802.16e cell. The authors took account the mobility of users among

¹This work was supported by a research contract with Maroc Telecom R&D No. 10510005458.06 PI.

the regions of the cell. So, the analysis of wireless systems, as IEEE802.16e WiMAX, often requires to model the effect of user mobility. The mobility model is a critical element for any study about the radio communications. Therefore we choose to use the Random Waypoint (RWP) mobility model. This model have been studied largely in Ad-hoc networks (Johnson & Maltz, 1996) and briefly in wireless networks (Hytti & Virtamo, 2007). In particular we consider (Johnson & Maltz, 1996), wherein the authors introduced the RWP mobility model to find the mean arrival and departure rates into a concentric cell as well as the mean sojourn time in the cellular networks.

Call Admission Control schemes play an important part in radio resource management (Kobane et al., 2007), (Ibrahimi et al., 2008) and (Ibrahimi et al., 2009). Their aims are to maintain an acceptable QoS to different calls by limiting the number of ongoing calls in the system, minimize the call blocking and dropping probabilities and in the same time efficiently utilize the available resources (Niyato & Hossain, 2005). Our study is motivated by an attempt to find the better CAC in IEEE802.16e WiMAX, for both traffics RT and BE that handles the intra or inter cell mobility issue in the downlink of IEEE802.16e WiMAX with AMC technic. Based on the RWP mobility model for an IEEE802.16e cell, we model the system capacity through a Continuous Time Markov Chain (CTMC). For this reason we propose to study two promising CAC algorithms that guarantee a QoS. Our propositions allow to Internet Service Providers (ISP) to choice a CAC dependently of its purpose to manage their networks.

The rest of this paper is organized as follows: In Section 2, we describe the system and mobility properties as well as both CAC algorithms. In section 3, we develop the system and RWP mobility model. Based on it, the sections 4 and 5, develop the analysis for the first and second CAC schemes respectively. In Section 6 we define the performance metrics used to achieve the performance match. The section 7 provides some numerical results before conclude in section 8.

2. Framework

2.1. IEEE802.16 basis principles

The IEEE802.16e Physical layer uses an OFDMA sub-carrier allocation policy for the data transmission. The uplink and downlink sub-frames divide the time and frequency space into sub-carriers. The minimum frequency-time unit of sub-channelization is one slot, and a frame is constructed by a number of slots. Different sub-carriers are allocated to a mobile transmission in function of the resource requested by the mobile. Moreover, a sub-channel can be used periodically by different mobiles due to theirs classes of traffic. Once a mobile is granted to transmit by a bandwidth response in the DL-MAP, base station assigns one or more subcarriers and hence defines the sub-channels that the mobile will be able to use for its data transmission.

An IEEE802.16e cell is organized in region. Each region use specific modulation and coding technics. Users belongs to a region in function of theirs Signal-to-Noise Ratio SNR. The number of subcarriers allocated to a mobile directly depends on the available modulation, the type of traffic and the requested bandwidth.

2.2 System description

In this paper we consider a continuous time performance model for a single IEEE802.16e cell. The cell is decomposed into several regions. Mobiles are uniformly distributed on the whole cell. The cell population is dispatched between the regions according with their area coverages. Each region is characterized by the modulation used for data transmissions. Due to the AMC scheme described in the previous section, the calls use a modulation chosen in function of the receiver SNR. We consider the Adaptive Modulation and Coding (AMC) with pathloss only, consequently, the SNR depends only on the distance between the base station and the calling mobile. The four classes of services defined in the standard are:

- Unsolicited Grant Service (UGS) that caters for real-time, fixed-size data packets with constant bit rate (CBR). This service is granted at regular time intervals without a request or polls.
- Real-Time Polling Service (rtPS) that caters for real time data packets that vary in size generated at periodic intervals, such as MPEG video and VoIP with silence suppression, where packet sizes are variable.
- Non-Real-Time Polling Service (nrtPS), designed for connections that do not have delay requirements. nrtPS is similar to BE. It only differs from BE in that it guarantees a minimum bandwidth, e.g., FTP.
- Best-Effort (BE) that service makes no guarantee of service. To guarantee a minimum bandwidth the connection must subscribe to the nrtPS service, e.g., web browsing data.

We gather these classes of traffic into two types: Real Time (RT), corresponding to UGS or rtPS classes, and Best Effort (BE), corresponding to classes nrtPS and BE. Thus the RT class gathers the non delay-tolerant calls.

The calls can change of roamed region on the basis of the Random Waypoint (RWP) model. We use the RWP model to determine the mobility behavior of a call over a convex area. This area (i.e. the cell) is decomposed into several concentric regions. The RWP model helps us to determine the incoming/migrating rates for each region (intra-mobility). Moreover we extends our theoretical results to obtain the handover rate to/from external cell (inter-mobility).

2.3 Connection Admission Control

In Fig. 1, we describe the first CAC algorithm for a new call of class- c . In this first algorithm, the calls receive the same bit rate depending on its type of traffic (without the priority between RT and BE calls). Consequently, the system allocate a number of subcarriers in function of the location of the call. A new class- c call arriving in region i is accepted if the required resources are available for it. Else, the call is blocked. If a call did not finish its service in the region i and moves to the neighbor region $j = i \pm 1$, the call is accepted in region j if the system accepts its modulation changing. The available resources of the system could be not enough to accept a bandwidth increase. In this case, the migrating call is dropped.

The Fig. 2 represents our last CAC algorithm for a new call of class- c . In this other CAC algorithm the BE calls have no bandwidth requirement. Since the BE calls tolerate throughput variation, they will use the sub-carriers left by the RT call occupancy. In fact, the BE calls are never blocked and received the same resources according with the Processor

Sharing (PS) (Benameur et al., 2001). Thus, the bit rate of a BE call depends on its region (i.e modulation). For the RT calls only, the CAC algorithm follows the same scheme than previously: a new RT call arriving in region i is accepted if enough resources are available for it. And since a RT call did not terminate its service in region i before migrating to a region j , $j = i \pm 1$, the call will be able to remain in the system if enough resources are available to afford the modulation change.

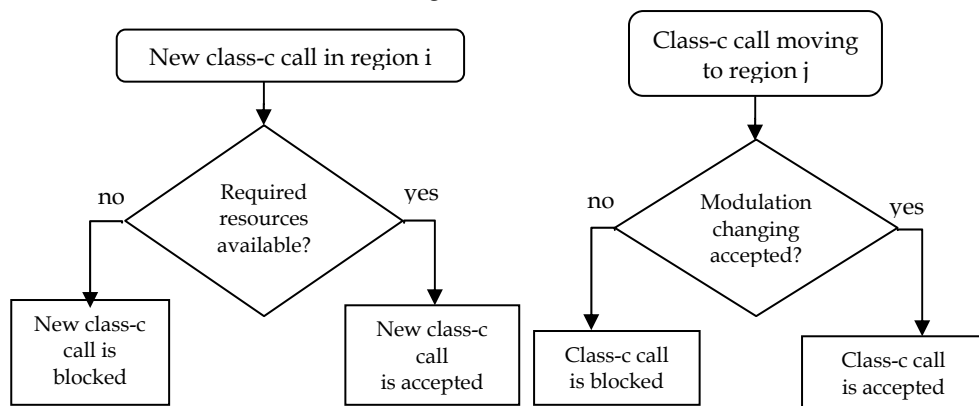


Fig. 1. First CAC Algorithm decision.

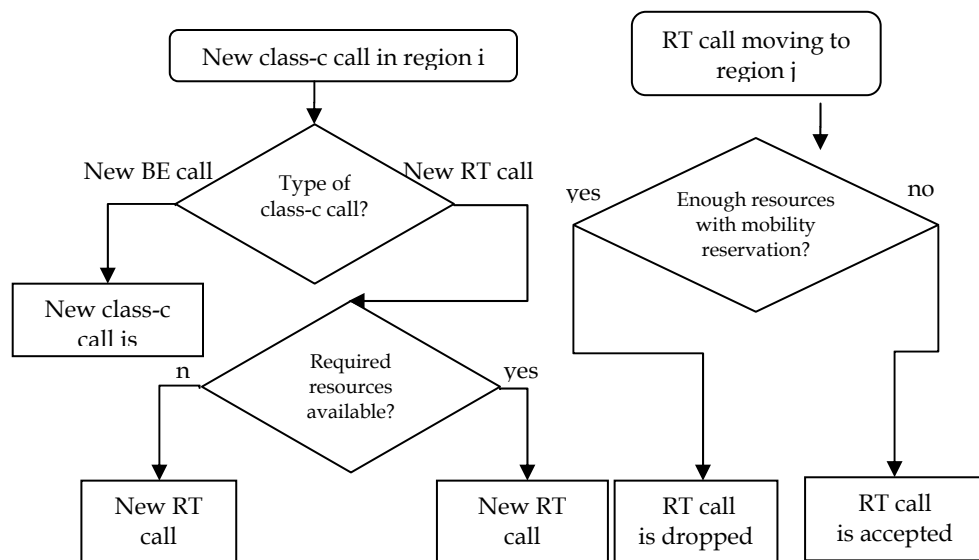


Fig. 2. Second CAC Algorithm decision.

Finally, remarks that for both CAC algorithms, the RT calls are independent of the consume resources: the RT-call remaining-time only depends on the behavior of the user. Conversely the BE calls remain in the system in function of the consumed resources: the more sub-carriers a BE call have, the faster it leaves the system.

In addition, our CAC algorithms seek to reduce the dropping probability: the probability that an on-progress service is dropped due to its mobility. As explained above, the call consumes bandwidth in function of the used modulation. By migrating to an outer region, a call may require additional resources. Thus, this call might undergo a drop due to lack of available resources. To prevent from these drops, our CAC algorithms introduce a reserved part of bandwidth. This reservation aims to satisfy the need of additional resources demanded in case of outer migration.

3. Model

3.1 Cell decomposition and instantaneous throughput

We consider without loss of generality, the Adaptive Modulation and Coding (AMC) with pathloss only. Then, the OFDMA cell is decomposed into r regions according to the AMC value corresponding to a certain value of SNR as depicted in Fig. 3. Let R_i ($i=1, \dots, r$) be the radius of the i -th region and S_i represents the corresponding surface. Each region corresponds to a specific modulation order (see Table 1). In OFDMA scheme, the total number N of sub-carriers is divided into L sub-channels (or groups) each containing k sub-carriers, such $k=N/L$.

In our study, we consider the multi-services WiMAX/OFDMA system with two types of traffics real-time (RT) and best-effort (BE). Also, we define the instantaneous bit rate (radio interface rate) for a call of class- c located in the region i as follows:

$$d_c^i = L_c^i \times k \times B \times e_i, \quad (1)$$

where K is the number of sub-carriers assigned to each sub-channel; B is the baud rate (symbol/sec); e_i is the modulation efficiency (bits/symbol) and L_c^i is the sub-channels allowed for class- c call in region i . The above bit rate can be degraded by the error channel due to collision, shadow fading effect, as defined in (Tarhini & Chahid, 2007),

$$R_c^i = d_c^i \times (1 - BLER_i), \quad (2)$$

where $BLER_i$ is the BLock Error Rate in region i . The Table 1 indicates the modulations and codings used in a IEEE802.16e cell as function of the user SNR. The SNR requirement for a BLER less than 10^{-6} depends on the modulation type as specified in the standard (Standard IEEE802.16, 2004). Then, we have $\gamma_1=24.4$ dB, $\gamma_2=18.2$ dB, $\gamma_3=9.4$ dB, $\gamma_4=6.4$ dB and $\gamma_0=\infty$.

Modulation	Coding rate	Received SNR (dB)	Cell ratio (%)
64-QAM	3/4	$[\gamma_1, \gamma_0)$	1.74
16-QAM	3/4	$[\gamma_2, \gamma_1)$	5.14
QPSK	1/2	$[\gamma_3, \gamma_2)$	20.75
BPSK	1/2	$[\gamma_4, \gamma_3)$	39.4

Table 1. IEEE802.16e AMC settings.

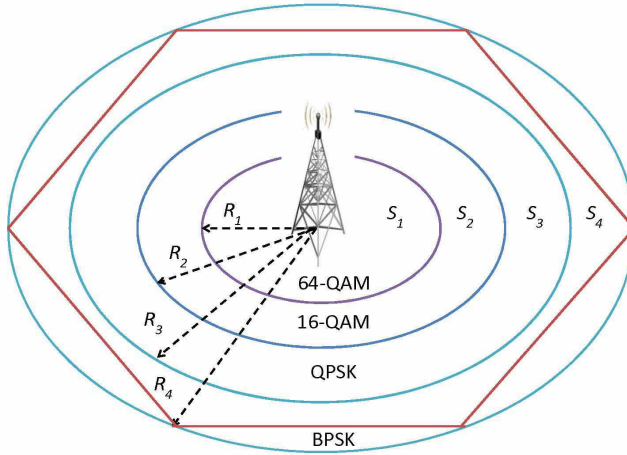


Fig. 3. OFDMA Cell decomposed into concentric regions.

3.2 System state and transitions

Our model of the system is based on the Continuous Time Markov chain (CTMC) technic. The different transition rates within the space of feasible states defined in the next sections are caused by one of the following events: arrival of a new call of class- c to region i ; migration of an ongoing call of class- c from region i to j ; termination of an ongoing call of class- c in the region i . Furthermore, we consider in our analysis the following assumptions:

1. The arrival process of new calls of class- c in region i is Poisson with rate $\lambda_{c,i}^0$;
2. The service time of a class- c call is exponentially distributed with mean $1/\mu_c$;
3. The mean dwell time or sojourn time in region i is exponentially distributed with mean $1/\Gamma_c^i$;
4. The mean arrival rate of migrating call of class- c from the region i to region j is $\lambda_c^{i,j}$.

Let $n_c^i(t)$ be the number of calls of class- c in progress at time t in region i . The state of the system at time t is defined by: $\vec{n}(t) = (n_{RT}^1(t), \dots, n_{RT}^r(t), n_{BE}^1(t), \dots, n_{BE}^r(t))$. Then, we model the process $\{\vec{n}(t), t > 0\}$ as 2r-dimension quasi-birth and death Markov chain. In the steady-state it has an unique stationary distribution, with:

$$\vec{n} = (n_{RT}^1, \dots, n_{RT}^r, n_{BE}^1, \dots, n_{BE}^r).$$

3.3 User mobility behaviour

We compute the arrival migration rates using RWP model. In the RWP model a node moves in a convex domain $\Omega \subset R^2$ along a straight line segment from one waypoint to another. The waypoint, denoted by P_i , are uniformly distributed in Ω , $P_i \sim U(\Omega)$. Transition from P_{i-1} to P_i is referred to as the i -th leg, and the velocity of the node on i -th leg is given by random

variable v_i . In particular in the RWP model, it is assumed that P_i and v_i are all independent and v_i are uniformly distributed. Here, the domain Ω corresponds to one cell and leg to path between both waypoints P_i and P_{i-1} . Also, the node corresponds to a mobile or user moving in the cell. With this notation the RWP process (for a single node) is defined by an infinite sequence of triples (Bettstetter et al., 2004),

$$\{(P_0, P_1, v_1), (P_1, P_2, v_2), \dots\}.$$

We note that the process RWP is time reversible. This means that the arrival rates across any line segment or border are equal in both directions. In other words, the average rates of calls moving from region i to region j per time unit is equal to the number of calls moving from region j to region i per time unit, i. e., $\lambda_c^{i,j} = \lambda_c^{j,i}$ as proved in (Norris, 1999).

Our aim is to compute the migration rates $\lambda_c^{i,j}$. As the velocity of the user (node) is assumed to have an uniform distribution from v_{\min} , ($v_{\min} > 0$) to v_{\max} , denoted by $f_v(v)$, where

$$f_v(v) = \begin{cases} \frac{1}{v_{\max} - v_{\min}}, & \text{if } v \in [v_{\min}, v_{\max}]; \\ 0, & \text{otherwise.} \end{cases}$$

Let T_i denote the transition time on i -th leg (path) defined as $T_i = \frac{l_i}{v_i}$, where

$l_i = |P_i - P_{i-1}|$. The variables l_i and v_i are independent random variables. The average time from one waypoint to another is given by:

$$E[T] = \bar{l} \int_{v_{\min}}^{v_{\max}} \frac{1}{v} f_v(v) dv = \bar{l} \frac{\ln(v_{\max}/v_{\min})}{v_{\max} - v_{\min}} = \bar{l} \cdot E[1/v]. \quad (3)$$

We consider the area A_i of each concentric cell of radius R_i as a convex disk of same radius in which the mobiles move according to the RWP model. Our aim is to compute the arrival rate into a cell of radius R_{i-1} . As introduced in (Hyyti & Virtamo, 2007), let $a_1 = a_1(R_{i-1}, \varphi)$ denote the distance from point $R_{i-1} = |(0, R_{i-1})| \in A_i$ to the border of A_i in direction φ (angle anti-clockwise away from the tangent at point $R_{i-1} = |(0, R_{i-1})|$). $a_2 = a_1(R_{i-1}, \varphi + \pi)$ denotes the distance to the border in the opposite direction (see Fig. 4). Also, we note the specific flux at R_{i-1} in direction φ by

$$\psi(R_{i-1}, \phi) = \frac{1}{2C_v} \cdot a_1 a_2 (a_1 + a_2), \quad (4)$$

where $C_v = \bar{l} A_i^2 \cdot E[1/v]$. We recall that the i -th disk surface is $A_i = \pi R_i^2$ and the mean length of a leg in this disk is

$$\bar{l} = \frac{1}{\pi R_i^2} \int_0^\pi a_1 a_2 (a_1 + a_2) d\phi. \quad (5)$$

According with the RWP model developed in cellular network context (Hyyti & Virtamo, 2007), the arrival rate for one user to region $i-1$ over all contour of disk of the radius R_{i-1} is given by

$$\lambda(R_{i-1}) = 2\pi R_{i-1} \int_0^\pi \sin(\phi) \psi(R_{i-1}, \phi) d\phi. \quad (6)$$

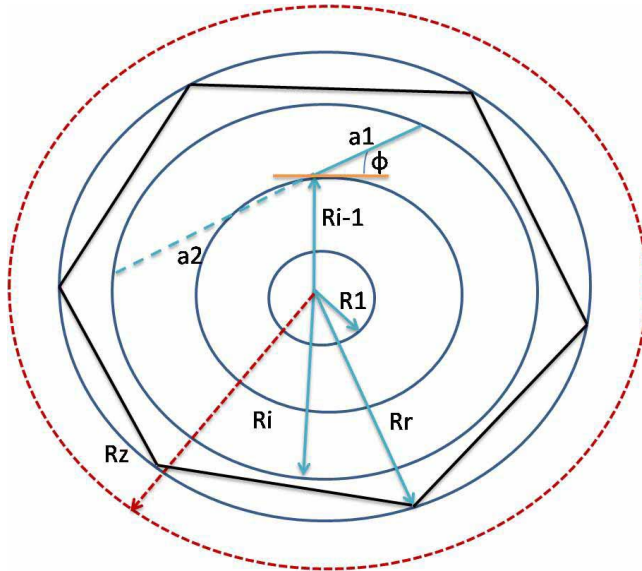


Fig. 4. RWP domains (disk of radius R_i) and $R_z = 2R_r - R_{i-1}$.

From the Fig. 6, we deduce easily the distances a_1 and a_2 as follow:

$$a_1(R_{i-1}, \phi) = \sqrt{R_i^2 - R_{i-1}^2 \cos^2(\phi)} - R_{i-1} \sin(\phi),$$

$$a_2(R_{i-1}, \phi) = \sqrt{R_i^2 - R_{i-1}^2 \cos^2(\phi)} + R_{i-1} \cos(\phi).$$

As the user mobility behaviors are independent, the total migration rate from region i ($i = 1, \dots, r$) to region j ($j = i \pm 1$) of class- c is given by

$$\lambda_c^{i,j} = \lambda(R_j).n_c^i. \quad (7)$$

Finally, the handover arrival rate is given by

$$\lambda_c^{ho} = \lambda_c^{r+1,r} = \lambda(R_r).n_c^r. \quad (8)$$

Now we can compute the mean sojourn time Γ_c^i of one mobile that proceed to class- c call in region i . This mobile can arrives from region j with rate $\lambda_c^{j,i}$ ($i = j \pm 1$) or from outside as new call with rate $\lambda_{c,i}^0$. Let p_c^i be the probability of finding a call of class- c in region i ($i = 1, \dots, r$). Then the mean sojourn time is given by (Hyyti & Virtamo, 2007):

$$\Gamma_c^i = \frac{p_c^i}{\lambda_c^{i+1,i} + \lambda_c^{i-1,i} + \lambda_{c,i}^0}, \quad \text{with } \lambda_c^{0,1} = 0, \quad (9)$$

where

$$p_c^i = P(n_c^i(t) \geq 1) = \sum_{\vec{n}, n_c^i \geq 1} \pi_k(\vec{n}). \quad (10)$$

where π_k is the probability distribution computed in the next system analysis and $k=1, 2$.

4. System analysis for the first CAC algorithm

4.1 Bandwidth occupancy

As described above in the Fig. 1, the interest QoS is to guarantee for a call of class- c a same bit rate independently of its position in the cell. In fact, we allocate to it the needed sub-channels by using equation (2) as

$$L_c^i = \frac{R_c}{k \times B \times e_i \times (1 - BLER_i)}. \quad (11)$$

We recall that the mean call duration of BE calls in the system depends on the transmitting payload in bits, i.e., $\mu_{BE} = \frac{R_{BE}}{E(Pay)}$, where $E(Pay)$ is the mean file size (Downey, 2001). Thus, we define the space of the admissible states as follows

$$E = \{\vec{n} \in N^{2r} \mid \sum_{i=1}^r \{n_{RT}^i L_{RT}^i + n_{BE}^i L_{BE}^i\} \leq L\}. \quad (12)$$

Let L_m be the reserved capacity for migrating or handoff calls of class-c and L_0 denotes the remaining capacity given by $L_0 = L - L_m$. Let $B_1(\vec{n})$ be the occupancy bandwidth when system state is \vec{n} , with

$$B_1(\vec{n}) = \sum_{i=1}^r \{n_{RT}^i L_{RT}^i + n_{BE}^i L_{BE}^i\}. \quad (13)$$

4.2 Equilibrium distribution

The call of class-c can come as new call or migrating/handoff call in region i of the cell. For \vec{n} the current system state, we define the arrival rate of call in region i , as

$$\lambda_c^i(\vec{n}) = \begin{cases} \lambda_{c,i}^0 + \lambda_c^{i-1,i} + \lambda_c^{i+1,i}, & \text{if } B_1(\vec{n}) < L_0; \\ \lambda_c^{i-1,i} + \lambda_c^{i+1,i}, & \text{if } L_0 \leq B_1(\vec{n}) < L. \end{cases} \quad (14)$$

We have two classes of services RT and BE. Each class-c call in region i ($i = 1, 2, \dots, r$) requires the effective bandwidth L_c^i . Then, we have $2r$ classes in the cell and the equilibrium distribution is given by BCMP theorem (Chao et al., 2001) for multiple classes

with possible class changes, with $\rho_c^i = \frac{\lambda_c^i(\vec{n})}{\Gamma_c^i + \mu_c}$ as

$$\pi_1(\vec{n}) = \frac{1}{G} \prod_{i=1}^r \frac{(\rho_{RT}^i)^{n_{RT}^i}}{n_{RT}^i!} \frac{(\rho_{BE}^i)^{n_{BE}^i}}{n_{BE}^i!}, \quad (15)$$

where $\vec{n} \in E$ and G is the normalizing constant given by

$$G = \sum_{n \in E} \prod_{i=1}^r \frac{(\rho_{RT}^i)^{n_{RT}^i}}{n_{RT}^i!} \frac{(\rho_{BE}^i)^{n_{BE}^i}}{n_{BE}^i!}.$$

So, this probability depends of the mean sojourn time Γ_c^i which depends itself on the probability p_c^i in equation (10). Also the latter depends of the above distribution and vice versa. We can use the fixed point theorem to resolve this problem as follow:

Algorithm : Probability convergence algorithm

- 1: Initialize the probability in (9): $p_{c,old}^i = p_c^i = 0.1$.
 - 2: Compute the mean sojourn time in (9).
 - 3: Calculate the steady-state probability $\pi_1(\vec{n})$ from (15).
 - 4: Derive the new value of probability from (10), denoted by $p_{c,new}^i$.
 - 5: Check the convergence of the probability between the old and the new values. if $|p_{c,new}^i - p_{c,old}^i| < \xi$, where ξ is a very small positive number, then the new probability is used to compute the performance metrics. Otherwise, go to step 2 with new value as initial value. The iterations are continued until to reach the convergence of probability.
-

5. System analysis for the second CAC algorithm

5.1 Bit rates per class and feasible system states

Here, we consider the CAC algorithm follows the scheme described in the Fig. 2. Consider there is a minimum capacity reserved for BE calls denoted by L_{BE} and reserved L_{RT}^m sub-channels for RT calls mobility. We denote by L_{RT} the remaining capacity for RT calls given by:

$$L_{RT} = L - L_{BE} - L_{RT}^m. \quad (16)$$

Let $B_2(\vec{n})$ be the bandwidth occupied by RT calls when system sate is \vec{n} , with

$$B_2(\vec{n}) = \sum_{i=1}^r n_{RT}^i L_{RT}^i, \quad (17)$$

where

$$L_{RT}^i = \frac{R_{RT}}{k \times B \times e_i \times (1 - BLER_i)}.$$

In this section, we guarantee for RT calls a QoS in terms to maintain a same bit rate everywhere in the covered area by the cell. Whereas a BE call receives the instantaneous bit rate denoted by R_{BE}^i in region i . The dynamic capacity $C(\vec{n})$ shared fairly among all BE calls simultaneously in progress in the system is

$$C(\vec{n}) = \begin{cases} L - B_2(\vec{n}) - L_{RT}^m, & \text{if } B_2(\vec{n}) < L_{RT}; \\ L_{BE}, & \text{otherwise.} \end{cases} \quad (18)$$

The BE calls share the reserved capacity with PS manner. The number of sub-channels L_{BE}^i allocated to a BE call in region i with PS policy is

$$L_{BE}^i(\vec{n}) = \left\lfloor \frac{C(\vec{n})}{\sum_{i=1}^r n_{BE}^i} \right\rfloor, \quad (19)$$

where $\lfloor x \rfloor$ indicates the largest integer that is less than or equal to x . Then the BE call receives in region i the bit rate $R_{BE}^i(\vec{n}) = L_{BE}^i(\vec{n}) \times k \times B \times e_i \times (1 - BLER_i)$. Therefore, the mean BE call duration is given by $\mu_{BE}^i = \frac{R_{BE}^i(\vec{n})}{E(Pay)}$. Since the system accept without limit the BE calls, the space of admissible states is

$$F = \{\vec{n} \in N^{2r} \mid \sum_{i=1}^r n_{RT}^i L_{RT}^i \leq L_{RT}\}. \quad (20)$$

We define the indication function as

$$\delta(X) = \begin{cases} 1, & \text{if } X \text{ is true;} \\ 0, & \text{otherwise.} \end{cases}$$

5.2 Transition rates

The transition rates from the state \vec{n} to other ones are introduced as described in the sequel.

Let \vec{n}_{i+}^{-c} be the state when a new class- c call is arrived and we denote this transition by $q_{(\vec{n}, \vec{n}_{i+}^{-c})}^{-c}$. Let \vec{n}_{i-}^{-c} be the state when a class- c call in region i terminates its service or changes

its modulation order and we denote this transition by $q_{(\vec{n}, \vec{n}_{i,j}^c)}^{\vec{c}}$. Let $\vec{n}_{i,j}^c$ ($j = i \pm 1$) be the state when a class- c call in region i moves to the neighbor region j and we denote this transition by $q_{(\vec{n}, \vec{n}_{i,j}^c)}^{\vec{c}}$. Then, we have

$$\begin{aligned} q_{(\vec{n}, \vec{n}_{i+}^c)}^{\vec{c}} &= \delta(B_2(\vec{n}) + L_{RT}^i \leq L_{RT}) \lambda_{RT,i}^0, \\ q_{(\vec{n}, \vec{n}_{i+}^c)}^{\vec{c}} &= \lambda_{BE,i}^0, \\ q_{(\vec{n}, \vec{n}_{i,j}^c)}^{\vec{c}} &= \delta(B_2(\vec{n}) + \Delta_{RT}^{i,j} \leq L_{RT} + L_{RT}^m) \lambda_{RT}^{i,j}, \\ q_{(\vec{n}, \vec{n}_{i,j}^c)}^{\vec{c}} &= \lambda_{BE}^{i,j}, \\ q_{(\vec{n}, \vec{n}_{i-}^c)}^{\vec{c}} &= n_{RT}^i (\mu_{RT} + \Gamma_{RT}^i), \\ q_{(\vec{n}, \vec{n}_{i-}^c)}^{\vec{c}} &= n_{BE}^i (\mu_{BE}^i(\vec{n}) + \Gamma_{BE}^i), \end{aligned}$$

where Γ_c^i is computed in Algorithm 1 by replacing the the probability π_1 by π_2 , and $\Delta_{RT}^{i,j} = L_{RT}^i - L_{RT}^j$.

Let Q be the matrix of the possible transitions, where $Q = (q_{(\vec{n}, \vec{n}')}^{\vec{c}})$ for $\vec{n} \in F$ and $\vec{n}' \in F$. The transition rate from state \vec{n} to \vec{n}' is denoted by $q_{(\vec{n}, \vec{n}')}^{\vec{c}}$. Its value must be obtained as the sum of all terms in each line in matrix Q is equal to zero for RT call and BE one as well as $i = 1, \dots, r$.

5.3 Steady-state distribution

Now, we recall that $\pi_2(\vec{n})$ denotes the steady-state probability when system is in the state \vec{n} ($\vec{n} \in F$) and by $\vec{\pi}$ the steady-state distribution vector, where $\vec{\pi} = \{\pi_2(\vec{n}) \mid \vec{n} \in F\}$. The steady-state probability vector is solution of the following system of equations:

$$\vec{\pi} Q = \vec{0}, \quad (21)$$

$$\vec{\pi} \vec{1} = 1. \quad (22)$$

where $\vec{1}$ is a column vector of ones and $\vec{0}$ is a row vector of zeros.

6. Performances metric

Once the equilibrium distribution probabilities 4.2 and 5.3 are calculated, we compute many interesting metrics of the system. In this section, we provide explicit expressions for various metrics like dropping probabilities, blocking probabilities, average sojourn time and average throughput.

6.1 First scheme

6.1.1 Blocking probabilities

A new call of class- c in region i is blocked with probability:

$$B_c^i = \sum_{\vec{n} \in E_c^i} \pi_1(\vec{n}), i = 1, \dots, r, \quad (23)$$

where $E_c^i = \{\vec{n} \in E \mid B_1(\vec{n}) + L_c^i > L_0\}$.

6.1.2 Dropping probabilities

Migrating call dropping probability in region i . The migrating call of class- c from region i to region j is dropped with probability

$$D_c^i = \sum_{\vec{n} \in E_c^{i,j}} \pi_1(\vec{n}), i = 2, \dots, r, \quad (24)$$

where $E_c^{i,j} = \{\vec{n} \in E \mid B_1(\vec{n}) + L_c^j - L_c^i > L\}$.

6.1.3 Average throughput

The average throughput in the system is

$$Th_1 = \sum_{\vec{n} \in E} \pi_1(\vec{n}) \sum_{i=1}^r (n_{RT}^i R_{RT} + n_{BE}^i R_{BE}). \quad (25)$$

6.2 Second scheme

6.2.1 Blocking probabilities

A new call of class-RT in region i is blocked with probability

$$B_{RT}^i = \sum_{\vec{n} \in F_{RT}^i} \pi_2(\vec{n}), i = 1, \dots, r, \quad (26)$$

where $F_{RT}^i = \{\vec{n} \in F \mid B_2(\vec{n}) + L_{RT}^i > L_{RT}\}$.

6.2.2 Dropping probabilities

Migrating call dropping probability in region i . The migrating call of class-RT from region i to region j is dropped with probability

$$D_{RT}^i = \sum_{\vec{n} \in F_{RT}^{i,j}} \pi_2(\vec{n}), i = 2, \dots, r, \quad (27)$$

where $F_{RT}^{i,j} = \{\vec{n} \in F \mid B_2(\vec{n}) + \Delta_{RT}^{j,i} > L_{RT} + L_{RT}^m\}$.

6.2.3 Average throughput

The average throughput in the system is

$$Th_2 = \sum_{n \in F} \pi_2(\vec{n}) \sum_{i=1}^r (n_{RT}^i R_{RT} + n_{BE}^i R_{BE}^i(\vec{n})). \quad (28)$$

7. Numerical applications

The following parameters and assumptions are used in our numerical applications. Considering OFDMA cell system with an FFT (*fast Fourier transform*) size of 2048 sub-carriers. The cell is decomposed into two regions ($r = 2$), $R_1 = 300$ m and $R_2 = 600$ m with AMC scheme: 16-QAM 3/4 ($e_1 = 3$ bits/symbol) and QPSK 1/2 ($e_2 = 1$ bit/symbol). The baud value $B = 2666$ symbols/sec, $BLER_i = 0$, and $K = 48$. These parameters correspond to the transmission modes with conventionally coded modulation (Liu & Zhou, 2005); The bit rate R_{RT} is equal to 128 Kbps and R_{BE} is 384 Kbps (Tarhini & Chahed, 2007); The total bandwidth L is 10 sub-channels; The mean call duration for RT calls is equal to 120 sec and the download of files with mean size for BE calls is equal to $E(Pay) = 5$ Mbits. We assume a mobile moves according to the RWP model on a convex disk of radius $R_z = 900$ m. It randomly chooses a new speed in each waypoint from an uniform distribution between $[v_{min}, v_{max}]$, where $v_{min} = 3$ km/h (low mobility) or $v_{min} = 20$ km/h (high mobility), $v_{max} = 90$ km/h.

7.1 Impact of first scheme

The Fig. 5 presents the blocking probabilities for each burst profile (i.e. modulation) in terms of reserved resources for mobility. As expected, the probabilities increase as the reserved threshold L_m increases and the modulation efficiency decreases. Moreover, an appreciable difference exists between class RT and class BE blocking probability. This is due to the required capacity per class type, the BE class call in our numerical environment requires more bandwidth than the RT class call. So, when threshold increases, the blocking probability increases due to the CAC mechanism which gives the priority to migrating/handoff call than the new call. So the blocking probability mainly depends on the required bandwidth associated with the call modulation efficiency. We also observe that as

the reserved bandwidth L_m increases, the calls are more blocked as the incoming region is away from the base station and as the calls demand an high bandwidth. In particular, we observe on the figure the blocking probabilities for two type of arrivals : BE calls in 16QAM and RT calls in QPSK. The blocking probabilities are exactly the same because the product between the required bandwidth and the modulation efficiency are the same for both.

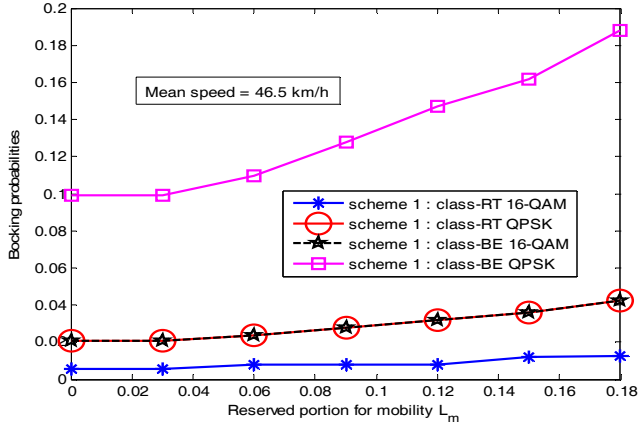


Fig. 5. Blocking probabilities versus threshold L_m and mean speed for $\lambda_{RT,i}^0 = \lambda_{BE,i}^0 = 0.3$ call/sec.

The Fig. 6 shows the average throughput of the whole cell versus the reserved threshold mobility. On this figure two singular behaviors have to be studied. The first main observation deals with the throughput increasing because of an higher mobility. As a mobile moves faster, it increases its probability to change region (and thus modulation) per unit of time. This fact implies also more calls dropped due to lack of resources. In fact the system will implicitly coerce the system calls to use better modulations. This last deduction explains how an higher mobility allows to reach a better throughput. Note that this remark is confirmed in (Peyre & Elazouzi, 2009) and also observed for the ad-hoc networks (Grossglaue & Tse, 2002). Moreover, we can criticize on the Fig. 6, the impact of resource reservation to ease the call mobility. For any mobility behavior, the average cell throughput decreases as the reservation share increases. But the throughput fall depends on the mobility behavior. In fact, by introducing a resource reservation, we helps more and more migrating calls to remain in the border regions. Theses calls use more subchannels to reach the same bit rate than previously. Consequently, the system reaches a lower throughput by prioritizing the user mobility management.

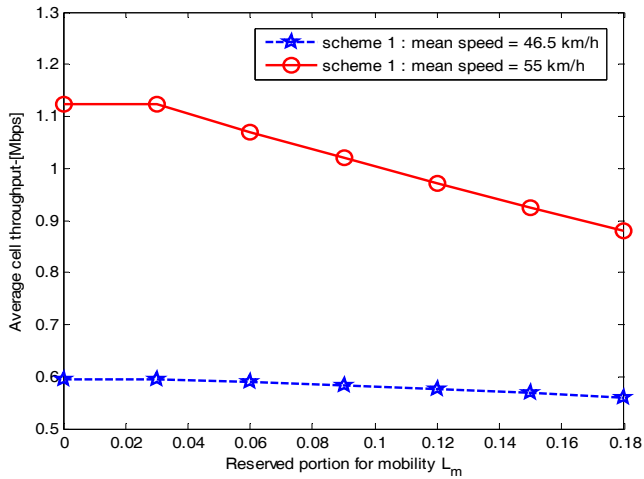


Fig. 6. Average throughput versus threshold L_m and mean speed for $\lambda_{RT,i}^0 = \lambda_{BE,i}^0 = 0.3$ call/sec.

The Fig. 7 represents the dropping probabilities versus reserved threshold L_m . We plot the results for both types of traffic in the border region (i.e. QPSK modulation), and for two mobility behaviors. The figure helps to appreciate the great impact of the resource reservation on the mobility management efficiency. Concerning the real-time traffics, we observe that an higher mobility causes an twenty-times dropping increase. To fight against this effect, our CAC algorithm permits to tune the resource reservation in order to decreases the call drops under a desired value. For examples, a reservation lower than ten percents of the total bandwidth decreases the number of dropped calls under a one-percent probability. For the Best Effort traffic, we observe exactly the same behavior. Nevertheless, the dropping probability for the BE traffic remains higher than for the RT because the BE calls require an higher bit rate. As they ask for more bandwidth, they undergo more drops. For this type of traffic, our CAC allows also to greatly reduce the dropping probability of the BE calls by increasing the resources reservation L_m .

7.2 Impact of second scheme

The Fig. 8 shows the blocking probabilities for the RT traffics in both regions versus the bandwidth reservation for the Best Effort call. We plot the results obtained for two resource reservation profiles for the user mobility management. This figure clearly shows the impact

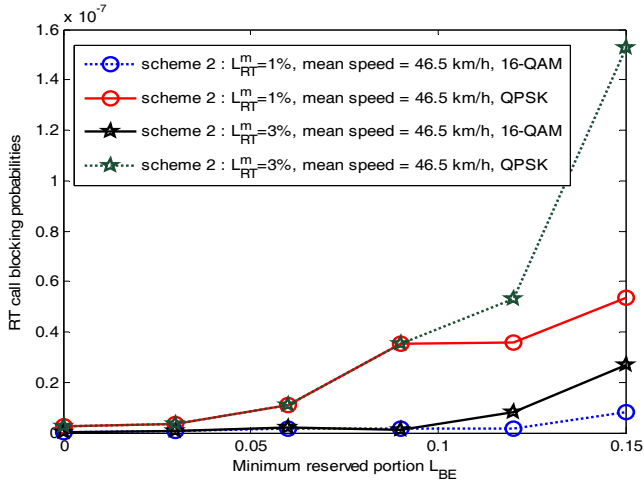


Fig. 7. Blocking probabilities versus threshold L_m and mean speed for $\lambda_{RT,i}^0 = \lambda_{BE,i}^0 = 0.3$ call/sec.

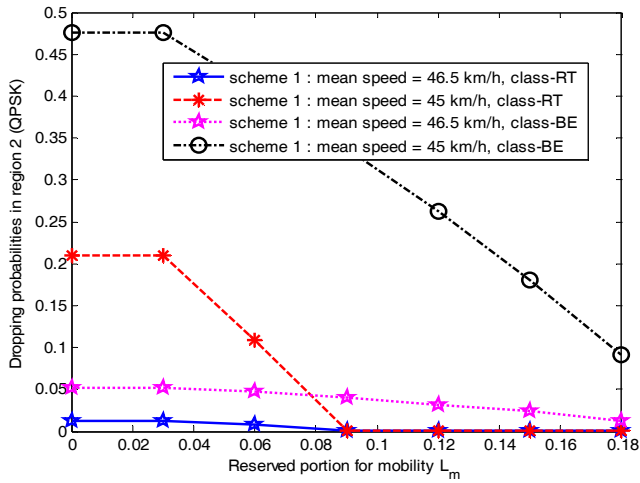


Fig. 8. Dropping probabilities versus threshold L_{BE} and L_{RT}^m for $\lambda_{RT,i}^0 = \lambda_{BE,i}^0 = 0.3$ call/sec.

of L_{BE} and L_{RT}^m on the blocking probabilities experienced in each region. In general, the blocking probabilities are heavily increased by the BE bandwidth reservation. In addition, this drawback is amplified by the increase of the mobility management resource reservation.

From this statement, we could compute the the value ranges for L_{BE} and L_{RT}^m which satisfy a maximum blocking probability threshold.

The Fig. 9 shows the dropping probabilities for the RT traffics in the border region versus the bandwidth reservation for the Best Effort call. We plot the results obtained for two resource reservation profiles for the user mobility management. The figure shows how the resource reservation L_{RT}^m fights against the dropping due to the BE bandwidth reservation. On this figure, the dropping probability is reduced as the BE reservation is greater than ten percents. So we can determine the possible values for L_{RT}^m from a desired maximum dropping probability.

The Fig. 10 shows the total cell throughput. This figure presents the impact of the different bandwidth reservation to increase the BE call throughput and to ease the call mobility. On the Fig. 10 we can appreciate the great impact of the BE bandwidth reservation L_{BE} .

Indeed, by reservation 15% of the total bandwidth, we double the total cell throughput. In addition, to increase the mobility-purpose reservation slightly decrease the total throughput. This observation leads to think that a small quantity of subchannels for Best effort calls allows to reach very higher throughput. In fact, this by reserving few subchannels to the BE calls, we also increase the blocking and dropping probabilities for the RT calls. Therefore, these amount of resources freed by the dropped call or let free by the blocking one allow to the BE calls to use even more resources.

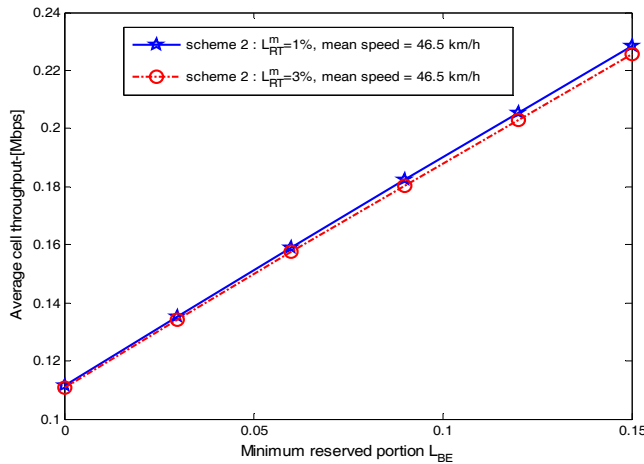


Fig. 9. Mean throughput versus L_{BE} and L_m for $\lambda_{RT,i}^0 = \lambda_{BE,i}^0 = 0.3$ call/sec.

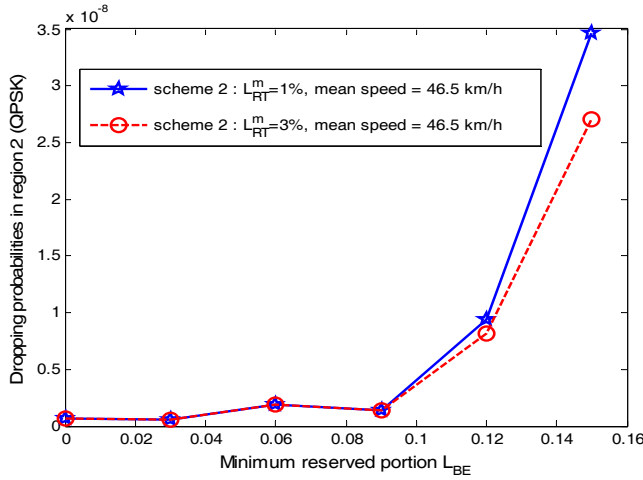


Fig. 10. RT dropping probabilities versus threshold L_{BE} and L_{RT}^m for $\lambda_{RT,i}^0 = \lambda_{BE,i}^0 = 0.3$ call/sec.

8. Conclusion

Recently many works have been introduced to study the WiMAX performances in order to improve some QoS for users. In this sense our work contributes in order to improve the QoS of users. In fact, by considering both traffics RT and BE, we proposed two strategies of QoS management. The first defines constant bit rates (CBR) for RT and BE calls. We also introduce a resource reservation to ease the mobility of the users. The second scheme replaces the CBR policy of the BE calls by a Processor Sharing (PS), and we add an other bandwidth reservation to improve the minimum Best Effort call throughput. Moreover, we define a realistic mobility model through the accurate Random Waypoint model. Based on these propositions we develop a continuous time Markov chain which determines the steady state of the system. From the model, we provide a large range of performance metrics. We conclude our analysis by criticize the impact of each CAC algorithm parameters on the system performances. By gathering all our results, we can easily choose one of the two proposed CAC algorithms to meet with the desired traffic prioritization policy. In addition, we analyzed the possible ways to tune the parameters of each CAC algorithms in order to specify the main thresholds (average throughput, blocking and dropping probabilities). As future works we seek to introduce thinking times in our RWP model. Our main objective is also to improve the CAC algorithms by determining the best way to roam the mobile user to a region, in function of its speed and the expected time spend in the next region.

9. References

- Benameur, N.; Ben Fredj, S.; Delcoigne, F.; Oueslati-boulahia, F.; Roberts, J. W.; & Moulineaux, I. L. (2001). Integrated admission control for streaming and elastic traffic. *Journal Lecture Notes in Computer Science*, Vol. 2156, (February, 2001), pp. 69-82, ISSN 0302-9347.
- Bettstetter, C.; Hartenstein, H. & Perez-Costa, X. (2004). Stochastic properties of the random waypoint mobility model. *ACM/KluwerWireless Networks: Special Issue on Modeling and Analysis of Mobile Networks*, Vol. 10, No. 5, (September, 2004).
- Chao, X.; Miyazawa, M.; Pinedo, M. & Atkinson, B (2001). Queuing networks: Customers, signals and product form solutions. *The Journal of the Operational Research Society*, Vol. 52, No. 5, (May, 2001), pages number (600-601).
- Downey, A.B. (2001). The structural cause of file size distributions. *IGMETRICS/Performance*, Vol. 29, No. 1, (June, 2001), pages number (328-329), ISSN 0163-5999.
- Grossglauser, M. & Tse, D. (2002). Mobility increases the capacity of ad-hoc wireless networks. *IEEE/ACM Transactions on Networking*, Vol. 10, (August, 2002) pages number (477-486), ISSN 1063-6692.
- Hytti, E. & Virtamo, J. (2007). Random waypoint mobility model in cellular networks. *Wireless Network*, Vol. 13, No. 2, (Avril, 2007), pages number (177-188), ISSN 1022-0038.
- Ibrahimi, K.; El-Azouzi, R.; & Bouyakhf, E.H. (2008). Uplink Call Admission Control in Multi-services W-CDMA Network. *Proceedings of 13th IEEE Symposium on Computers and Communications (ISCC'08)*, July 2008, Marrakech, Morocco.
- Ibrahimi, K.; El-Azouzi, R.; S.K. Samanta; & Bouyakhf, E.H. (2009). Adaptive Modulation and Coding scheme with intra- and inter-cell mobility for HSDPA system. To appear in the Sixth International Conference on Broadband Communications, Networks, and Systems (IEEE/ICST BROADNETS), Septembre 2009, Madrid, Spain.
- Johnson, D.B. & Maltz, D.A. (1996). Dynamic source routing in ad hoc wireless networks. In *the book Mobile Computing*, pages number (153-181). Kluwer Academic Publishers.
- Kobbane, A.; El-Azouzi, R.; Ibrahimi, K. & Bouyakhf, E.H (2007). Capacity Evaluation of multi-service W-CDMA : A Spectral Analysis Approach. *Proceedings of the Sixth IASTED International Conference on Communication Systems and Networks, CSN*, August 2007, Palma of Mallorca, Spain.
- Liu, G.O. & Zhou, Z. (2005). Queuing with adaptive modulation and coding over wireless links: Cross-layer analysis and design. *IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS*, Vol. 4, No. 3, (May 2005).
- Niyato, D. & Hossain, E. (2005). Call admission control for qos provisioning in 4G wireless networks: issues and approaches. *IEEE Network*, Vol. 19, No. 5, (2005) pages number (5-11), ISSN 0890-8044.
- Norris, J.R. (1999). *Markov chains*. Cambridge University Press, ISBN = 0-521-68181-3.
- Peyre, T. & El-Azouzi, R. (2009). IEEE802.16 multi-class capacity including AMC scheme, mobility and QoS differentiation for initial and bandwidth request ranging. *To appear in the proceedings of WCNC*, Hungary, April 2009, Budapest.
- Peyre, T.; Ibrahimi, K. & El-Azouzi, R. (2008). IEEE802.16 multi-class capacity including AMC scheme and QoS differentiation for initial and bandwidth request ranging. *Proceedings of Valuetools*, Greece, October 2008, Athens.

- Piggin, P. & Depth, W. (2004). Broadband wireless access. *IEEE Communications Engineer*, Vol. 2, Issue 5, (Octobre 2004), pages number (36-39).
- Seo, H.H.; Ryu, B.H.; Hwang, E.S.; Cho, C.H. & Lee, N.W. (2004). A study of code partitioning scheme of efficient random access in OFDMA-CDMA ranging subsystem. *Proceedings of JCCI*, pages 262, April 2004.
- Standard IEEE802.16. (2004). *IEEE standard for local and metropolitan area networks, part 16: Air interface for fixed broadband wireless access systems*, New York, USA.
- Tarhini, C. & Chahed, T. (2007). On capacity of OFDMA-based IEEE802.16 WiMAX including adaptative modulation and coding and inter-cell interference. *Proceedings of the 15th IEEE Workshop on Local and Metropolitan Area Network (LANMAN)*, Princeton NJ, June 2007, USA.
- Yaghoobi, H. (2004). Scalable OFDMA physical layer in IEEE 802.16 Wirelessman. *Intel Technology Journal*, Vol. 8, (August 2004), pages number (201-212).

Queueing-Model-Based Analysis for IEEE802.11 Wireless LANs with Non-Saturated Nodes

Shigeo Shioda and Mayumi Komatsu
Chiba University
Japan

1. Introduction

The IEEE 802.11 protocol has gained widespread popularity as a standard MAC-layer protocol for wireless local area networks (WLANs). The IEEE 802.11 standard defines Distributed Coordination Function (DCF) as a contention-based MAC mechanism, but it does not have quality-of-service (QoS) functionality. The IEEE 802.11 standard group has approved the 802.11e standard for MAC layer QoS enhancements to the former 802.11 standard, where the Enhanced Distributed Channel Access (EDCA) function of 802.11e is a QoS enhancement of the DCF.

While the IEEE 802.11e claims to support the QoS, several challenging problems still remain on the support of real-time applications with strict QoS requirements. Under the DCF, the real-time bidirectional applications like Voice over IP (VoIP) cannot efficiently utilize the bandwidth of WLANs. The inefficient bandwidth utilization is mainly caused by the uplink/downlink unfairness problem in WLANs (Cai et al., 2006). The DCF assigns the same number of access opportunities to each individual mobile terminal as well as the access point (AP), but each mobile terminal serves one uplink flow while the AP needs to serve all downlink flows. Thus, a downlink flow necessarily gets comparatively lower bandwidth than an uplink flow gets. The unfairness between uplink and downlink flows likely builds up the queue at the access point (AP) and causes packet loss of downlink flows even at moderate load, where unused bandwidth still remains for uplink flows in a WLAN. This implies that, under the use of bidirectional applications, the AP is likely to become performance bottleneck over the standard WLANs. Note that the occupancy of the AP buffer strongly depends on the throughput of the WLAN, at which rate the AP can successfully transfer frames over the WLAN. Thus, the performance of bidirectional applications over WLANs needs to be analyzed by taking into account both the occupancy of the AP buffer and the throughput of the IEEE 802.11 DCF (or EDCA).

In this article, we present a mathematical model for evaluating the MAC-layer performance such as per-flow throughput as well as the network-layer performance such as packet loss at each station. Our proposal combines a Markov chain model for evaluating the throughput of the IEEE 802.11 DCF and a queueing model for analyzing the network-layer performance of each station. The Markov chain model used in our proposal is primarily based on the model by Malone (Malone et al., 2007), which allows us to analyze the throughput of IEEE 802.11 under unsaturated nodes, but we have made an important extension of their model in order to consider the effect that arriving IP packets are queued in the buffer of a station when the

station has frames to transmit. For analyzing the network-layer performance, we apply the GI/M/1 model, where the service time corresponds to the MAC-layer packet service time, which is the time interval between the instant that a packet reaches the head of the queue and the instant that the packet is successfully transferred. It was shown in (Zhai et al., 2004) that the exponential distribution is a good approximation for the MAC-layer packet service time, the mean of which can be evaluated through the analysis of the IEEE 802.11 DCF.

Through extensive simulations using the simulator ns2, we show that our model can accurately predict how many VoIP conversations can be multiplexed over a WLAN without loss of any packets. Our model allows us to evaluate the IEEE 802.11 DCF with contention window (CW) differentiation, which is a service differentiation scheme provided by EDCA. By using this feature of our proposal, in this article, we also investigate how much the CW differentiation could improve the bandwidth utilization by making contention window of the AP smaller than mobile terminals.

This article is organized as follows: in Section 2 we present review on related work. In Section 3, we propose an analytical model to evaluate the performance of the IEEE 802.11 DCF under non-saturated conditions. In Section 4, we present a queueing model to analyze the queueing delay and packet loss ratio at the buffer of the AP or mobile terminals. In Section 5, we show the results of simulation experiments to show the accuracy of the proposed model. In Section 6, we conclude the article with a few remarks.

2. Related Work

The performance of the IEEE802.11 has been widely studied in the literature. Bianchi (Bianchi, 2000) proposed a two-dimensional Markov chain model to analyze the performance of the IEEE 802.11 DCF under the so-called *saturation* condition, in which all stations always have data to send. Robinson *et al.* (Robinson & T.S.Randhawa, 2004) and Xiao (Xiao, 2005) extended Bianchi's DCF model to analyze the performance of the EDCA function of IEEE802.11e under the saturation condition.

Since persistent saturation continues only during a short time period in actual operation, it is important to evaluate the performance of IEEE 802.11 under non-saturation conditions. Ergen *et al.* (Ergen & Varaiya, 2005) proposed an extension of the Bianchi's DCF model by introducing additional states to the Bianchi's Markov chain to represent idle states of a station. Malone *et al.* (Malone et al., 2007) developed a different extension of the Bianchi's DCF model; their model allows stations to have different packet-arrival rates. Daneshgaran *et al.* (Daneshgaran et al., 2008) proposed an analytical model for non-saturated conditions in order to account for packet transmission failures due to errors caused by propagation through the channel. Foh *et al.* (Foh et al., 2007) proposes to use a queueing model to evaluate the performance of IEEE 802.11 under non-saturated conditions. In their queueing model, customers in the system represent active stations, where being "active" means having frames to send. The Zhao *et al.* (Zhao et al., 2008) proposed approximating the attempt rate, at which a station attempts to send a frame, in a non-saturated setting by scaling the attempt rate of saturated setting with the probability that a packet arrives.

As we have explained, in the use of bidirectional applications, packets are likely to be delayed and dropped in the buffer of the AP. The queueing delay and the packet loss in the buffer of the AP would largely affect the performance of real-time applications. All of the studies mentioned in the above, however, could not analyze the queueing delay and the packet loss at the buffer of the AP or each station. Several proposals have been made to conduct cross-layer analysis where the performance of the network layer such as queueing delay or packet

loss at stations is jointly evaluated with the MAC-layer performance such as the throughput (Cheng et al., 2007; Tickoo & Sikdar, 2004; Xiang et al., 2007; Zhai et al., 2004). For example, Zhai *et al.* (Zhai et al., 2004) integrated the Bianchi's Markov-Chain with a queueing model. Tickoo *et al.* (Tickoo & Sikdar, 2004) proposed a similar model where a simplified Bianchi's model was used. The proposal by Xiang *et al.* (Xiang et al., 2007) corresponds to the extension of the Zhai's model to non-saturated conditions. The existing proposals concerning the cross-layer analysis approximate the Bianchi's Markov-chain by a simplified model. Our analytical model, which is categorized into the cross-layer analysis, attempts to directly integrate Bianchi's (or Malone's) Markov-chain with the queueing model.

3. Model of Non-saturated Stations

In this section, we present a bi-dimensional Markov model for evaluating the performance of IEEE 802.11 DCF under non-saturated conditions. We represent the state of each station by a pair of integers $(s(t), b(t))$, where $s(t)$ and $b(t)$ respectively denote the back-off stage and counter of a given station (say station A) at time t . We also let $\{t_1, t_2, \dots\}$ denote state transition instants of station A. Note that $\{(s(t), b(t)), t \geq 0\}$ is not a continuous-time Markov process because the inter-state-transition time is not exponentially distributed. The state at state-transition instants $\{(s(t_n), b(t_n)), n \geq 1\}$, however, would define a Markov chain, where $\{t_n\}_{n \in \mathbb{N}}$ form imbedded Markovian points. In the following, we focus on the state transitions on imbedded Markovian points and simply represent the state of a station by (s, b) , omitting the time parameter t .

3.1 Per-station Markov Model

Assume that there are n stations (one access point and $n - 1$ terminals) in the system. The back-off stage starts at 0 at the first attempt to transmit a packet and increases by 1 every time a transmission attempt results in a collision up to the maximum value. We denote the maximum back-off stage of station l ($l = 1, \dots, n$) by m_l . The maximum back-off stage is related to CW_{max} through $2^{m_l} W_0 = CW_{max} + 1$ where $W_0 = CW_{min} + 1$. The probability that a transmission attempt of station l results in a collision is assumed to be p_l . The back-off stage is reset at 0 after a successful transmission. At the back-off stage s , the back-off counter is initially chosen uniformly between $[0, W_s - 1]$, where $W_s = 2^s W_0$. The counter decreases by one at the start of every time slot when the medium is sensed idle. Note that the back-off counter is suspended when the medium is busy due to the transmission (or collision) by other stations. When the back-off counter reaches zero, the station attempts to transmit a frame at the start of the next time slot.

When the back-off stage of station l reaches the maximum value m_l , it remains m_l even if the station consecutively fails to send frames. Note that the frame is discarded and the back-off stage is reset at 0 when the number of consecutive-frame-retransmission exceeds the retry limit. In this article, however, we do not consider the influence on the frame discard due to consecutive transmission failures because the frame discard due to the consecutive retransmission failures rarely occur in usual cases. This simplification was also used in Bianchi (Bianchi, 2000) and Malone (Malone et al., 2007).

In non-saturated conditions, a station may not have a frame to transmit just after transmitting a frame and resetting the back-off stage and timer. In this paper, such a station is referred to as being "post-backoff". As used in Malone (Malone et al., 2007), we introduce notation $(0, k)_e$ for $k \in [0, W_0 - 1]$ to represent a post-backoff station with back-off timer k . A station in state $(0, k)_e$ makes a transition into $(0, k - 1)$ at the start of the next time slot if (at least) one frame

has arrived during the current time slot; otherwise it enters $(0, k-1)_e$. We assume that the transition probability from state $(0, k)_e$ to state $(0, k-1)$ of station l is q_l .

A station in state $(k, 0)$ ($0 \leq k \leq m_l$) attempts to transmit a frame at the beginning of the next time slot. In the case of a successful transmission, it makes a transition into one of post-backoff states $((0, k)_e, k = 0, \dots, W_0 - 1)$ with probability $1 - r_l$, and it makes a transition into one of backoff states with stage 0 $((0, k), k = 0, \dots, W_0 - 1)$ with probability r_l . In the case of a collision, it enters one of states with back-off stage $k+1$ (when $0 \leq k < m_l$) or m_l (when $k = m_l$). More precisely,

$$\begin{aligned} P[(0, l)_e | (k, 0)] &= (1 - r_l)(1 - p_l) / W_0, \\ P[(0, l) | (k, 0)] &= r_l(1 - p_l) / W_0, \\ P[(k+1, l) | (k, 0)] &= r_l(1 - p_l) / W_{k+1}, \quad \text{for } 0 \leq k < m_l \\ P[(m, 0) | (m, 0)] &= r_l(1 - p_l) / W_{m_l}. \end{aligned} \quad (1)$$

Parameter r_l is the probability that station l has at least one frame after frame transmission. If the back-off counter of the station in post-backoff state reaches 0 but it has no frame, it remains in post-backoff state $(0, 0)_e$. A station in state $(0, 0)_e$ receives at least one frame with probability q_l during the current time slot. If it receives at least one frame during the current time slot and the medium is sensed idle, it attempts to transmit a frame at the start of the next time slot. In the case of a successful transmission, it makes a transition into one of post-backoff states $((0, k)_e, k = 0, \dots, W_0 - 1)$ with probability $1 - q_l$, and it makes a transition into one of backoff states with stage 0 $((0, k), k = 0, \dots, W_0 - 1)$ with probability q_l . In the case of a collision, it enters one of states with back-off stage 1. If a station in state $(0, 0)_e$ receives a frame during the current time slot but the medium is sensed busy at the start of the next time slot, it enters one of backoff-states with stage 0. More precisely,

$$\begin{aligned} P[(0, 0)_e | (0, 0)_e] &= 1 - q_l + \frac{q_l(1 - p_l)P_{idle}}{W_0}, \\ P[(0, k)_e | (0, 0)_e] &= q_l(1 - p_l)P_{idle} / W_0, \quad \text{for } k > 0 \\ P[(0, k) | (0, 0)_e] &= q_l(1 - P_{idle}) / W_0, \quad \text{for } k \geq 0 \\ P[(1, k) | (0, 0)_e] &= q_l p_l P_{idle} / W_1, \quad \text{for } k \geq 0. \end{aligned}$$

3.2 Analysis of the Markov Chain

Figure 1 shows the state transition diagram of the Markov chain. Fortunately, the stationary distribution of the Markov chain can be analytically obtained (see Appendix A). To show this, let $b(i, k)$ denote the stationary probability of being in state (i, k) , and let $b(i, k)_e$ denote the stationary probability of being in $(i, k)_e$. We can show that the stationary distribution of the state $(0, 0)_e$, $b(0, 0)_e$, is given through the following equation:

$$\begin{aligned} 1/b(0, 0)_e &= 1 - q_l + \frac{q_l(W_0 + 1)}{2(1 - r_l)} \left(\frac{q_l W_0}{1 - (1 - q_l)W_0} \right. \\ &\quad \left. + (1 - P_{idle})(1 - r_l) - r_l P_{idle}(1 - p_l) \right) \\ &\quad + \frac{p_l q_l^2}{2(1 - p_l)(1 - r_l)} \left(\frac{W_0}{1 - (1 - q_l)W_0} - \frac{P_{idle}(1 - p_l)r_l}{q_l} \right) \\ &\quad \times \left\{ 1 + 2W_0 \frac{1 - p_l - p_l(2p_l)^{m-1}}{1 - 2p_l} \right\}, \end{aligned} \quad (2)$$

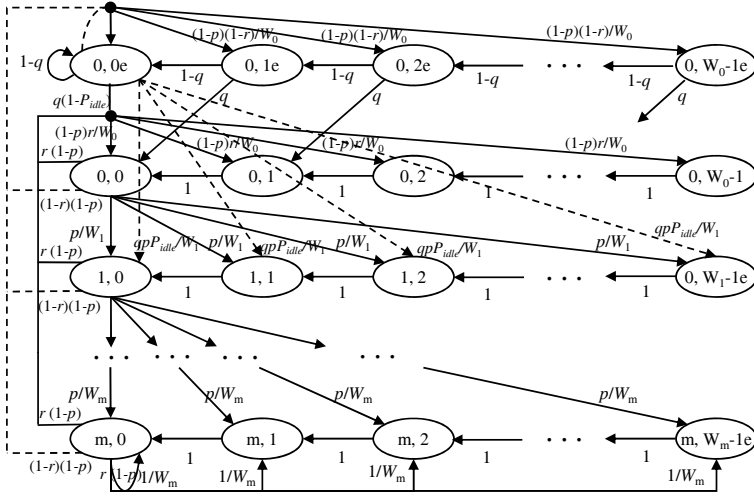


Fig. 1. State transition diagram.

where P_{idle} is the probability that the medium is idle when the station in state $(0,0)_e$ attempts to transfer a frame. Malone *et al.* assumed that $P_{idle} = 1 - p_l$, and we use this assumption in this article. We can explicitly obtain the stationary distribution of other states.

A station in state $(k,0)$ ($0 \leq k \leq m$) attempts to transmit a frame when the medium is idle at the beginning of the next time slot. A station in state $(0,0)_e$ also attempts transmission at the beginning of the next time slot if (at least) one frame arrives during the current time slot. The probability that station l attempts transmission, τ_l , is then given by

$$\begin{aligned} \tau_l &= q_l P_{idle} b(0,0)_e + \sum_{i \geq 0} b(i,0) \\ &= b(0,0)_e \left(\frac{q_l^2 W_0}{(1-p_l)(1-r_l)(1-(1-q_l)W_0)} - \frac{q_l r_l P_{idle}}{1-r_l} \right). \end{aligned} \quad (3)$$

As shown in (2), the stationary distribution of each state contains unknown parameter p_l , q_l , and r_l . If packets arrive at station l according to a Poisson process with mean rate λ_l , we can estimate p_l and q_l through the following equations:

$$\begin{aligned} p_l &= 1 - \prod_{j \neq l} (1 - \tau_j), \\ q_l &= \left(\prod_j (1 - \tau_j) \right) (1 - e^{-\lambda_l T_s}) + \left(1 - \prod_j (1 - \tau_j) \right) (1 - e^{-\lambda_l T_c}). \end{aligned} \quad (4)$$

To obtain r_l , we assume that the station l can be modeled as a queue with infinite-buffer, and observe that the mean inter-arrival time of packets at station l should be equal to the mean frame-transmission interval of station l if the queue is stable. Now assume that a station

enters one of post-backoff or backoff states via absorbing state $(0,0)_a$ after successful frame transmission. (The sojourn time in $(0,0)_a$ is assumed to be zero.) Note that the mean return time to $(0,0)_a$ is equal to the mean frame-transmission interval. With denoting E_s the expected time spent per state, it follows from the fact $b(0,0)_a = \tau_l(1 - p_l)$ that

$$\text{mean frame-transmission interval} = \frac{E_s}{b(0,0)_a} = \frac{E_s}{\tau_l(1 - p_l)}.$$

Since the mean inter-arrival time of packets is $1/\lambda$,

$$\frac{1}{\lambda} = \frac{E_s}{\tau_l(1 - p_l)}, \quad (5)$$

from which we obtain

$$r_l = \left\{ 1 - q + \frac{q_l(W_0 + 1)(1 - P_{idle})}{2} + \frac{T_F - 1/\lambda_l}{E_s} \frac{q_l^2 W_0}{1 - (1 - q)^{W_0}} \right\} \\ \left/ \left\{ 1 - q + \frac{q_l(W_0 + 1)(1 - P_{idle})}{2} + \frac{T_F - 1/\lambda_l}{E_s} (1 - p)qP_{idle} \right\} \right., \quad (6)$$

where T_F is the mean MAC-layer packet service time, which is defined as the time interval between the instant that a packet reaches the head of the queue and the instant that the packet is successfully transferred, and it is approximately represented by (8). If the right hand side of (6) exceeds 1, we set $r_l = 1$. Note that the right hand side of (6) exceeds 1 only when station l is congested and thus the frame loss frequently occurs due to the buffer overflow at station l . The expected time spent per state E_s is given as follows:

$$E_s = \left(\prod_i (1 - \tau_i) \right) T_s + \left(1 - \prod_i (1 - \tau_i) \right) T_c,$$

where T_s is the length of time slot, and T_c is the expected time taken for a collision. In this article, we assume that RTS/CTS is disabled and thus

$$T_c = \frac{\text{ACK} + 2 \times \text{PHY}}{R_b} + \frac{\text{DATA}}{R_d} + \text{SIFS} + \text{DIFS},$$

where

- SIFS: SIFS duration
- DIFS: DIFS duration
- ACK: length of ACK frame (without physical header)
- PHY: length of physical header
- DATA: length of data frame (without physical header)
- R_b : basic rate
- R_d : data rate

Equations (3), (4), and (6) are simultaneous equations concerning p_j, q_j, r_j, τ_j for $j = 1, \dots, n$ which can be numerically solved by iterative substitution.

Remark 1. The difference between our model and the model by Malone *et al.* (Malone *et al.*, 2007) is in (1) where Malone *et al.* assumes $r_l = q_l$ but our model does not. Parameter q_l is the probability that at least one frame arrives at station l during a time slot while r_l is the probability that station l has at least one frame after successful frame transmission. Since r_l is almost equal to the probability that at least one frame arrives during the mean MAC-layer packet service time, r_l is usually larger than q_l and both parameters are the same only when station l has no buffer. In this sense, $r_l = q_l$ is equivalent to the buffer less model where each station is able to have at most one frame.

3.3 Throughput and MAC-layer packet service time

The throughput of a flow is the ratio of the length of a data frame to the inter-frame-transmission time; that is,

$$\text{throughput} = \frac{DATA}{E_s/b(0,0)_a} = \frac{(1-p_l)\tau_l DATA}{E_s}. \quad (7)$$

The MAC-layer packet service time is the interval between the instant that the station enters one of back-off state and the instant that it successfully transmits a frame. The MAC-layer packet service time can be approximated by the interval between the instant entering state $(0,0)$ and the instant of successful frame transfer, and its mean is given as follows:

$$T_F = \left(\frac{1+p_l W_0 (2p_l)^m}{2(1-p_l)} + \frac{W_0 (1-(2p_l)^{m+1})}{2(1-2p_l)} \right) E_s. \quad (8)$$

4. Queueing Modeling for IP Layer Analysis

At mobile terminals, the network layer receives packets from the transport layer. At the AP, network layer receives packets from mobile terminals or a router connected via wired line. Received packets make a queue in the buffer and are sequentially delivered to destinations via the IEEE 802.11-MAC layer. The queueing analysis is required to evaluate the queueing delay in the buffer or the packet loss due to buffer overflow, which have large impact on the end-to-end quality of service of applications. In this section, we show how we could evaluate these performance metrics by the queueing analysis.

4.1 Evaluation of GI/M/1/K+1 model

A queueing model is mainly characterized by the arrival process and the service time distribution. In the current model, the service time corresponds to the MAC-layer packet service time. It was reported in (Zhai *et al.*, 2004) that the exponential distribution is a good approximation for the MAC-layer packet service time, and thus in this article we use this approximation. The exponential distribution is fully characterized by the mean value, which is given by (8). To evaluate the performance under constant-bit-rate traffic like VoIP, we assume that packets arrive according to a renewal process at each station. Under the renewal-process arrival and the exponential service distribution, the queueing behavior of each station is modeled as a GI/M/1/K+1 model, where the system is able to have at most $K+1$ customers (one is server and other K customers in the buffer). Note that it is not difficult to extend the analysis under renewal arrivals to that under some non-renewal (correlated) arrival processes including Markov Modulated Arrival Process (MMPP) or Markovian Arrival Process (MAP) (Neuts, 1981).

The analysis of $GI/M/1/K+1$ model is often conducted by the imbedded Markov-Chain technique (Gross & Harris, 1998), where customer arrival instants form imbedded Markovian points. Let π_j denote the steady state probability that j customers (packets) stay in the system at arrival instants, and let $P = \{p_{ij}\}$ represents the transition probability matrix:

$$p_{ij} = P[X_{n+1} = j | X_n = i],$$

where X_n denotes the number of customers (frames) in the system at the n th arrival instant. The balance equation is $\pi = \pi P$ where $\pi = (\pi_0, \dots, \pi_{K+1})$. We define

$$\beta_k \stackrel{\text{def}}{=} \int \frac{(\mu T)^k}{k!} e^{-\mu x} A(dx),$$

where $A(x)$ is the distribution function of the inter-arrival time. Note that β_k is the probability that k customers departs from the queue during the inter-arrival time. It is easy to see that

$$P = \begin{pmatrix} 1 - \beta_0 & \beta_0 & 0 & \dots & 0 & 0 \\ 1 - \beta_1 - \beta_0 & \beta_1 & \beta_0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 - \sum_{j=0}^{K-1} \beta_j & \beta_{K-1} & \beta_{K-2} & \dots & \beta_0 & 0 \\ 1 - \sum_{j=0}^K \beta_j & \beta_K & \beta_{K-1} & \dots & \beta_1 & \beta_0 \\ 1 - \sum_{j=0}^K \beta_j & \beta_K & \beta_{K-1} & \dots & \beta_1 & \beta_0 \end{pmatrix}.$$

It follows from the balance equation that for $j \geq 1$

$$\pi_j = \sum_{k=0}^{K+1-j} \beta_k \pi_{k+j-1} + \beta_{K+1-j} \pi_{K+1},$$

from which

$$\pi_{j-1} = \frac{1}{\beta_0} \left\{ \pi_j - \sum_{k=1}^{K+1-j} \beta_k \pi_{k+j-1} - \beta_{K+1-j} \pi_{K+1} \right\}. \quad (9)$$

Observe that the right-hand-side of the above equation is represented in terms of $\{\pi_j, \dots, \pi_{K+1}\}$, which enables us to recursively obtain the steady state distribution.

4.2 Average Delay and Loss Ratio

Once we obtain the steady state distribution at arrival instants $\{\pi_j\}_{j=0}^{K+1}$, we can evaluate the queueing delay and the loss ratio. For example, the average queueing delay $E[D]$ is given by

$$E[D] = \frac{1}{\mu} \sum_{j=1}^{K+1} k \pi_j.$$

The distribution function of the queueing delay is

$$\begin{aligned} P[D \leq x] &= \sum_{j=0}^{K+1} P[D \leq x | k] \pi_j \\ &= \sum_{j=0}^{K+1} \left(1 - \sum_{k=0}^{j-1} e^{-\mu x} \frac{(\mu x)^k}{k!} \right) \pi_j. \end{aligned} \quad (10)$$

The packet loss ratio is equal to π_{K+1} .

Basic rate	1Mbps
Date rate	11Mbps
PHY header	192bits
MAC header	288bits
ACK length	112bits + PHY header
SIFS	10 μ s
DIFS	50 μ s
Slot time	20 μ s

Table 1. DCF parameters used in the numerical examples

5. Numerical Experiments: Evaluation of the Admissible Limits of Voice Flows

5.1 Conditions of Numerical Experiments

In this section, we see the accuracy of the proposed analytical model by comparing numerical analysis and computer simulation results. We used network simulation tools ns2 to obtain simulation results. In the simulation, there were n mobile terminals in an IEEE 802.11b-based wireless LAN. Each mobile terminal conducted a bidirectional voice conversation through the AP with a node outside the WLAN, and thus there were n uplink and n downlink voice flows under n mobile terminals. Each voice flow generated G.711-codec traffic; 200 byte packets (160-byte data and 40-byte RTP/UDP/IP header) were generated every 20 ms in a voice flow. The parameters of the DCF used in the numerical examples is depicted in Table 1. The buffer-sizes of the AP and mobile terminals were all set at 30 in packet.

In the experiments, we evaluated the throughput and packet-loss ratio of each flow. We also investigated how many voice conversations could be multiplexed in the wireless LAN without having packet loss, which we refer to as the “multiplexable limit of voice conversations” and denote by N_{max} in this article. As mentioned in Section 1, the uplink/downlink unfairness in WLANs makes the AP the performance bottleneck under the standard IEEE 802.11 DCF. The CW differentiation between the AP and mobile terminals would provision a fair resource sharing between the uplink and downlink traffic. In the experiments, we investigate how much the CW differentiation enhances the multiplexable limit of voice flows.

5.2 Results of Numerical Experiments

5.2.1 Throughput

We first evaluated the throughput of uplink and downlink voice flows when the contention window parameters of all stations were set at $(CW_{min}, CW_{max}) = (31, 1023)$, which are the default setting of the IEEE 802.11. Figure 2 compares analytical and simulation results concerning the total throughputs of uplink flows as well as the total throughputs of downlink flows. For reference, we also show the results evaluated by the analytical model of Malone (Malone et al., 2007). The result was given in terms of application level throughput defined by (7), where we exclude the lengths of PHY, MAC, IP, UDP, and RTP headers from the length of data frame. The throughput estimated by our analytical model agrees well with simulation results. Figure 2 shows that the uplink flows obtained larger throughput than the downlink, indicating that the AP was the performance bottleneck.

Figure 3 shows the result when the contention window parameters of the AP were set at $(CW_{min}, CW_{max}) = (7, 1023)$. Note that parameter setting $(CW_{min}, CW_{max}) = (7, 1023)$ gives

higher priority to the AP over mobile terminals and thus, under this parameter setting, the unfairness between uplink and downlink flows should be improved. Actually, the difference between uplink and downlink flows in the total throughput became smaller than the case when $(CW_{min}, CW_{max}) = (31, 1023)$. The throughput estimated by our analytical model agrees well with simulation results when the number of mobile terminals was less than 13, but some discrepancy was observed when the number of mobile terminals was larger than 15. This discrepancy may come from (5) where we neglect the packet loss at the buffer of stations. We also evaluated the throughput when the contention window parameters of the AP were set at $(CW_{min}, CW_{max}) = (3, 7)$. The results are shown in Figure 4. In this parameter setting, the downlink flows obtained larger throughput than the uplink, indicating that mobile terminals were the performance bottleneck.

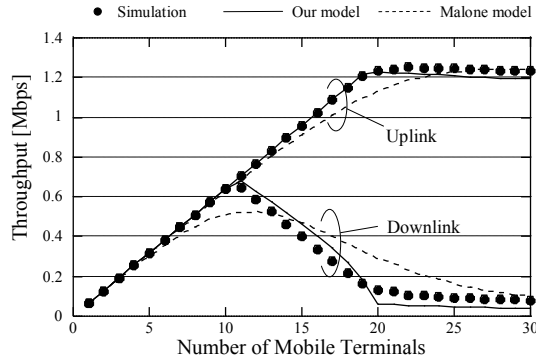


Fig. 2. Throughput versus the number of voice flows: $(CW_{min}, CW_{max}) = (31, 1023)$ at the AP.

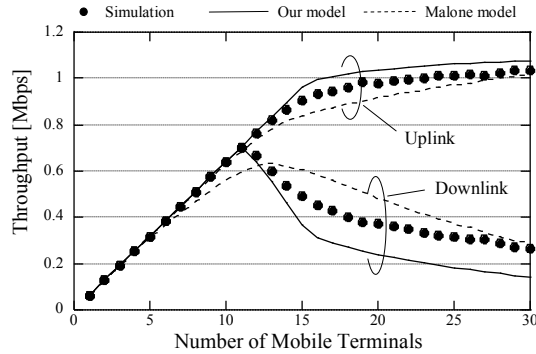


Fig. 3. Throughput versus the number of voice flows: $(CW_{min}, CW_{max}) = (7, 1023)$ at the AP.

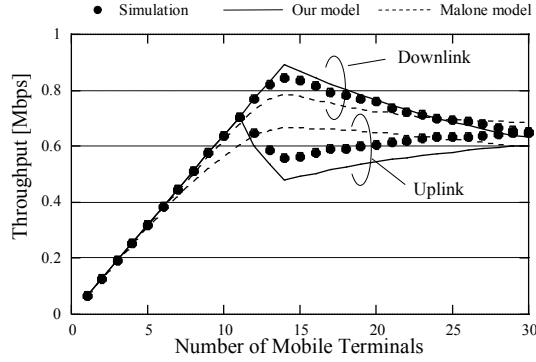


Fig. 4. Throughput versus the number of voice flows: $(CW_{min}, CW_{max}) = (3, 7)$ at the AP.

5.2.2 Multiplexable limit of voice conversation

From the total throughput of uplink or downlink flows, we see whether the AP or mobile terminal is overloaded or not. To explain this, let T_{up} and T_{down} respectively denote the total throughput of uplink and downlink flows. Since one voice flow generates 64kbps traffic, if

$$T_{down} = n \times 64\text{kbps} \quad (11)$$

is not satisfied, then the sufficient throughput for downlink flows is not obtained and thus the AP is overloaded, while if

$$T_{up} = n \times 64\text{kbps} \quad (12)$$

is not satisfied, then mobile terminals are overloaded. We define the multiplexable limit for downlink (uplink) flows by the maximum number of voice flows satisfying (11) ((12)) and denote it by N_{max}^{\downarrow} (N_{max}^{\uparrow}). The multiplexable limit of voice conversation N_{max} is equal to $\min\{N_{max}^{\downarrow}, N_{max}^{\uparrow}\}$.

(CW_{min}, CW_{max}) of the AP	(31, 1023)			(7, 1023)			(3, 7)		
	N_{max}^{\downarrow}	N_{max}^{\uparrow}	N_{max}	N_{max}^{\downarrow}	N_{max}^{\uparrow}	N_{max}	N_{max}^{\downarrow}	N_{max}^{\uparrow}	N_{max}
Simulation	10	15	10	11	11	11	12	11	11
Our model	10	19	10	11	15	11	13	11	11
Malone model	1	1	1	1	1	1	1	1	1

Table 2. Multiplexable limit of voice conversation.

Table 2 summarizes the multiplexable limits of voice conversations under three different combinations of congestion window parameters of the AP. The congestion window parameters of mobile terminals were all set at $(CW_{min}, CW_{max}) = (31, 1023)$. For all cases, the multiplexable limits of voice conversations N_{max} estimated by our model agreed with the simulation results although some discrepancy was observed in the estimation of N_{max}^{\downarrow} or N_{max}^{\uparrow} . Under the analytical model by Malone *et al.* (Malone et al., 2007), (11) and (12) were not satisfied when more than one voice conversation were multiplexed. Thus, according their model, $N_{max} = N_{max}^{\downarrow} = N_{max}^{\uparrow} = 1$, which was far from the simulation results.

The unbalance between N_{ad}^{\downarrow} and N_{ad}^{\uparrow} when $(CW_{min}, CW_{max}) = (31, 1023)$ comes from the uplink/downlink unfairness in WLANs. The table shows that the discrepancy was resolved as the congestion window parameters of the AP became smaller. The multiplexable limit of voice conversation, however, did not increase so much even when the uplink/downlink unfairness was improved.

5.2.3 Packet Loss Ratio

We also evaluated the packet loss ratios of uplink and downlink voice flows by our analytical model and simulation. Results were depicted in Figure 5 when the contention window parameters of the AP were $(CW_{min}, CW_{max}) = (31, 1023)$, in Figure 6 when $(CW_{min}, CW_{max}) = (7, 1023)$, and in Figure 7 when $(CW_{min}, CW_{max}) = (3, 7)$. These figures indicate that results by our analytical model agree well with the simulation results. The discrepancy between analytical results and simulation may come from that assumption that the mobile terminals have large buffer to temporarily keep frames, which is not satisfied in the setting of ns2.

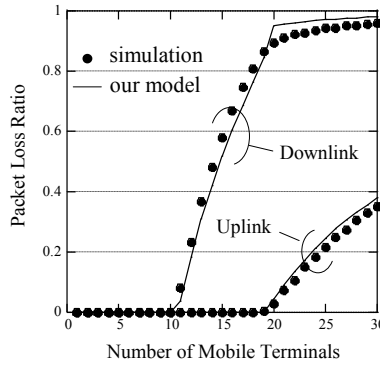


Fig. 5. Packet loss ratio: $CW_{min}=31, CW_{max}=1023$.

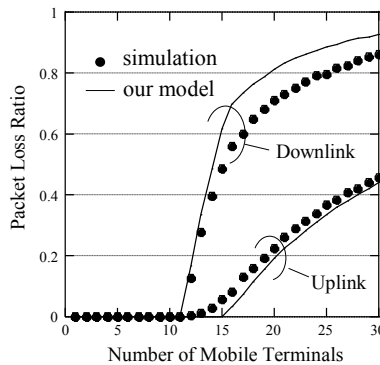


Fig. 6. Packet loss ratio: $CW_{min}=7, CW_{max}=1023$.

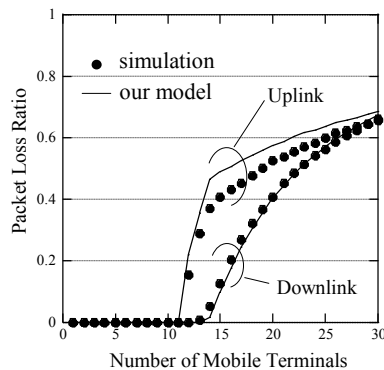


Fig. 7. Packet loss ratio: $CW_{min}=3$, $CW_{max}=7$.

6. Conclusion

In this article, we proposed an analytical model for jointly evaluating the performance of the IEEE 802.11-DCF MAC layer and of the network layer. We find that our model accurately evaluate the per-flow throughput as well as the packet loss ratio when a number of uplink and downlink voice flows are multiplexed over a WLAN. There exists some discrepancy between the prediction by our model and simulation results especially when the number of multiplexed voice flows is quite large. The cause of the discrepancy needs to be further explored. In the current model, the frame loss due to exceeding the retry limit is not taken into consideration, which also remains a future work.

7. References

- Bianchi, G. (2000). Performance analysis of the IEEE802.11 distributed coordination function, *IEEE Journal on Selected Areas in Communications* **18**(3): 535–547.
- Cai, L., Shen, X., Mark, J., Cai, L. & Xiao, Y. (2006). Voice capacity analysis of wlan with unbalanced traffic, *IEEE Trans. Veh. Technol* **55**(3): 752–761.
- Cheng, Y., Ling, X., Song, W., Cai, L., Zhuang, W. & Shen, X. (2007). A cross-layer approach for WLAN voice capacity planning, *IEEE J. Select. Areas Commun.* **25**(4): 678–688.
- Daneshgaran, F., Laddomada, M., Mesiti, F. & Mondin, M. (2008). Unsaturated throughput analysis of IEEE 802.11 in presence of non ideal transmission channel and capture effects, *IEEE Trans. Wireless Communications* **7**(4): 1276–1286.
- Ergena, M. & Varaiya, P. (2005). Throughput analysis and admission control in IEEE 802.11a, *Mobile Networks and Applications* **10**(5): 705–716.
- Foh, C., Zukerman, M. & Tantra, J. (2007). A Markovian framework for performance evaluation of IEEE 802.11, *IEEE Trans. Wireless Communications* **6**(4): 1276–1285.
- Gross, D. & Harris, C. (1998). *Fundamentals of Queueing Systems*, 3rd Ed., John Wiley & Sons.
- Malone, D., Duffy, K. & Leith, D. (2007). Modelling the 802.11 distributed coordination function in non-saturated heterogeneous conditions, *IEEE/ACM Trans. Networking* **15**(1): 159–172.
- Neuts, M. (1981). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, The Johns Hopkins University Press.

- Robinson, J. & T.S.Randhawa (2004). Saturation throughput analysis of IEEE802.11e enhanced distributed coordination function, *IEEE J. Select. Areas Commun.* **33**(5): 917–928.
- Tickoo, O. & Sikdar, B. (2004). Queueing analysis and delay mitigation in IEEE802.11 random access MAC based wireless networks, *IEEE INFOCOM*, pp. 1404–1413.
- Xiang, B., Yu-Ming, M. & Jun, X. (2007). Performance investigation of IEEE802.11e EDCA under non-saturation condition based on the M/G/1/K models, *IEEE ICIEA*, pp. 298–304.
- Xiao, Y. (2005). Performance analysis of priority schemes for IEEE802.11 and IEEE802.11e wireless LANs, *IEEE Trans. Wireless Communications* **4**(4): 1506–1515.
- Zhai, H., Kwon, Y. & Fang, Y. (2004). Performance analysis of IEEE 802.11 MAC protocols in wireless LANs, *Wireless Communications and Mobile Computing* **4**: 917–931.
- Zhao, Q., Tsang, H. & Sakurai, T. (2008). A simple model for nonsaturated IEEE 802.11 DCF networks, *IEEE Communication Letters* **12**(8): 563–565.

A. Analysis of the Markov chain of Figure 1

Since $b(i+1, 0) = pb(i, 0)$ for $i \geq 1$ and $b(1, 0) = b(0, 0)p + b(0, 0)_e qpP_{idle}$, we obtain

$$\sum_{i \geq 1} b(i, 0) = \frac{b(1, 0)}{1 - p} = \frac{b(0, 0)p + b(0, 0)_e qpP_{idle}}{1 - p}. \quad (13)$$

The balance equation concerning state $(0, W_0 - 1)_e$ yields

$$b(0, W_0 - 1)_e = b(0, 0)_e \frac{q(1 - p)P_{idle}}{W_0} + \frac{(1 - p)(1 - r)}{W_0} \sum_{i \geq 0} b(i, 0).$$

Substituting (13) into the above yields

$$b(0, W_0 - 1)_e = b(0, 0)_e \frac{q(1 - rp)P_{idle}}{W_0} + b(0, 0) \frac{1 - r}{W_0}. \quad (14)$$

From the balance equation concerning state $(0, k)_e$, we have

$$\begin{aligned} b(0, k)_e &= (1 - q)b(0, k + 1)_e + b(0, W_0 - 1)_e, \quad \text{for } W_0 - 1 > k > 0, \\ qb(0, 0)_e &= (1 - q)b(0, 1)_e + b(0, W_0 - 1)_e, \end{aligned} \quad (15)$$

from which for $k > 0$

$$b(0, k)_e = b(0, W_0 - 1)_e \frac{1 - (1 - q)^{W_0 - k}}{q}, \quad (16)$$

and

$$qb(0, 0)_e = b(0, W_0 - 1)_e \frac{1 - (1 - q)^{W_0}}{q}. \quad (17)$$

Substituting (14) into (17) yields

$$\frac{b(0, 0)_e}{b(0, 0)} = \frac{1 - r}{q} \frac{1 - (1 - q)^{W_0}}{qW_0 - P_{idle}(1 - rp)(1 - (1 - q)^{W_0})}. \quad (18)$$

It follows from (16) and (17) that

$$\sum_{k=1}^{W_0-1} b(0, k)_e = b(0, 0)_e \left\{ \frac{W_0 q}{1 - (1 - q)^{W_0}} - 1 \right\},$$

and thus

$$\sum_{k=0}^{W_0-1} b(0, k)_e = b(0, 0)_e \frac{W_0 q}{1 - (1 - q)^{W_0}}. \quad (19)$$

Next we consider the stationary probability of state $(0, k)$. The balance equation concerning state $(0, W_0 - 1)$ yields

$$\begin{aligned} b(0, W_0 - 1) &= \sum_{k \geq 0} b(k, 0) \frac{(1 - p)r}{W_0} + b(0, 0)_e \frac{q(1 - P_{idle})}{W_0} \\ &= \left\{ b(0, 0) + \frac{b(0, 0)p + b(0, 0)_e q p P_{idle}}{1 - p} \right\} \frac{(1 - p)r}{W_0} + b(0, 0)_e \frac{q(1 - P_{idle})}{W_0} \\ &= \left\{ \frac{b(0, 0) + b(0, 0)_e q p P_{idle}}{1 - p} \right\} \frac{(1 - p)r}{W_0} + b(0, 0)_e \frac{q(1 - P_{idle})}{W_0} \\ &= b(0, 0) \frac{r}{W_0} + b(0, 0)_e \frac{q}{W_0} \{1 - (1 - pr)P_{idle}\}. \end{aligned} \quad (20)$$

It comes from the balance equation concerning state $(0, k)$ that for $W_0 - 1 > k \geq 0$

$$\begin{aligned} b(0, k) &= b(0, k + 1) + b(0, W_0 - 1) + q b(0, k + 1)_e \\ &= b(0, k + 1) + b(0, W_0 - 1) + b(0, W_0 - 1)_e (1 - (1 - q)^{W_0 - k - 1}) \\ &= (W_0 - k) b(0, W_0 - 1) + b(0, W_0 - 1)_e \sum_{n=1}^{W_0 - 1 - k} \{1 - (1 - q)^n\} \\ &= (W_0 - k) (b(0, W_0 - 1) + b(0, W_0 - 1)_e) - b(0, W_0 - 1)_e \frac{1 - (1 - q)^{W_0 - k}}{q}. \end{aligned} \quad (21)$$

Combining (14), (20), and (21) yields

$$\begin{aligned} \sum_{k=0}^{W_0-1} b(0, k) &= b(0, 0) \left\{ \frac{W_0 + 1}{2} - \frac{1 - r}{q} + \frac{(1 - q)(1 - (1 - q)^{W_0})(1 - r)}{q^2 W_0} \right\} \\ &\quad + b(0, 0)_e \left\{ \frac{q(W_0 + 1)}{2} - (1 - rp)P_{idle} + \frac{P_{idle}(1 - q)(1 - rp)(1 - (1 - q)^{W_0})}{q W_0} \right\} \end{aligned} \quad (22)$$

By representing $b(0, 0)$ in (22) in terms of $b(0, 0)_e$ through (18), we obtain

$$\begin{aligned} \sum_{k=0}^{W_0-1} b(0, k) &= b(0, 0)_e \frac{q}{1 - r} \left\{ \frac{q W_0}{1 - (1 - q)^{W_0}} - P_{idle}(1 - rp) \right\} \\ &\quad \times \left\{ \frac{W_0 + 1}{2} - \frac{1 - r}{q} + \frac{(1 - q)(1 - (1 - q)^{W_0})(1 - r)}{q^2 W_0} \right\} \\ &\quad + b(0, 0)_e \left\{ \frac{q(W_0 + 1)}{2} - (1 - rp)P_{idle} + \frac{P_{idle}(1 - q)(1 - rp)(1 - (1 - q)^{W_0})}{q W_0} \right\} \end{aligned}$$

$$\begin{aligned}
&= b(0,0)_e \left[1 - q - \frac{qW_0}{1 - (1-q)^{W_0}} + \frac{q(W_0 + 1)}{2(1-r)} \right. \\
&\quad \times \left. \left(\frac{qW_0}{(1 - (1-q)^{W_0})} + (1 - P_{idle})(1-r) - rP_{idle}(1-p) \right) \right].
\end{aligned} \tag{23}$$

Since $b(i,k) = (W_i - k)/W_i b(i,0)$ for $i > 0$, it follows that

$$\sum_{k=0}^{W_i-1} b(i,k) = b(i,0) \frac{W_i + 1}{2},$$

from which we have

$$\begin{aligned}
\sum_{i=1}^m \sum_{k=0}^{W_i-1} b(i,k) &= \sum_{i=1}^m b(i,0) \frac{W_i + 1}{2} + \sum_{i=m+1}^{\infty} b(i,0) \frac{W_m + 1}{2} \\
&= \frac{b(1,0)}{2} \left\{ \sum_{i=1}^m p^{i-1} (W_0 2^i + 1) + \sum_{i=m+1}^{\infty} p^{i-1} (W_0 2^m + 1) \right\} \\
&= \frac{b(1,0)}{2} \left\{ \frac{1}{1-p} + \frac{2W_0(1 - (2p)^m)}{1-2p} + \frac{W_0(2p)^m}{1-p} \right\} \\
&= \frac{b(1,0)}{2(1-p)} \left\{ 1 + 2W_0 \frac{1-p - p(2p)^{m-1}}{1-2p} \right\}.
\end{aligned} \tag{24}$$

It comes from (13) and (18) that

$$b(1,0) = b(0,0)_e \frac{pq^2}{1-r} \left(\frac{W_0}{1 - (1-q)^{W_0}} - \frac{P_{idle}(1-p)r}{q} \right). \tag{25}$$

By substituting (19), (23), (24), and (25) into the normalization condition

$$\sum_{i=0}^m \sum_{k=0}^{W_i-1} b(i,k) + \sum_{k=0}^{W_m-1} b(0,k)_e = 1,$$

we finally have

$$\begin{aligned}
1/b(0,0)_e &= 1 - q + \frac{q(W_0 + 1)}{2(1-r)} \left(\frac{qW_0}{(1 - (1-q)^{W_0})} + (1 - P_{idle})(1-r) - rP_{idle}(1-p) \right) \\
&\quad + \frac{pq^2}{2(1-p)(1-r)} \left(\frac{W_0}{1 - (1-q)^{W_0}} - \frac{P_{idle}(1-p)r}{q} \right) \left\{ 1 + 2W_0 \frac{1-p - p(2p)^{m-1}}{1-2p} \right\}
\end{aligned} \tag{26}$$

Once we have obtained $b(0,0)_e$, the stationary probabilities of other states are easy to calculate.

Increasing the Time Connected to Already Deployed 802.11 Wireless Networks while Traveling by Subway

Jaeouk Ok, Pedro Morales, Masateru Minami and Hiroyuki Morikawa
The University of Tokyo
Japan

1. Introduction

Recently, an increasing number of people retrieve various contents via the Internet using a publish/subscribe service such as podcasts. Typical usage of those applications is to subscribe to as many favorite sites as possible, and selectively enjoy the automatically downloaded latest episodes. Newly available portable multimedia players enable people to easily enjoy downloaded video/audio contents in a variety of places, and the added communication functionality such as 3G or 802.11 (IEEE Standard 802.11, 1999) makes it possible to retrieve the latest episodes as soon as they are published on the web.

Service like podcasts are especially beneficial to people traveling by subway¹ considering their idle time in a subway train. However, most of subway tunnels in Tokyo are unfortunately covered by neither 3G nor 802.11 as of December 2008. Wireless connection can be established when a subway train stays under coverage areas at a station², but it will be repeatedly interrupted each time the subway train passes through non-coverage areas in the tunnels while traveling along a railroad. This interruption limits the time under coverage areas, which reduces the maximum possible connected time. This time is further reduced by current implementation exploiting the intermittent connectivity poorly. In this chapter, we focus on the 802.11 wireless connection management while traveling by subway because of its higher throughput, lower subscription cost, and larger variety of 802.11-enabled portable devices than 3G.

We aim to increase the time connected to already deployed 802.11 wireless networks for podcast-like applications while traveling by subway in Tokyo. To understand the target environment, we investigated the commercial 802.11 HOTSPOT networks (NTT Communications HOTSPOT, [Online]) deployed in Tokyo Metro. One of the findings

¹Tokyo Metro carries average 6.22 million passengers per day as of 2007 (Tokyo Metro, [Online])

²In Tokyo, for example, approximately 97% of subway stations are densely covered by three different service providers as of December 2008 (NTT Communications HOTSPOT; NTT DoCoMo Mzone; NTT EAST FLET'S SPOT, [Online])

against the common belief regarding long distance mobility is that the main factor to the diminishment of available connected time is link layer connection management, not IP layer mobility support because of the deployed VLANs across the target networks. We propose an optimized solution for this subway environment to increase the connected time by reducing the following two types of delay. The one delay is experienced when establishing the wireless connection after coming out of non-coverage area in the tunnel, and the other when switching the wireless connection to the next AP while crossing overlapping coverage areas at stations.

Our method reduces the establishment delay by building a chain that links the last AP in the previous station before the tunnel with the first AP in the next station after the tunnel, called *border APs* in this chapter. By referring to this chain when leaving a station, a client can reduce the delay to establish the connection through the use of passive scan only on the channel corresponding to the upcoming border AP. The switching delay is reduced by building a list of the APs available at each station for which a client has connection authorization, called *preferred APs* in this chapter. By referring to this list when crossing overlapping coverage areas in a station, a client can reduce the delay to switch the connection through the use of unicast scan using *Authentication Request* frames to the limited number of preferred APs. This is feasible by taking advantage of the key attributes of the target environment: strong mobility pattern along a railroad and limited number of APs at each station. In our analysis, the delay obtained by our method is 94.4% smaller than the one obtained by passive scan when establishing terminated wireless connections, and 94.2% smaller than the one obtained by active scan when switching wireless connection to the next AP.

The rest of this chapter is organized as follows. Section 2 describes the findings from investigating the commercial 802.11 HOTSPOT networks deployed in Tokyo Metro. We propose a method to increase the connected time to already deployed 802.11 wireless networks while traveling by subway in Tokyo in Section 3. Section 4 analyzes the increased connected time by our proposed method in comparison with related work. Section 5 presents the effectiveness of our system through implementation and experiments. Section 6 concludes the chapter, and shows future work.

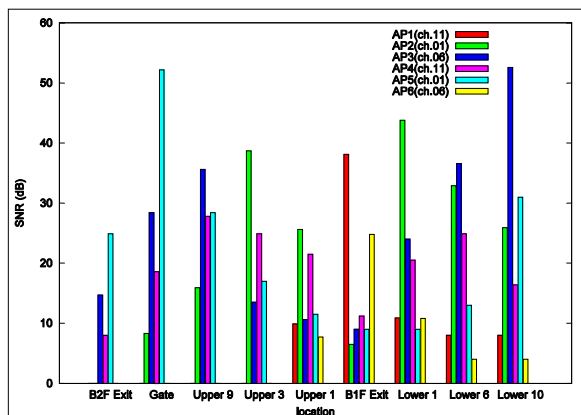


Fig. 1. Averaged SNR values of beacon frames measured at different locations in Waseda (T4) station. The names of location refer to Fig. 2.

2. Target Environment Investigation

In order to find out the factors that decrease the possible connected time, we investigate the commercial 802.11 HOTSPOT network in Tokyo Metro. All experiments were performed with a windows XP machine while walking at stations and moving by subway in November 2008. NetStumbler (NetStumbler, [Online]), Wireshark (Wireshark, [Online]), and built-in Wireless Auto Configuration were used to monitor beacon frames, analyze IP packets and manage 802.11 wireless network connections. The findings are classified by the points of our interest: link layer handoff, IP mobility support, and restrictions on application layer. This section discusses each of them in detail.

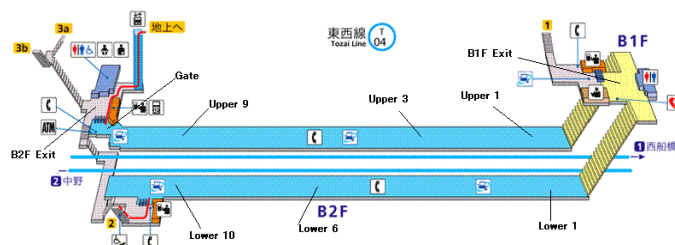


Fig. 2. Waseda (T4) station structure (Tokyo Metro, [Online]) There are six 802.11g APs installed on channel 1, 6 and 11 to collaboratively cover the station including entrances, ticket gates, platforms, etc.

2.1 Link Layer Handoff

To find the necessity of link layer handoff in 802.11 HOTSPOT³, we studied coverage areas at stations by measuring the averaged SNR values of beacon frames at different locations in 10 stations⁴. In each station, multiple APs ranging from four to seven are installed to collaboratively cover entrances, ticket gates, platforms, passages for transfer, etc. Figure 1 shows an example of the measured results in Waseda (T4) station on Tozai line, whose structure is as shown in Figure 2. From the figure we observe that: 1) the coverage area of a single AP is not large enough to cover the entire station, and 2) each location is under coverage areas of multiple APs. Therefore, while a client moves around at stations, it is necessary to switch the wireless connection across overlapping wireless coverage areas. For example, let us assume that a subway train comes in from right in Figure 2. The clients in the train get associated to AP2 according to Figure 1. For some clients staying in the train, the established wireless connection will be switched from AP2 to AP3, when the train moves to the left for the next station. For others getting off the subway train at Lower 1 on the platform and walk out B1F Exit, the wireless connection will be switched from AP2 to AP1. Unlike stations with overlapping coverage areas, there are non-coverage areas in each tunnel between stations. To show non-coverage areas in tunnels, we measured the time

³In Tokyo Metro, NTT Communications HOTSPOT provides 173 subway stations among 179 with IEEE 802.11 b/g wireless access service. (Tokyo Metro, [Online])

⁴Yoyogi-uehara (C1), Yoyogi-koen (C2), Meiji-jingumae (C3), Omote-sando (C4), Takatanobaba (T3), Waseda (T4), Kagurazaka (T5), Kasumigaseki (M15), Ginza (M16), Tokyo (M17), and Otemachi (M18) station

stamps of received beacon frames while moving by subway through seven stations from Nihombashi (T10) to Waseda (T4) station on Tozai line. The measured results are shown in Table 1. From the table we observe that: 1) there exist non-coverage areas in each tunnel, and 2) the time length under non-coverage areas is approximately one third of the total moving time on a railroad. This is explained by different speeds of subway when passing above two areas: high speed within large non-coverage areas in the tunnels and low speed within small coverage areas in the stations. Though there is large period of time spent under coverage of 802.11 wireless networks even while traveling by subway, the wireless connection is repeatedly interrupted due to non-coverage areas in the tunnel. Therefore, it is necessary to establish disconnected wireless connection whenever the client enters the following coverage areas.

Station	Coverage Area	Non-coverage Area
Nihombashi (T10)	83 sec	20 sec
Otemachi (T9)	66 sec	56 sec
Takebashi (T8)	69 sec	36 sec
Kudanshita (T7)	70 sec	29 sec
Iidabashi (T6)	81 sec	53 sec
Kagurazaka (T5)	81 sec	55 sec

Table 1. Non-coverage areas in the tunnels

2.2 IP Mobility Support

To find the necessity of IP mobility support, we studied IP network configuration by dumping and analyzing IP packets. After a successful association using the proper ESSID (i.e., 0033) and WEP key, a global IP address is assigned via Dynamic Host Configuration Protocol (DHCP) (Droms, 1997). The DHCP server has multiple ranges of address pool. It is possible to have various addresses assigned with the same client and AP at different times. To show the range of DHCP address pool, we collected the obtained TCP/IP address configuration with a single AP at different times in Yoyogi-uehara (C1) station on Chiyoda line. Part of the collected information is shown in Table 2.

Host address	Subnet Mask	Default Gateway
210.162.9.65	255.255.254.0	210.162.9.1
211.0.159.54	255.255.254.0	211.0.159.1
61.127.100.37	255.255.254.0	61.127.100.1

Table 2. The range of DHCP address pool

Once an IP address is assigned in the beginning of a session, the same address is repeatedly assigned not only from different APs in the same station, but also from the APs in other stations by another DHCP request during the DHCP lease length. We confirmed this fact by starting a session at Yoyogi-uehara (C1) station, and receiving the same IP address from two different APs in the same station and also receiving it in other stations such as Yoyogi-koen (C2), Meiji-jingumae (C3), Omote-sando (C4), etc.

From above two experiments, we observe that HOTSPOT implements Virtual Local Area Network (VLAN) (IEEE Standard 802.1Q-2003, 2003) to accommodate multiple subnets in

each physical LAN, and therefore, IP layer mobility support is unnecessary. The initially assigned IP address does not need to be changed, as long as it performs DHCP lease renewal every lease length (i.e., 5 minutes). In order to prevent DHCP renewal from failing in the non-coverage areas in the tunnels, it is necessary to perform DHCP renewal with a shorter interval: (lease length) - (the maximum time spent in the non-coverage areas). As shown in Table 1, the time spent in non-coverage areas in the tunnels are much shorter than the lease length, therefore, we can perform the DHCP discovery only once in the beginning of the session, and skip it later on by renewing DHCP with the above shorter interval.

2.3 Restrictions on Application Layer

NTT Communications HOTSPOT implements web-based authentication to complement the open system authentication. Despite a successful TCP/IP address configuration via DHCP, a client is unable to send and receive data traffic outside the network. To gain Internet access outside the network, users are required to enter a username and password in an authentication web page, to which the first attempt to access any web page after launching a web browser is automatically redirected. We also found that the web authentication expires when there is no traffic sent or received during DHCP lease length (i.e., 5 minutes). Again, as the time spent in non-coverage areas in the tunnels are much shorter than the lease length, it needs to be done only once in the beginning of the session in our subway mobility scenario. Moreover, a download manager such as DownThemAll (DownThemAll, [Online]) is necessary to automatically *resume* broken downloads due to the intermittent connectivity.

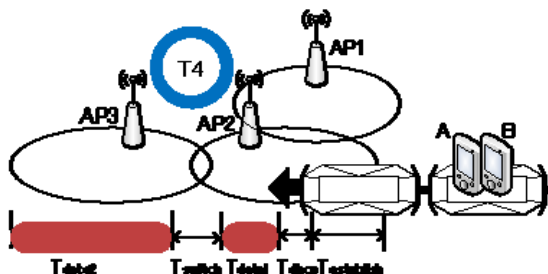


Fig. 3. Mobility scenario

3. Increasing Time Connected while Traveling by Subway

In this section, we describe our proposed method to increase the time connected to 802.11 wireless networks while traveling by subway. We look into the detailed composition of the main factor to the diminishment of available connected time, and propose a method to address them.

3.1 Problem Statement

To clarify the problems arising from utilizing 802.11 wireless connection while traveling by subway, consider the following scenario depicted in Figure 3. Assume that two clients (A and B) took a subway train in the previous T5 station downloading subscribed podcasts with 802.11-enabled portable devices. The wireless connection is terminated while passing through

the non-coverage area in the tunnel. When the train is entering T4 station, the devices discover that they become under coverage area, scan available APs, and get connected to AP2. When the train stops at T4 station, client A gets off and walk to the coverage area of AP1, but client B stays in the train and moves to the next station passing through the coverage area of AP3. Given that we do not modify the commercially deployed APs to enlarge coverage areas, the maximum connected time of devices passing through T4 station (e.g., client B) is limited shown in Figure 3, and only some parts T_{data} of it is used to download subscribed podcasts. This limited time is also used to recognize the coverage area of the firstly appearing AP and establish link layer connection to it ($T_{establish}$), to check the availability of a previously used IP address via DHCP Request message (T_{dhcp}), and switch link layer connection to a closer AP (T_{switch}). Since DHCP request can be skipped by DHCP renew with shorter interval as shown in the previous section, the main factor to the diminishment of available connected time is composed of $T_{establish}$ and T_{switch} for the link layer connection management.

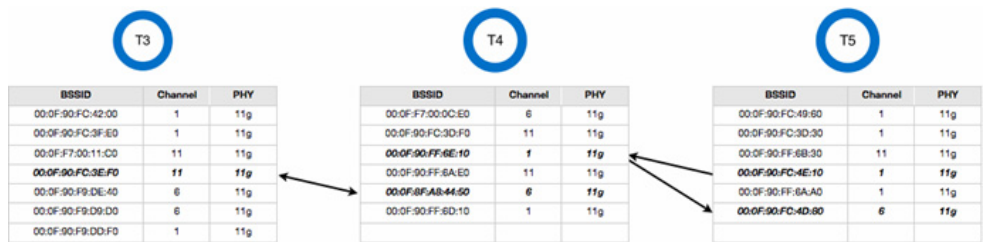


Fig. 4. An example of a chain of border APs and a list of preferred APs in T3, T4 and T5 stations

3.2 Proposed Method

In order to reduce the delay experienced when managing link layer connection, we take advantage of the following key attributes of the target environment: 1) strong mobility pattern along a railroad, and 2) limited number of APs at each station.

3.2.1 Fast Connection Establishment after Coming out of Non-coverage Areas

In order to reduce the delay experienced when establishing link layer connection ($T_{establish}$), a client needs to discover that it came out of non-coverage areas in the tunnel, and find the AP to associate to with less scanning delay. Because the standard does not define the behavior when no AP is found during scanning process, periodical active scanning is commonly implemented for a general purpose. For example, Wireless Auto Configuration in Windows XP, executes active scanning every 60 seconds (Microsoft Technet, [Online]) when no preferred AP is found during channel scanning phase. However, this periodic scanning makes $T_{establish}$ up to in the magnitude of tens of seconds in subway mobility scenario due to this coverage area recognition delay. Therefore, it is necessary to perform channel scanning continuously, when no AP is found in the tunnel.

We reduce $T_{establish}$ by continuously performing selective passive scan, which does not generate excessive management traffic under non-coverage area in the tunnel. Taking advantage of a strong mobility pattern along a railroad, we build a chain that links the last AP in the previous station before the tunnel with the first AP in the next station after the

tunnel, called border APs. By referring to this chain when leaving a station, a client can reduce $T_{\text{establish}}$ through the use of passive scan only on the channel corresponding to the upcoming border AP. For example, back in Figure 1, AP2 in T4 station is the border AP appearing at the end of non-coverage area in the tunnel from T5 station. A part of a chain of border APs regarding T3 (1 border AP), T4 (2 border APs), and T5 (2 border APs) stations is shown in Figure 4. The border APs are related by arrows with the information of BSSID, channel number, and PHY type.

The connection establishment process is performed in the following steps. When a client can not find any APs after being disconnected from a border AP (e.g., 00:0F:90:FC:4E:10 AP in T5 station), it assumes that it is passing through non-coverage area in the tunnel and looks up its chain of border APs to find next border AP (e.g., 00:0F:90:FF:6E:10 AP in T4 station). Then, it sets up its interface to the desired channel and PHY type (e.g., channel 1 and 11b/g), and waits for a beacon from the next border AP. If a beacon from the border AP arrives within the maximum time spent in the non-coverage area, a client executes authentication and association phase. Otherwise, standard active scan is executed to handle this unexpected case where the pre-built chain of border APs can not reflect the real situation due to changes in infrastructure or unexpected non-coverage area, etc.

The delay of selective passive scan is as follows. Under the best case scenario the beacon arrives as soon as the device enters the coverage area, so the delay is 0. The worst case happens when the device just missed beacon when it enters the coverage area, so the delay is the same as beacon interval, 100 msec by default. Therefore, the highest $T_{\text{establish}}$ will be composed of one beacon interval for the selective passive scan, one RTT for authentication, and one RTT for association phase.

3.2.2 Fast Connection Switching across Overlapping Coverage Areas

In order to reduce the delay experienced when switching link layer connection T_{switch} across overlapping coverage areas, a client needs to find the next AP to handoff with less scanning delay. Because a client has flexible mobility patterns in a station, referring to a chain of border APs can not tell clients staying in the train to go to next station from ones getting off to go to other direction, leading to executing active scan. Active scan takes long, because it tries to acquire the information about all nearby APs regardless of a client's connection authorization. Therefore, it is necessary to perform channel scanning only to the target APs that the client has connection authorization.

We reduce T_{switch} by performing multiple open system authentication only to target APs at each station, called AuthScan (Ok et al., 2008). Taking advantage of limited number of APs at each station, we build a list of the APs available at each station for which a client has connection authorization, called preferred APs. By referring to this list when crossing overlapping coverage areas in a station, a client can reduce T_{switch} through the use of unicast scan using *Authentication Request* frames to the limited number of preferred APs. For example, back in Figure 1, there are only six APs of which a client has connection authorization in Waseda (T4) station. A part of a possible preferred AP list regarding T3 (7 APs), T4 (6 APs), and T5 (6 APs) stations is shown in Figure 4. The information is composed of BSSID, channel number, and PHY type. In addition, other common beacon information such as capability information, SSID, supported rates, PHY parameter sets, and WPA parameters needs to be saved for the successful handoff completion.

The connection switching process is performed in the following steps. When detecting the need for switching the current link based on its policy (e.g. signal strength, transmission rate, missed beacon number, retransmission number, etc), a client looks up its preferred AP list and selects one except its currently associated AP in the same station as a target AP. Then, it sets up its interface to the desired channel and PHY type, transmits an *Authentication Request* frame to the target AP, and waits for an *Authentication Response* during *MinChannelTime*, a time parameter involved in active scan long enough to guarantee the reception of a *Probe Response* frame. This process is repeated for all the remaining APs in the list. After the next AP is selected by comparing the Received Signal Strength Indications (RSSIs) measured when receiving *Authentication Response* frames from each AP, a client executes authentication and association phase. The algorithmic flow of our system including selective passive scan and AuthScan is shown in Figure 5.

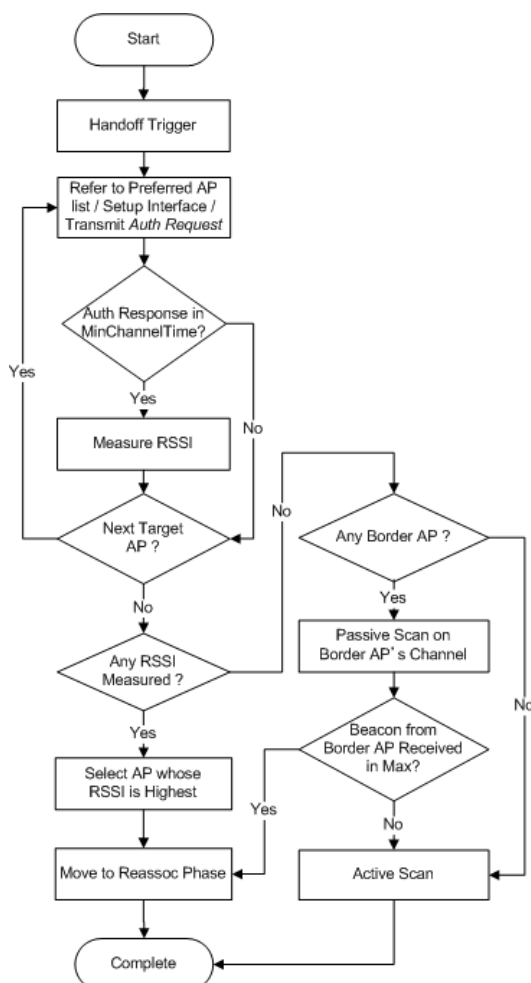


Fig. 5. The algorithmic flow of our system including AuthScan and selective passive scan

The delay of AuthScan is as follows. Assuming at least one of the target APs is available and can fulfill the handoff policy, it takes $M * RTT + (N - M) * MinChannelTime$, where N is the number of target APs that exclude the associated AP before handoff in the preferred AP list at each station, and M is the number of *Authentication Response* frames received. The total connection switching delay T_{switch} will be composed of $M * RTT + (N - M) * MinChannelTime$ due to the scanning process, one RTT for authentication, and one RTT for association phase. This achieves channel scanning with lowest delay among the approaches to work under subway's intermittent connectivity without modifying deployed APs as shown in the next section.

4. Increased Connected Time Comparison

In this section, we introduce the related work and analyze the increased connected time by our proposed method in a subway mobility scenario.

4.1 Limitations of Related Work

Many approaches have been proposed to address the channel scanning issues, and can be classified into two groups as below. The first group tries to eliminate the necessity for the channel scanning phase. Some approaches decouple the time-consuming channel scan from the actual handoff phase and scan earlier for maintaining a list of candidate APs with their handoff metrics before the connection to the current AP is terminated (Ramani & Savage, 2005; Wu et al., 2007). Other approaches enable scanning while communicating with the currently associated AP utilizing multiple NICs (Brik et al, 2005; Ok et al, 2007). Though the total handoff delay of these approaches in the first group is shorter than that of our proposed method, none of the above related work can be applied in our subway environment. (Ramani & Savage, 2005; Wu et al., 2007; Brik et al, 2005) can not generate a list of candidate APs with their handoff metrics before the connection to the current AP is terminated under subway's intermittent connectivity. (Ramani & Savage, 2005; Ok et al, 2007) require modification to APs, and (Brik et al, 2005; Ok et al, 2007) require extra hardware on a client.

The second group tries to improve the efficiency of the channel scanning phase. The first way to do this is to reduce the number of channels that are effectively going to be scanned by the client. This can be achieved through various methods such as a cache (Shin et al., 2004), Neighbor Graph (NG) (Mishra et al., 2004), sensor overlay network (Waharte et al., 2004), etc. Another way to improve the efficiency is by reducing the time waiting at each channel. In order to do this, a client is provided with an AP list, and only scans target APs in a unicast fashion to reduce the time to wait at each channel (Ok et al., 2008; Kim et al., 2004; Jeong et al., 2003; Huang et al., 2006). The performance of these approaches depend on the number of channels to scan or deployed APs, but this is not an issue in our target environment, where there are limited number of APs available. Among these, AuthScan achieves the lowest handoff delay, one RTT less than selective unicast scan (Kim et al., 2004) without modifying standard.

4.2 Performance Comparison

To show the increased connected time by our method, we estimate the interrupted time under coverage areas ($T_{establish}$ and T_{switch}) by various methods. Firstly, we compare $T_{establish}$ of passive scan and selective passive scan, which do not generate excessive management traffic under

non-coverage area in the tunnel. Assuming that a client starts to scan at the beginning of coverage area, total delay to establish wireless connection of each method is as follows.

- Passive Scan: $18^5 * Beacon\ Interval + 2 * RTT$
- Selective Passive Scan: $1 * Beacon\ Interval + 2 * RTT$

For example, in the case of T4 station where a single border AP exists on channel 1, as depicted in Figure 3, $T_{establish}$ of above two methods are compared in Table 3, where RTT is 0.6 msec, beacon interval is 100 msec.

Method	$T_{establish}$
Passive Scan	$18 * 100 + 2 * 0.6 = 1801.2\ msec$
Selective Passive Scan	$1 * 100 + 2 * 0.6 = 101.2\ msec$

Table 3. Connection establishment delay

Secondly, we compare T_{switch} of active scan, selective active scan, selective unicast scan, and AuthScan, which expedite scanning process by generating extra management traffic across overlapping coverage area at station. Assuming that there are M target APs on N channels and all of them response to requests, total delay to switch wireless connection of each methods are as follows, where $MaxChannelTime$ is a time parameter involved in active scan long enough to guarantee the reception of the *Probe Response* frames from multiple APs available in the same channel.

- Active Scan: $MaxChannelTime * N + MinChannelTime * (18 - N) + 2 * RTT$
- Selective Active Scan: $N * MaxChannelTime + 2 * RTT$
- Selective Unicast Scan: $M * RTT + 0 * MinChannelTime + 2 * RTT$
- AuthScan: $M * RTT + 0 * MinChannelTime + 1 * RTT$

For example, in the case of T4 station where five target APs exist on three channels as depicted in Figure 3, T_{switch} of above four methods are compared in Table 4, where RTT is 0.6 msec, beacon interval is 100 msec, $MaxChannelTime$ is 15 msec, and $MinChannelTime$ is 1024 μ sec (Jeong et al., 2003).

Method	T_{switch}
Active Scan	$3 * 15 + 15 * 1.024 + 2 * 0.6 = 61.56\ msec$
Selective Active Scan	$3 * 15 + 2 * 0.6 = 46.2\ msec$
Selective Unicast Scan	$5 * 0.6 + 2 * 0.6 = 4.2\ msec$
AuthScan	$5 * 0.6 + 1 * 0.6 = 3.6\ msec$

Table 4. Connection switching delay

⁵Additional 4 channels (52, 56, 60, and 64ch) in 5.3GHz (W53) and 11 channels (100, 104, 108, 112, 116, 120, 124, 128, 132, 136, and 140ch) in 5.6GHz (56W) were added to the conventional 4 channels in 5.2 GHz (52W) for 11a in 2005 and 2007, respectively. Therefore, the total number of channels to scan sums up 33 channels. However, we focus on the conventional 18 channels, which most of devices in Japan support, in this chapter.

From the two tables, we can observe that the combination of selective passive scan and AuthScan achieves the lowest interrupted connection time. The delay obtained by selective passive scan is 94.4% smaller than the one obtained by standard passive scan when establishing terminated wireless connections. The delay obtained by AuthScan is 94.2% smaller than the one obtained by active scan when switching wireless connection to the next AP. The increased connected time will become larger in proportion to the number of wireless connection establishment and switching while traveling by subway. Besides the aforementioned parameters, there are hardware induced delays such as interface setup time. These delays are not considered in the previous analysis, because they vary from maker to maker and, therefore, are unsettled. In fact, the real delay observed in experiments is even larger than the values obtained in the analysis.

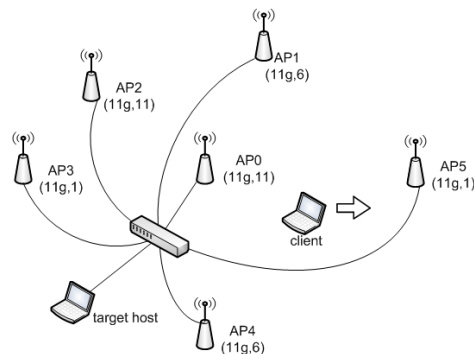


Fig. 6. Experimental setup

5. Implementation and Experiments

We have implemented a prototype of our proposed method on an IBM Thinkpad X31 (CPU Pentium M 1.7GHz, 1GB RAM) with an Atheros AR5212-based wireless interface. It runs Debian Linux 4.0 Etch with a 2.6.18-5 kernel, modified madwifi (MadWifi, [Online]) as the wireless interface driver, and modified wpa_supplicant (Linux WPA/WPA2/IEEE 802.1X Supplicant, [Online]) as the application software. We also implemented an active scan enabled client⁶ to scan all 18 channels in an active manner for a comparison purpose. Since the delay of selective passive scan in $T_{\text{establish}}$ is probabilistically determined between 0 and 100 msec as shown in Section 3, our experiments focus on the delay to switch wireless connection T_{switch} of our prototype.

In order to evaluate the performance of our prototype in an actual network, we set up the following experimental environment. We build six overlapping BSSs in an office⁷: AP0(11g, ch.11), AP1(11g, ch.6), AP2(11g, ch.11), AP3(11g, ch.1), AP4(11g, ch.6), and AP5(11g, ch.1), as described in Figure 6. The APs and a target host (IBM Thinkpad X31) are connected by 100

⁶Because of regulatory reasons, wpa_supplicant is implemented to passively scan channel 12, 13, 14, 34, 38, 42 and 46.

⁷There are 19 802.11 a/b/g APs on seven channels sharing the same medium with our experimental APs.

Base-T cable, and all APs are working as a bridge between the wireless and wired network in link layer level under open system authentication.

We evaluate our prototype's performance by measuring 1) its average delay to switch wireless connection on the application level, and 2) RTTs during handoff in comparison with active scan in the following experiment scenario. A client with an IEEE 802.11 a/b/g NIC is associated to AP0. The client moves towards AP5 while using *ping* command to transmit ICMP Echo Request frames to the target host in the same subnet. We set the ICMP frame size as 480 bytes, and the interval between frames as 10 msec. Then, we reduce the transmission power of AP0, while increasing that of AP5 to emulate a mobility scenario in a limited space. As the client gets closer to AP5, the degradation of signal strength from AP0 triggers handoff to AP5. Figure 7 shows the average delay to switch wireless connection from ten runs of the handoff scenario since the sending of the *Authentication Request* frame to the first AP scanned (2AQ in the graph). The x-axis shows the steps in the authentication scanning process. They correspond to the sending of the *Authentication Request* frame (AQ), reception of the *Authentication Response* frame (AS), sending of the *Reassociation Request* frame (RQ) and reception of the *Reassociation Response* frame (RS). The number before each of them is the actual AP name being checked. The y-axis is the time delay in milliseconds measured in the application in order to get the handoff delay from the user's perspective. We added a checkpoint right before calling the driver *ioctl*, in the case of the AQ and RQ, and right after receiving the driver's informational event for the AS and RS.

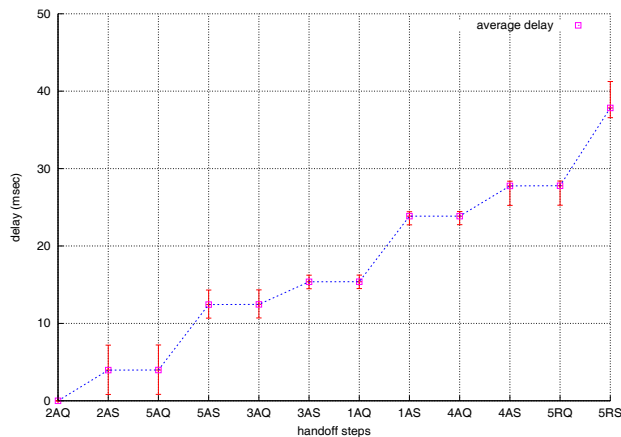


Fig. 7. The average handoff delay checking five APs

The total handoff delay in the application level obtained when checking five APs with open system authentication is 37.84 msec in average. This includes hardware induced delay such as interface setup time, delay introduced by system calls and events that flow between userland and kernel space, and 1 RTT for *Authentication Request* and *Authentication Response*. It takes 3.60 msec in average to check APs on the same channel (from AP 0 to AP2, from AP5 to AP3, from AP1 to AP4), while it takes 8.46 msec in average to check APs on the different channel (from AP 2 to AP5, from AP3 to AP1). Therefore, the AuthScan client saves 4.86 msec to setup the interface into the desired channel, each time checking APs on the same channel consecutively in our experiment.

Figure 8 shows one example of the way RTT changes during handoff from AP0 to AP5. The upper graph corresponds to one run by an AuthScan client, and the lower by an active scan client. The x-axis shows the ICMP sequence number and the y-axis shows the RTT in milliseconds. Handoff takes place between the 682nd and 686th frames (A) in the case of the AuthScan client (i.e., 3 frames dropped, 30 msec disrupted), and between the 3637th and 3924th frames (B) for the active scan client (i.e., 286 frames dropped, 2860 msec disrupted). Therefore, the AuthScan client drops approximately 98.95% fewer frames than the active scan client in the experiment.

6. Conclusion

In order to increase the time connected to already deployed 802.11 wireless networks while traveling by subway in Tokyo, we have developed a system equipped with two scanning modes: 1) passively scanning on a selected channel, and 2) scanning with multiple open authentication. Through analysis and experiments, we have shown that our method increases the time connected to 802.11 wireless networks by establishing wireless connection when coming out of non-coverage area in the tunnel and switching its wireless connection across overlapping coverage area at station with less delay.

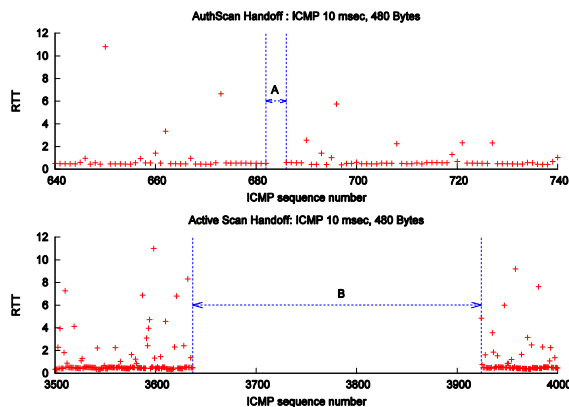


Fig. 8. Profile of RTTs during handoff from AP5 to AP0

The main contribution of this chapter is two-fold:

- We investigated the commercial 802.11 HOTSPOT wireless networks deployed in Tokyo Metro, and clarified the main factor to the diminishment of available connected time.
- We proposed an optimized solution for the subway's intermittent connectivity environment, and analyzed the increase connected time by our method. In addition, we showed the effectiveness of our system through experiments in comparison with standard active scan.

Our proposed method will work under similar subway 802.11 wireless network environments in any other cities. Our future efforts will be oriented to build a more sophisticated chain of border APs, and list of preferred APs. A chain of border APs including interrupted time under non-coverage area in the tunnels can save power by sleeping the interface before performing selective passive scan. A list of preferred APs built per AP at each station can save unicast scanning time even further.

7. References

- Brik, V.; Mishra, A. & Banerjee, S. (2005). Eliminating handoff latencies in 802.11 WLANs using multiple radios: Applications, experience, and evaluation, *Proceedings of ACM/USENIX IMC 2005*, Berkeley, CA, October 2005
- DownThemAll. [Online].
Available: <http://www.downthemall.net/>
- Droms, R. (1997). Dynamic Host Configuration Protocol, RFC 2131, Internet Society
- Huang, P.; Tseng, Y. & Tsai, K. (2006). A Fast Handoff Mechanism for IEEE 802.11 and IAPP Networks, *Proceedings of IEEE VTC 2006-Spring*, Melbourne, Australia, May 2006
- IEEE Standard 802.11. (1999). IEEE. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications
- IEEE Standard 802.1Q-2003. (2003) IEEE Standards for Local and Metropolitan Area Networks, Virtual Bridged Local Area Networks
- Jeong, M.; Watanabe, F. & Kawahara T. (2003). Fast active scan for measurement and handoff, Technical report, DoCoMo USA Labs, Contribution to IEEE 802, May 2003
- Kim, H.; Park, S.; Park, C.; Kim, J. & Ko, S. (2004). Selective Channel Scanning for Fast Handoff in Wireless LAN using Neighbor Graph, *Proceedings of ITC-CSCC 2004*, Sendai, JAPAN, July 2004
- Linux WPA/WPA2/IEEE 802.1X Supplicant. [Online].
Available: <http://hostap.epitest.fi/>
- MadWifi - a Linux kernel device driver for Wireless LAN chipsets from Atheros. [Online].
Available: <http://madwifi.org/>
- Microsoft Technet. [Online].
Available: <http://technet.microsoft.com/en-us/library/cc757419.aspx>
- Mishra, A; Shin, M. & Arbaugh, W. (2003). An Empirical Analysis of the IEEE 802.11 MAC Layer Handoff Process, *ACM SIGCOMM Computer Communication Review*, Vol. 3, April 2003, pp. 93-102
- Mishra, A.; Shin, M. & Arbaugh, W. (2004). Context Caching using Neighbor Graphs for Fast Handoffs in a Wireless Network, *Proceedings of IEEE INFOCOM 2004*, Hong Kong, March 2004
- NTT Communications HOTSPOT. [Online].
Available: <http://www.hotspot.ne.jp/>
- NTT DoCoMo Mzone. [Online].
Available: <http://www.nttdocomo.co.jp/service/data/mzone/>
- NTT EAST FLET'S SPOT. [Online].
Available: <http://flets.com/spot/>
- NetStumbler. [Online].
Available: <http://www.netstumbler.com/>
- Ok, J.; Morales, P.; Darmawan, A. & Morikawa, H. (2007). Using Shared Beacon Channel for Fast Handoff in IEEE 802.11 Wireless Networks, *Proceedings of IEEE VTC2007-Spring*, Dublin, Ireland, April 2007
- Ok, J.; Morales, P. & Morikawa, H. (2008). AuthScan: Enabling Fast Handoff across Already Deployed IEEE 802.11 Wireless Networks Mobility, *Proceedings of IEEE PIMRC 2008*, Cannes, France, September 2008
- Ramani, I. & Savage, S. (2005). SyncScan: Practical Fast Handoff for 802.11 Infrastructure Networks, *Proceedings of IEEE INFOCOM 2005*, Miami, FL, March 2005

- Shin, S.; Forte, A. G.; Rawat, A. S. & Schulzrinne, H. (2004). Reducing MAC layer handoff latency in IEEE 802.11 wireless LANs, *Proceedings of MobWiac 2004*, Philadelphia, PA, September 2004
- Tokyo Metro. [Online].
Available: <http://www.tokyometro.jp/global/en/about/outline.html>
- Waharte, S.; Ritzenthaler, K. & Boutaba, R. (2004). Selective Active Scanning for Fast Handoff in WLAN using Sensor Networks, *Proceedings of MWCN 2004*, Paris, France, October 2004
- Wireshark. [Online].
Available: <http://www.wireshark.org/>
- Wu, H.; Tan K.; Zhang, Y. & Zhang, Q. (2007). Proactive Scan: Fast Handoff with Smart Triggers for 802.11 Wireless LAN, *Proceedings of IEEE INFOCOM 2007*, Anchorage, Alaska, May 2007

Asymmetric carrier sense in heterogeneous medical networks environment

Bin Zhen, Huan-Bang Li, Shinsuke Hara[†] and Ryuji Kohno^{††}

*National Institute of Information and Communications Technology, 3-4, Hikarino-oka,
Yokosuka, 239-0847, Japan*

[†]Osaka City University, 3-3-138 Sugimoto, Osaka, 530-0001 Japan

^{††}Yokohama National University, 79-5 Tokiwadai, Yokohama, 240-8501, Japan

Summary: Complementary WLAN and WPAN technologies, as well as other wireless technologies will play a fundamental role to support ubiquitous healthcare delivery. This chapter investigates energy based clear channel assessment (CCA) of IEEE WLAN (802.11b) and WPAN (802.15.4b) system when they coexist in a close space. We derive closed-form expressions of energy based, qualify the asymmetric CCA in both AWGN channel and fading channels, and show the impact of noise uncertainty on CCA operation. In the heterogeneous medical networks environment, WPAN is oversensitive to the 802.11b signals and WLAN is insensitive to the 802.15.4b signals. The asymmetric CCA issue in heterogeneous networks is different from the traditional “hidden node” or “exposed node” issues in homogeneous network. Energy based CCA can effectively avoid possible packet collisions when they are close within the “heterogeneous exclusive CCA range”. However, beyond this range, WPAN can still sense 802.11b signals, but WLAN lose its sense to 802.15.4b signals. This leads to WPAN traffic in a position secondary to the WLAN traffic. A two-band CCA scheme, with an additional CCA detector in auxiliary channel, is proposed to combat the asymmetric CCA issue in the heterogeneous networks.

1. Introduction

Integration of heterogeneous wireless technologies is required to for revolutionary healthcare delivery in hospital, small clinic, residential care center, and home [1-4]. The medical environment is a diverse workspace, which encompasses everything from the patient admission process, to examination, diagnosis, therapy, and management of all these procedures. The concept of “wireless hospital” combines all medical, diagnostic and clinical data together whenever needed through wireless integration [4]. There is desire to use IEEE version of wireless local area networks (WLAN) and wireless personal area networks (WPAN) technologies in the unlicensed industrial, scientific and medical (ISM) bands as a common communication infrastructure [2, 3]. The WLAN technology is typically used for office oriented applications and patient connection to the outside world, while the WPAN

technology is usually used for wearable sensors around patients to collect vital information for ubiquitous healthcare service [2-6].

The use of complementary heterogeneous WLANs and WPANs in the shared ISM band results in coexistence, interference and spectrum utilization issues. The coexistence of wireless technologies in ISM band has been a hot topic [6-9]. Adaptive frequency hopping was proposed for Bluetooth devices to avoid interference from WLAN [7]. A model for analyzing the effect of 802.15.4 on 802.11b performance was provided by Howitt and Gutierrez [8]. The degradation of WLAN performance is small given that the WPAN activity is low. However, the high duty cycle of WLAN traffic can drastically affect the WPAN performance [6]. A distributed adaptation strategy for WPAN based on Q-learning has been proposed to minimize the impact of interference from 802.11b [9]. However, the spatial reuse issue in the heterogeneous networks has not drawn much attention. Some researches have shown that the spatial reuse and aggregate throughput in the homogeneous WLAN mesh network is closely related to physical channel sensing. Yang and Vaidya showed that the aggregate throughput can suffer significant loss with an inappropriate choice of carrier sense threshold [10]. Ma *et al*, by means of Markov chain model, evaluated how carrier sense threshold affects the throughput and packet collision [11]. Zhai and Fang found that the optimal carrier sensing threshold for one-hop flows does not work for multihop flows [12]. Zhu *et al* reported that a tunable sensing threshold can effectively leverage the spatial reuse and demonstrated it through testbed measurement [13, 14]. In [15], Zhu *et al* proposed a heuristic algorithm to adaptively tune the threshold of carrier sense to enhance throughput per user. Jamieson found carrier sense can be inefficient at low data rate when capture effect is most prevalent [16]. Ramachandran and Roy showed cross-layer dependence between carrier sense and system performance [17]. Simulators widely used for performance evaluation, like NS-2 and OPNET, do not contain detailed physical layer module like carrier sense. For the lack of carrier sensing knowledge between WPAN and WLAN, Golmie *et al* simply simulated two carrier sensing cases: the WPAN can only detect packets of its own type and the WPAN can also detect WLAN's transmission, in their coexistence study for medical applications [6].

In this chapter we study the coexistence issue in the heterogeneous medical networks environment from carrier sensing point. The remainder of the chapter is organized as follows. Section II briefly reviews various carrier sense methods and the considered WLAN and WPAN systems. In section III, a mathematical analysis of energy based carrier sense in both AWGN channel and fading channel is presented. Section IV describes the impact of asymmetric carrier sense in heterogeneous networks environment and presents a two-band carrier sense to combat the asymmetry. Section V finally concludes the chapter.

2. Review of systems and clear channel assessment

A. Wireless medical sensor networks

Wireless medical sensor networks can be considered as a special part of general wireless sensor networks (WSN), which are mainly implemented by low-rate WPAN technologies. As compared in Table I, both share some common features which include limited resources (e.g. computation power, memory, battery, bandwidth), low/modest duty cycle, energy efficiency, plug-and-play, diverse coexistence environments, and heterogeneous device

ability. But we can also find significant differences between them in the sensor device, dependability, networking, traffic pattern and channel.

Firstly medical sensors consider safety, quality and reliability as top priority, while general WSN are cost sensitive for market reason. The safety to human/animal body is therefore the first factor taken into considered. Thus medical sensors must be conscious of specific absorption ratio (SAR) to protect human tissue. Wearable IEEE WPAN devices are suggested to be separated at least 30cm distance from human body. Safe to human is the top priority of medical sensors. The radio emission should be as weak as possible. And medical sensors should be lightweight and small to achieve non-invasive and unobtrusive monitor. This limits the available resource which includes memory, battery power and computation ability in the medical sensors. The requirement is more stringent than the general WSN.

Secondly the medical sensor networks have more frequency bands to select than general WSN, which usually work in ISM band. Although the specific medical bands are less noisy, they are narrow band and conditional license. For example, the wireless medical telemetry service (WMTS) band can only be used in the licensed hospital and clinic, but not at home. On the contrary, the wideband ubiquitous ISM is somehow noisy since the frequency band should be shared with other systems.

Thirdly, the traffic pattern in medical sensor networks is featured by periodical real time data (e.g. EEG and ECG) and some top priority burst data (e.g. alarm and alert) [17]. In contrast, general WSN typically consider versatile traffic. The medical information, especially the alarm notification, have very strict requirement in terms of Quality-of-Service (QoS) since they are life critical. The transmission of vital signal has a life or dead meaning. This means more stringent QoS requirement than general WSN.

Fourthly, security of data is traditionally utmost important. Patient data needs to be protected in all stages of data acquiring, data transmission and data storage. It therefore importance to secure data at physical layer and MAC lay. However, security is not free, extensive sources are needed to secure data at the link layer. In the resource limited WSN, security becomes an overhead of existing network QoS. Because both of them are paramount in the healthcare service, the balance of security and QoS is a new issue. The general WSN do not require strict QoS and security simultaneously.

Fifthly, to improve reliability, general WSN tend to distribute redundant sensors as backup for sensing, transmission and forwarding. In contrast, there is little redundancy in medical WSN for medical reasons. For example, vital signals, like EEG (Electroencephalography) and ECG (Electrocardiogram), are location dependent and can only be measured by deterministic location. Therefore it is difficult to allocate redundant sensors in the limited area. Especially, it makes no sense to allocate sensors outside of the interest/effect area.

In summary, the lack of redundancy, priority traffic, dominant periodical data and balance of guaranteed QoS and security in versatile coexistence environment challenge the reliability design of wireless medical sensor networks.

	Medical wireless sensor networks	General wireless sensor networks
Common features	Limited resources: battery, computation, memory, energy efficiency Diversity coexistence environment low/modest data rate, low/modest duty cycle Dynamic network scale, plug-and-play, heterogeneous devices ability, dense distribution	
Sensor/ actuator	Single-function device Fast relative movement in small range device lifetime, days, <10 years (implant sensor) Safe (low SAR) and quality first	Multi-function device Rare or slow movement in large range network lifetime and device lifetime, months, <10 years Cost sensitive
Dependability	Reliability (first), guaranteed QoS Strongly security (except emergency)	expected QoS, redundancy-based reliability Required security
Networking	Small scale star network No redundancy in device Deterministic node distribution	Large scale hierarchical network redundant distribution Random node distribution
Traffic	Periodical RT (dominant), burst (priority) Uni-directional traffic M:1 communication	Burst (dominant), periodical Uni-directional or bi-directional traffic M:1 or point-point communication
channel	Specific medical channel, ISM band Body surface or through body	ISM band Obstacle is unknown

Table 1. Comparison between wireless medical sensor network and general wireless sensor networks

B. Systems overview

We consider IEEE 802.15.4b and 802.11b as examples for the ubiquitous medical services [2, 3, 6]. The former is a good candidate technology for low data rate and low cost medical sensors. Both systems operate in the unlicensed 2.4GHz ISM band, and both are based on carrier sense multiple access with collision avoidance (CSMA/CA) protocol. Carrier sense is more generally known as clear channel assessment (CCA) in the standards. The physical CCA can be either energy based, or feature based, or a combination of two. As shown in Fig. 1, there are only 4 WPAN channels locate in the guard bands of WLAN. Table II lists the parameters of both systems [18, 19]. The bit error rate (BER) of WLAN systems with AWGN channel is give by [18]

$$BER_{11b} = \frac{128}{255} \times \sum_{l=1}^6 M1_l \times \sqrt{Q(M2_l \times SNR)}, \quad (1)$$

where SNR is the signal-to-noise ratio, $Q(\cdot)$ is Gaussian Q-function, $M1=[24 \ 16 \ 174 \ 16 \ 24 \ 1]$, and $M2=[4 \ 6 \ 8 \ 10 \ 24 \ 16]$. The BER of WPAN systems is given by [19]

$$BER_{15.4b} = \frac{8}{15} \times \frac{1}{16} \times \sum_{k=2}^{16} -1^k \binom{16}{k} e^{20 \cdot SNR(1/k-1)}. \quad (2)$$

The radio path loss for indoor channels in the working frequency band is given by

$$\begin{cases} pl = 40.2 + 20 \log_{10} d & d \leq 8 \\ pl = 58.5 + 33 \log_{10} (d/8) & d > 8' \end{cases} \quad (3)$$

where d is separation distance between transmitter and receiver.

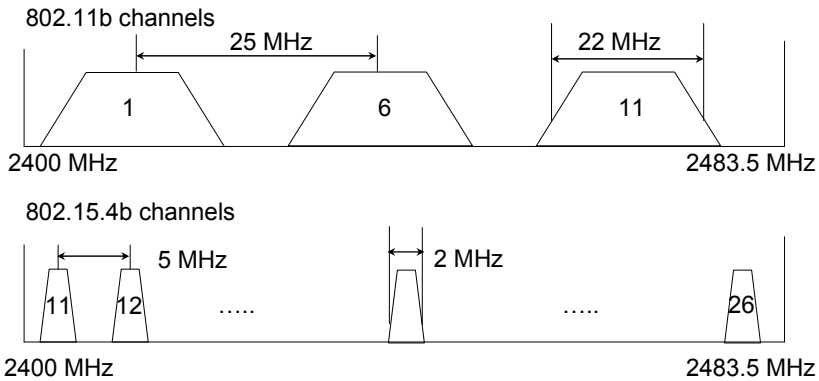


Fig. 1. Channel plan for IEEE 802.11b and 802.15.4b in 2.4GHz ISM band

Parameters	802.15.4b	802.11b
Transmission power (dBm)	0	16
Channel bandwidth (MHz)	2	22
Adjacent channel separation (MHz) ¹	5	25
Background noise (dBm) ²	-94.9	-84.6
Spread code (chips)	32/4 bits	11
Data rate (Mbps)	0.25	11
CCA window (μ s)	120	15

Table 2. System parameters of IEEE WLAN and WPAN

C. Clear channel assessment

Several IEEE WLAN and WPAN standards adopt CSMA protocol for channel access. CCA is a physical layer activity and is an essential element of the CSMA protocol. The concept of CCA was first proposed as an enhancement to the ALOHA protocol. The CCA detects an incoming packet and ensure a free medium before transmission. The CCA module processes received radio signals in a suitable time duration termed CCA window. It then reports the medium state, either busy or idle, by comparing the detection with a threshold. The energy based CCA integrates signal strength from radio front end during the CCA window. The feature based CCA looks for the known features, *e.g.* the modulation and spreading characteristics, of the signal over the channel. Modulated signals are in general coupled with sine wave carriers, pulse trains, repeating spreading, or cyclic prefixes, which result in built-in periodicity. This periodicity can be used to detect signal of a particular modulation type.

The feature based CCA performs far better than the energy based CCA. However, a prior knowledge of the signal characteristic is necessary. And the CCA module would need a dedicated detector for every potential coexistence signal class. The main advantages of energy based CCA are its simplicity, generality, and low power consumption. It is a universal mechanism that can be deployed in all systems. Unlike feature based CCA, there is no need for waiting time for the specific features of the signal and synchronization [17]. The downside of energy based CCA is that it is prone to false detection.

We applied energy based CCA to 802.15.4b and 802.11b systems to deliver ubiquitous healthcare service in this heterogeneous networks environment. There are several reasons for this. First, there have been nearly 10 wireless technologies with different modulations, band plans, and transmission powers in the 2.4GHz ISM bands due to its global availability.

¹ Distance between the central frequencies of non-overlapped adjacent channel.

² We assumed -174dBm/MHz thermal noise, 8dB implementation losses, and 8dB radio noise figure.

These include WLAN (802.11, 802.11b, 802.11g and 802.11n), WPAN (Bluetooth, 802.15.3, 802.15.4b, and chirp spread spectrum PHY of 802.15.4a), and passive radio frequency identification *etc.* Significantly, some medical equipments like electric surgical knife, magnetic resonance imaging (MRI), heat treatment machines, and microwave ovens also use this frequency band. Different from the communication device, the medical equipments emit radio signals unintentional. All the communication device and medical equipments are expected to be collocated in the integrated medical environments. Secondly, a medical sensor based on 802.15.4b is unlikely to have all the knowledge. And feature based CCA is usually complex and power-hungry due to the waiting time and synchronization [17].

3. Energy based clear channel assessment in heterogeneous networks

In mathematics, CCA is a test of two hypotheses:

$$\begin{cases} H_0 : y[n] = w[n] & \text{signal absent} \\ H_1 : y[n] = x[n] + w[n] & \text{signal present} \end{cases} \quad (4)$$

where $x[n]$ is the targeted signal: $w[n]$ is the white Gaussian noise with variance σ^2 ; and $n=1, \dots, N$ is the sample index in total N independent samples in the CCA window. Under common detection performance criteria, *e.g.* Neyman-Pearson (NP) criteria, likelihood ratio yields the optimal hypothesis testing solution. The CCA metric is compared to a threshold Γ to make a decision. CCA performance is characterized by a resulting pair of detection and false alarm probabilities, (P_d and P_{fa}), which are associated with the particular threshold Γ .

For simplicity, we assume the energy based CCA is realized by a simple non-coherent module that integrates the square of the received signal and sums its samples in analog or digital domain. In particular, the energy detection consists of a quadrature receiver with y_I and y_Q representing samples of signals on the I (in-phase) and Q (quadrature) branches respectively. Figure 2 depicts a block diagram of energy based CCA.

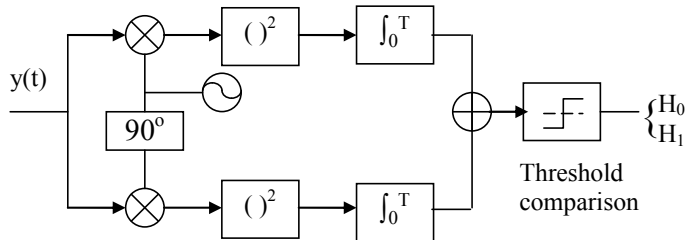


Fig. 2. Block diagram of energy based CCA detector

The energy based CCA metric can be given by

$$Y = \frac{1}{N} \sum_{n=1}^N (|y_I[n]|^2 + |y_Q[n]|^2), \quad (5)$$

where N is the number of independent samples in the CCA window. In an AWGN channel, each $|y_I[n]|$ and $|y_Q[n]|$ has a normal distribution with mean μ and variance σ^2 and Y can be evaluated as generalized chi-square function $Y \sim \chi^2(\lambda, 2N)$, where $2N$ is the degrees of freedom and $\lambda = \sigma^2 + \mu^2$. Under the H_0 hypothesis, each normal distribution has $\mu=0$. Thus, Y has a χ^2 distribution. Under the H_1 hypothesis in the present of signal with an signal-to-noise ratio (SNR) $\gamma = \frac{\mu^2}{\sigma^2}$, Y has a non-central χ^2 distribution. We have mean and variance as follows [20, 21]

$$\begin{cases} H_0: & \mu_0 = \sigma^2, \quad \sigma_0^2 = \frac{2}{N} \sigma^4 \\ H_1: & \mu_1 = \mu^2 + \sigma^2, \quad \sigma_1^2 = \frac{2}{N} (2\mu^2 + \sigma^4) \end{cases} \quad (6)$$

When N is large, using central limit theory, the energy based CCA metric in Eq. (4) can be approximated as Gaussian random process. Then P_d and P_{fa} can be expressed in terms of Gaussian Q-function

$$P_d = Q\left(\frac{\frac{\Gamma}{\sigma^2} - (1 + \gamma)}{\sqrt{\frac{2}{N}(1 + 2\gamma)}}\right), \quad P_{fa} = Q\left(\frac{\frac{\Gamma}{\sigma^2} - 1}{\sqrt{\frac{2}{N}}}\right), \quad (7)$$

where Γ is the decision threshold. Eq. (7) clearly shows that the decision threshold is related with noise level. If the noise is completely known, eliminating Γ in Eq. (7) gives

$$N = 2 \left(\frac{Q^{-1}(P_{fa}) - Q^{-1}(P_d) \sqrt{1 + 2\gamma}}{\gamma} \right)^2, \quad (8)$$

where $Q^{-1}()$ is inverse Gaussian Q-function. The energy based CCA can meet any desired P_d and P_{fa} simultaneously by increasing the number of samples in the CCA window N . Given a limited N , the CCA ability is obviously determined by the SNR of the signal and noise variance σ^2 . There is an inherent tradeoff between P_d and P_{fa} . We define the CCA error floor at the optimal threshold, which can be found by equating $1 - P_d$ and P_{fa} . Using Eq. (7), we obtain the CCA error floor

$$P_{CCA_ef} = Q\left(\sqrt{N} \frac{\gamma}{1 + \sqrt{1 + 2\gamma}}\right). \quad (9)$$

Note that the error floor depends on the number of symbol chips and the SNR. When $\text{SNR} \ll 1$, Eq. (9) can be approximated as

$$P_{\text{CCA}_{ef}} = Q\left(\sqrt{N} \frac{\gamma}{2}\right). \quad (10)$$

A linear decrease in SNR requires a quadratic increase in N to maintain the same error floor.

A. Asymmetric energy based CCA

Table III lists the numbers of signal chips in the CCA windows of IEEE WLAN and WPAN. Figure 3 shows the CCA error floor in the heterogeneous networks environment when the noise is known. Per Eq. (9), the error floors decrease with increment in signal chips in the CCA window. Given the defined CCA windows, the CCA abilities, *e.g.* sensitivity and range, to determine the channel state are different, this is termed asymmetric CCA. Under the same SNR conditions, the lowest error floor is where WPAN is used to sense 802.11b signals; the highest error floor is where WLAN is used to sense 802.15.4b signals. The performance difference is nearly 10dB.

The CCA asymmetry can be attributed to differences in the underlying signals over channel (power, symbol rate and background noise) and CCA window. In physics, a higher data rate and a longer CCA window means more signal pulses in baseband can be collected. Better CCA performance is a natural result. Asymmetric CCA can be further reinforced by other factors. For example, the difference in transmission powers which is usually stronger for WLAN, and the difference in channel bandwidth, which are 22MHz and 2MHz for the WLAN and WPAN, respectively. For both WPAN and WLAN, the performances to detect the signal of its own type are similar. There is not big difference in the numbers of symbols in the CCA window.

Table IV compares communication with CCA when both have an error probability of 1% with AWGN channel. As expected, the CCA range is larger than the communication range. For WPAN, sensing 802.11b signals has 4dB greater link margin compared to sensing the signals of its own type. In contrast, for WLAN sensing 802.15.4b signals requires a 4.8dB higher SNR.

Asymmetric CCA makes channel sensing insensitive or oversensitive to other signals in the mixed WLAN and WPAN environment. The asymmetric CCA in the heterogeneous networks is different from the traditional “hidden node” or “exposed node” issues in the homogeneous network. In the homogeneous network, two devices belong to the same system are reciprocal in ability to sense each other (we do not consider the minor difference due to implementation.). However, in the heterogeneous networks, the sensing abilities of different systems are unequal and depend on the underlying signals over channel and the separation distances. As shown in Fig. 3, WLAN signals are well sensed by both of them, but WPAN signals could be ignored by the WLAN systems when they are separated by enough space.

<div>Sensed signals</div> <div>Device</div>	802.15.4b signals	802.11b signals
802.15.4b	32*8	120*11
802.11b	15*2	15*11

Table 3. Number of signal chips in the CCA window

<div>Devices</div> <div>Signal</div>	802.15.4b	802.11b
Communication	-0.8	5.6
802.15.4b CCA	-3.2	2.6
802.11b CCA	-7.2	-2.2

Table 4. SNRs (dB) to achieve 1‰ communication BER and CCA error floors in AWGN channel

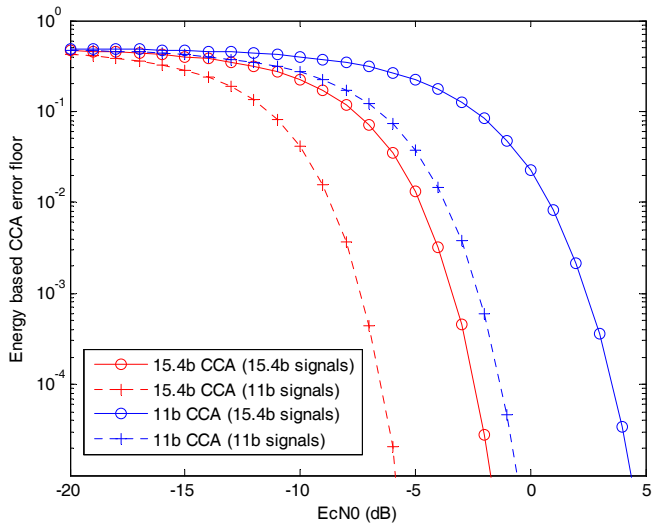


Fig. 3. Error floor of energy based CCA with AWGN channel in heterogeneous networks

Usually, NP criteria is adopted in CCA because a miss detection of a busy channel is riskier than a false alarm of a free channel. Eq. (7) can be re-written as

$$P_d = Q\left(\frac{Q^{-1}(P_{fa}) - \sqrt{2N}\gamma}{\sqrt{1+2\gamma}}\right). \quad (11)$$

As expected, P_f is independent of γ since there is no signal under H_0 . When the channel is varying due to fading and shadowing, Eq. (12) gives a CCA performance conditioned on the instantaneous SNR. The average CCA performance can be derived by averaging Eq. (11) over fading statistics

$$\overline{P_d} = \int_0^\infty Q\left(\frac{Q^{-1}(P_{fa}) - \sqrt{2N}\gamma}{\sqrt{1+2\gamma}}\right) f(\gamma) d\gamma, \quad (12)$$

where $f(\gamma)$ is the probability of distribution function (PDF) of SNR under fading.

The medium-scale variance of SNR can be characterized by log-normal distribution [22]. The log-normal shadowing is usually described in-term of its dB-spread, σ_{dB} , which is related to σ by

$$\sigma = \sigma_{dB} \ln(10)/10. \quad (13)$$

Under Rayleigh fading, the SNR γ has an exponential PDF

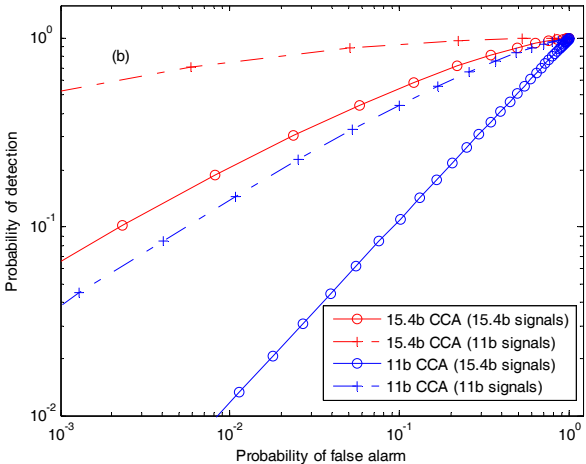
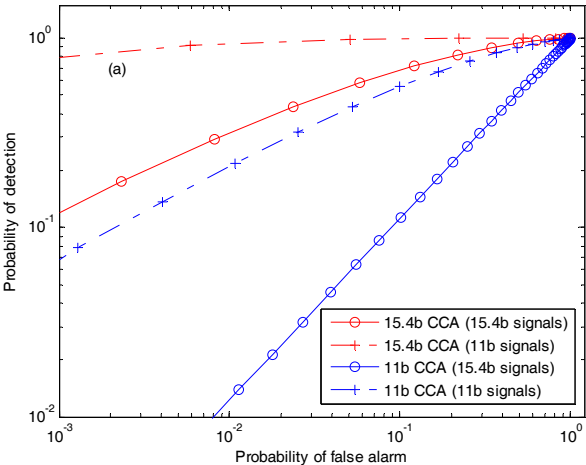
$$f(\gamma) = \frac{1}{\gamma} \exp\left(-\frac{\gamma}{\bar{\gamma}}\right) \quad \gamma \geq 0, \quad (14)$$

where $\bar{\gamma}$ denotes to average SNR. If the SNR follows a Rician distribution, the PDF of γ becomes

$$f(\gamma) = \frac{K+1}{\bar{\gamma}} \exp\left(-K - \frac{(K+1)\gamma}{\bar{\gamma}}\right) I_0\left(2\sqrt{\frac{K(K+1)\gamma}{\bar{\gamma}}}\right) \quad \gamma \geq 0, \quad (15)$$

where K is the Rician factor and $I_0(\cdot)$ is the modified Bessel function with order zero. Because it is difficult to have close-form expressions of Eq. (12) over fading channels, we evaluated them numerically in this chapter.

Figure 4 plots ROCs of energy based CCA over AWGN, log-normal shadowing, Rayleigh fading, and Rician fading channels. The asymmetric CCA abilities of WPAN and WLAN remain the same in fading channels. WLAN systems are insensitive to WPAN signals, while WPAN systems are oversensitive to WLAN signals. Comparing with the AWGN curves, we observe that channel fading degrades the performance of energy based CCA, and the degradations are closely related with the CCA parameters and SNR. In other words, meeting the desired performance demands a longer CCA window. Especially Rayleigh fading and Rician fading degrade the CCA performance of all systems significantly.



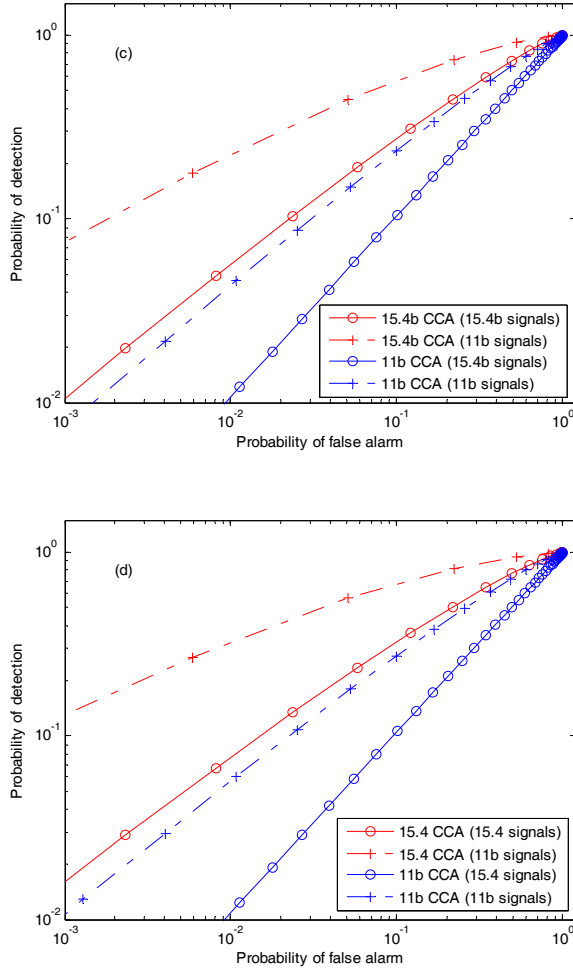


Fig. 4. ROC of energy based CCA in the case of SNR=-9.5 dB measured by WPAN over a) AWGN channel; b) log-normal shadowing ($\sigma_{dB}=6$ dB) channel; (c) Rayleigh fading channel, and d) Rician fading ($K=1.5$) channel.

B. Noise uncertainty

NP criteria requires a desired P_{fa} to determine the decision threshold. We can re-write Eq. (7) to have

$$\Gamma = \sigma^2 \left(1 + Q^{-1}(P_{fa}) \sqrt{\frac{2}{N}} \right)^2 \quad (16)$$

which shows the desired threshold should be proportional to the noise energy. In practice, the noise power cannot be estimated accurately because of noise uncertainty [23].

The background noise power fluctuates from time to time due to changes of environment and mobility of device. Another source of uncertainty is the error in quantization which is usually implemented by A/D converter. Assume the noise estimation is expressed as

$$\hat{\sigma}^2 = k\sigma^2, \quad k \geq 0. \quad (17)$$

When $0 < k < 1$, the noise is underestimated; when $k > 1$, it is overestimated. The Γ biases to the desired value due to error in noise estimation. An overestimation of noise decreases P_{fa} at expense of boosting P_d . It is *vice versa* for an underestimation of noise level. We can obtain the total CCA error by $1-P_d+P_{fa}$ using Eq. (7). Figure 5 shows the impact of noise uncertainty with 3 dB error. We used an 802.15.4 device to sense the 802.15.4 signals. The x-axis is the true SNR condition of CCA. As expected, both overestimation and underestimation deteriorate the CCA performance because of an un-optimal threshold. As shown in Fig. 5 and other numeric results, the performance loss of energy based CCA can

be approximated as
$$\begin{cases} 10 * \lg(k) & 0 < k \leq 1 \\ -10 * \lg(k) & k > 1 \end{cases}$$
 in both cases when SNR is high. In

practical, we are usually more interested in noise underestimation. In order to guarantee the P_{fa} the Γ is purposely biased. This increases the probability of miss detection of CCA and therefore the probability of packet collision. In other words, the noise level estimation of free channel also plays an important role in the CCA operation.

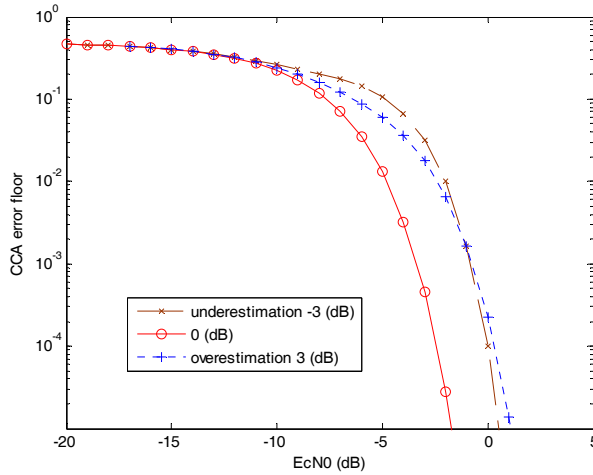


Fig. 5. Total error of energy based CCA with AWGN channel in the case of a 3dB noise uncertainty

4. Two-band clear channel assessment

A. Impact of asymmetric CCA in heterogeneous networks

Table V lists the required minimum SNRs and their corresponding distances to achieve reliable energy based CCA ($P_{FA} < 1\%$ and $P_D > 90\%$) over AWGN channel. The corresponding distances were computed using Eq. 1 to Eq. 3 and the parameters listed in Table I. For WPAN, the sensing of 802.11b signals is reliable at an SNR as low as -9.25dB. This SNR is 9.65dB lower than the critical SNR which is the least SNR to achieve BER < 0.1% for communication. The CCA range is 180 meters longer than the communication range. In contrast, sensing 802.15.4b signal by WLAN requires a high SNR up to 9.75dB, which is 3.15dB more than the critical SNR for communication. The CCA range is 42 meters shorter than the communication range. In the fading channels, all distances decreases depending on the fading condition.

We can define a “heterogeneous exclusive CCA range” (HECR), in which systems in the heterogeneous environment can reliably sense the activities of each other. In the considered scenario, the HECR is the maximum distance that WLAN can sense 802.15.4b signals. Given the system parameters and assumptions, the HECR for IEEE WLAN and WPAN is 25m in AWGN channel. Peaceful and fair coexistence between them can be expected when they are located within the HECR. However, it becomes different when they are separated beyond the HECR. For WPAN systems, the CCA range of WLAN signals is more than twice as long as the communication range. And it is longer than the CCA range of its own signal type. That is, the WPAN is oversensitive to the WLAN signals. It can even sense a WLAN packet that is outside of the keep-out range of receiver in the worst case.³ Although the oversensitive CCA avoids the ‘hidden node’ issue, it suffers from the ‘exposed node’ issue. This results in poor spatial reuse of frequency channels and low aggregation throughput since WPAN sometimes unnecessarily withdraw packet before transmission. As simulated in [11], the threshold optimized to maximize aggregate throughput is higher than the optimal threshold for a single hop. For WLAN systems, the CCA range of WPAN signals is about a quarter of the communication range. Packet collision may occur when WLAN traffic occurs immediately after the WPAN traffic.

	802.11b CCA		802.15.4b CCA	
	WPAN signals	WLAN signals	WPAN signals	WLAN signals
SNR (dB)	9.75	-4.5	-6	-9.25
Distance (m)	25	200	155	280

Table 5. Minimum SNRs (dB) and corresponding distances (m) to achieve CCA ($P_{FA} < 1\%$, $P_D > 90\%$) with AWGN channel

³ The keep-out range denotes to the minimum separation which WPAN and WLAN do not interfere each other.

Although the HECR of 25 meters is not sufficient for outdoor applications, it seems to be good enough for most indoor applications. Typical bedside medical applications define by IEEE 1073 are within this range [24]. This is different from those of most coexistence studies in which it is usually assumed that WLAN cannot sense the activities of WPAN [6, 8, 9]. The HECR over fading channels is expected to be shorter than 25 meters depending on the fading parameters. But the asymmetric CCA issue still exists. Putting the oversensitive and insensitive CCAs together results in an unfair share of channel between WLAN and WPAN when they are separated beyond HECR. There is a preferential treatment of WLAN traffic. The WLAN is over-protected, while the WPAN is vulnerable.

B. Two-band clear channel assessment

The asymmetric CCA issue in heterogeneous medical networks environment must be solved. It is needed to recognize the signal source, WPAN signals or WLAN signals, when the medium is busy. When the channel is occupied by WLAN signals, an 802.15.4b device should increase its CCA threshold to lower the CCA sensitivity. On the other hands, when the channel is busy in WPAN signals, an 802.11b device should lower its CCA threshold to heighten the CCA sensitivity.

Figure 6 illustrates the mechanism of a two-band CCA. The 22 MHz WLAN channel depicted in dash-dotted line and the 2 MHz WPAN channel depicted in dotted line overlap each other. There are about 3 MHz space between adjacent WPAN channels, which are inside the WLAN channel. As shown in red-solid line in Fig. 7, we termed it auxiliary channel in this chapter. We added a new energy based CCA detector in the auxiliary channel to provide additional information. For example, the auxiliary CCA detector can be the same as the 802.15.4b CCA detector except that the working frequency is tuned to the auxiliary channel. For both WLAN systems and WPAN systems, there are two energy based CCA detectors tuned in its original channel and in the auxiliary channel, respectively. The two detectors have the same ability to conduct carrier sensing given the same configuration. Besides, there is a little increase in the complexity of device and the power consumption of CCA.

The two-band CCA in the WPAN/WLAN device conduct channel sensing simultaneously. When the channel is free, both CCA detectors indicate a free channel. Table VI lists the output states of the two-band CCA detector when the channel is busy. The channel state indications are the same in both systems. When the channel is occupied by WLAN signals, both CCA detectors indicate a busy channel. In contrast, when the channel is occupied by WPAN signals, the auxiliary CCA detector indicates a free channel. Therefore, the signal source can be easily distinguished.

In the two-band CCA, the performance of every energy detector can be analyzed as in Section III. The total performance can be expected as the AND of the two CCA detector.

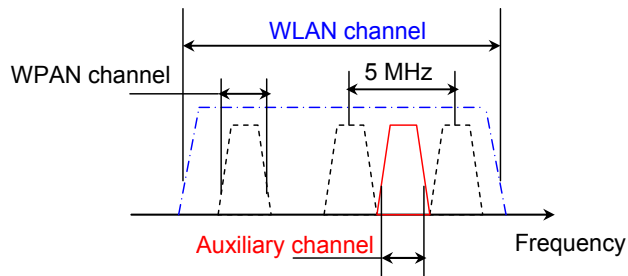


Fig. 6. Mechanism of two-band clear channel assessment

	802.11b		802.15.4b	
	CCA _{ch_11b}	CCA _{ch_aux}	CCA _{ch_15.4}	CCA _{ch_aux}
WPAN signals	√	×	√	×
WLAN signals	√	√	√	√

Table 6. Output states of the two-band CCA detector when the channel is busy

5. Conclusion

In this chapter, we have investigated the coexistence issue in the heterogeneous medical networks environment for ubiquitous health services from network access point. The energy based CCA was considered because the 2.4GHz ISM band is too crowded to apply feature based CCA for simple medical sensors.

Using central limit theorem, we have derived closed-form expressions for energy based CCA. We have shown and qualified impact of noise uncertainty on CCA and the asymmetric CCA in AWGN channel and fading channels. In the considered heterogeneous medical networks environment for ubiquitous healthcare purposes, WPAN is oversensitive to 802.11b signals and WLAN is insensitive to 802.15.4b signals. When WPAN and WLAN are located within the HERC, energy based CCA can effectively avoid possible packet collisions. The HERC is sufficient for most indoor medical applications. However, when they are farther apart, WLAN lose its sense to 802.15.4b signals. The asymmetric CCA puts WPAN traffic in a secondary position in the heterogeneous networks. We have proposed a two-band energy-based CCA with an additional CCA detector tuned in the auxiliary channel to combat the asymmetric CCA issue.

6. References

- R. Istepanian, E. Jovano and Y Zhang, "Guest editorial introduction to the special section on M-Health: beyond seamless mobility and global wireless healthcare connectivity," *IEEE Trans. Information Technology in Biomedicine*, vol.8, no.4, p.405-414, 2004.
- A. Soomro and D. Cavalcanti, "Opportunities and challenges in using WPAN and WLAN technologies in medical environments," *IEEE Communication Magazine*, vol.45, no.2, p.114-122, 2007.
- D. Cypher, N. Chevrollier, N. Montavont, and N. Golmie, "Prevailing over wires in healthcare environments: benefits and challenges," *IEEE Communication Magazine*, vol. 44, no. 4, p. 56-63, 2006.
<http://www.wilho.net/>
- A. Lymberis "Smart wearable systems for personalized health management: current R&D and future challenges," *Inter. Conf. of IEEE*, vol.4, p.3716-3719, 2003.
- N. Golmie, D. Cypher, and O. Rebala, "Performance analysis of low rate wireless technologies for medical application," *Computer Communication*, vol.28, no.10, p.1255-1275, 2005.
- N. Golmie, N. Chevrollier, and O. Rebala, "Bluetooth and WLAN coexistence: challenges and solutions," *IEEE Wireless Communication Magazine*, vol.10, no.6, p.22-29, 2003.
- I. Howitt and J.A. Gutierrez, "IEEE 802.15.4 low rate -wireless personal area network coexistence issues," *IEEE Wireless Communication & Network Conf.*, vol.3, p.1481-1486, 2003.
- S. Pollin, M. Ergen, and A. Dejonghe, "Distributed cognitive coexistence of 802.15.4 with 802.11," *Inter. Conf. on Cognitive Radio Oriented Wireless Networks and Communications*, p.1-5, 2006.
- Y. Xiao and N.H. Vaidya, "On physical carrier sensing in wireless ad hoc networks," *IEEE Conf. on Computer Communications*, vol.4, p.2525-2535, 2005.
- H. Ma, H. Alazemi and S. Roy, "A stochastic model for optimizing physical carrier sensing and spatial reuse in wireless ad hoc networks," *IEEE Conf. on Mobile adhoc and Sensor systems*, 2005.
- H. Zhai and Y. Fang, "Physical carrier sensing and spatial reuse in multirate and multihop ad hoc network," *IEEE Conf. on Computer Communications*, p.276-285, 2006.
- J. Zhu, S. Roy, X. Guo, and W.S. Conner, "Leveraging spatial reuse in 802.11 mesh networks with enhanced physical carrier sensing," *IEEE Conf. on Communications*, vol.7, p. 4004-4011, 2004.
- J. Zhu, B. Metzler, X. Guo, and Y. Liu, "Adaptive CSMA for scalable network capacity in high-density WLAN: a hardware prototyping approach," *IEEE Conf. on Computer Communications*, p.1-10, 2006.
- Y. Zhu, Q. Zhang, Z. Niu and J. Zhu, "On optimal physical carrier sensing: theoretical analysis and protocol design," *IEEE Conf. on Computer Communications*, p.2351-2355, 2007.
- K. Jamieson, B. Hull, A. Miu and H Balakrishnan, "Understanding the real-world performance of carrier sense," *ACM SIGCOMM workshop on Experimental Approach to Wireless Network Design and Analysis*, p. 52-57, 2005.
- I. Ramachandran and S. Roy, "On the impact of clear channel assessment on MAC performance," *IEEE Global Communications Conf*, p. 1-5, 2006.

- IEEE 802.15.4b standard, "Wireless medium access control and physical layer specification for low rate wireless personal area networks", 2006.
- IEEE 802.11b standard, "Wireless medium access control and physical layer specification for low rate wireless local area networks", 2001.
- V.I. Kostylev, "Energy detection of a signal with random amplitude," *IEEE Conf. on Communications*, vol.3, p. 1606-1610, 2004.
- S.M. Kay, "Fundamentals of statistical signal processing: detection theory," *Prentice-Hall PTR*, New Jersey, 1998.
- V. Erceg, *et al.*, "An empirically based path loss model for wireless channels in suburban environments," *IEEE J. on Selected Areas in Communications*, vol. 17, no. 7, pp. 1205-1211, 1999.
- A. Sonneschein and P.M. Fishman, "Radiometric detection of spread-spectrum signal in noise of uncertain power," *IEEE Trans. Aerospace Electronic Systems*, vol.28, no.3, p.654-660, 1992.
- IEEE P1073, "IEEE standard for medical device communications-overview and framework," 1996.

Multi-Agent Design for the Physical Layer of a Distributed Base Station Network

Philippe Leroux and Sébastien Roy
Université Laval
Canada

1. Introduction

As wireless networks are becoming more omnipresent and pervasive, appropriate resource allocation and organization becomes an increasingly pressing challenge. There exist on the consumer market two important types of wireless network technologies. On the one hand, cellular mobile networks are highly centralized and hierarchical. By contrast, wireless local area networks (WLANs) are deployed in an ad-hoc unstructured manner, thus avoiding the need for elaborate and costly planning. However, WLANs such as those falling under the highly successful 802.11 standard do not manage interference effectively and tend to collapse at high offered traffic loads. It can be seen that cellular and WLAN represent two radically different approaches in radio resource management, characterized by different sets of advantages and drawbacks.

The purpose of this chapter is to demonstrate the feasibility of a connection-oriented self-organized wireless system which offers efficient radio resource management and provides the best aspects of both cellular (reliable, connection-oriented operation even at high offered loads) and WLANs (ad-hoc deployment and distributed intelligence). This is achieved based on the multi-agent concept and local synergistic micro interaction (between neighboring transceivers) from which a global organization emerges.

The notion of Multiple Agent (MA) considered is of the “ant” variety, whereby small minimalist agents sense their environment and react to it in an interdependent manner. Social insects and mostly ants or bees are the most cited biological examples. In the literature such approaches have already been used to solve many combinatorial/optimization problems (Beongku et al., 2003; Brueckner & Parunak, 2003; Muraleedharan & Osadciw, 2003).

This design philosophy differs from more traditional approaches which consist in postulating criteria expressed by equations and models in order to formulate the problem in such a way that an optimal solution is derived within the defined context. In the agent approach, precise mathematical formulation of the problem is neither required nor very useful. The approach thus becomes attractive for tackling complex multidimensional problems which would otherwise be intractable. Therefore, our goal is not to demonstrate an optimal design, but to illustrate how a Multi Agent System (MAS) can be empirically designed and fine-tuned to fit a specific application. Moreover, it will be seen that such a dynamically adaptive solution, in spite of its empirical nature, offers many advantages over a rigid analytically-derived counterpart.

We will focus herein on Parunak's methodology (Parunak, 1997) because it offers an intuitive modeling framework, which is well suited to the empirical design approach.

Considering wireless networks, this chapter describes a flexible distributed base station (DBS) framework which removes many limitations of current networks in order to augment the solution space. For example, a plurality of DBS can simultaneously provide a network link to the same mobile, thus leveraging macrodiversity to improve link quality and/or achieve power savings. These DBS are designed with auto-organization in mind, such that the network structures itself autonomously. This is where MAS come in, offering the desired distributed intelligence, adaptability, scalability and auto configuration properties. However, the DBS architecture is challenging in at least three aspects:

1. It requires the continuously-updated solving of a large combinatorial problem, namely finding a good allocation of DBS resources to mobiles requiring service.
2. Interference must be handled in a transparent way so that mobiles can gain the best benefit of macrodiversity without being restrained by interfering mobiles.
3. Power control is an important aspect for both energy consumption and network capacity given that it is tightly-coupled with interference patterns.

These three aspects are entangled together such that an optimal allocation is a complex combinatorial problem. It is NP hard unless some heavy simplifying assumptions are made (on the geometry, on propagation, or other aspects). Moreover, in the context of mobility, an optimal solution at one point in time is not optimal if it cannot easily adapt to changing parameters (mobiles' positions, fading, etc.).

Yet, this complex context is well suited for a MAS design. Indeed, MA need an active environment in which to generate interaction. And each event of allocating power, channel or connections to mobiles, that a DBS generate, has consequences on other mobiles' links. This creates the required active environment in which agents can sense parameters such as the received power, interference and link quality, and where decisions can be made locally to generate new actions. In turn, the effect of these actions are sensed by other agents. The next section describes the challenges of the proposed DBS architecture. Then, MA design concepts used in this study are described. The fourth section details the proposed design of three categories of interacting agents respectively for :

1. macrodiversity connection management,
2. channel allocation, and
3. power level control.

Finally, the system is emulated. Results, including simulation of complex cases with randomly distributed DBS and mobile traffic, show first the resource allocation quality that can be obtained, and second the effectiveness of MAS design in terms of auto-configuration/scalability and dynamic adaptation properties.

A final brief discussion will extend Parunak's agent design principles to summarize the lessons learned from designing MAS for the application at hand.

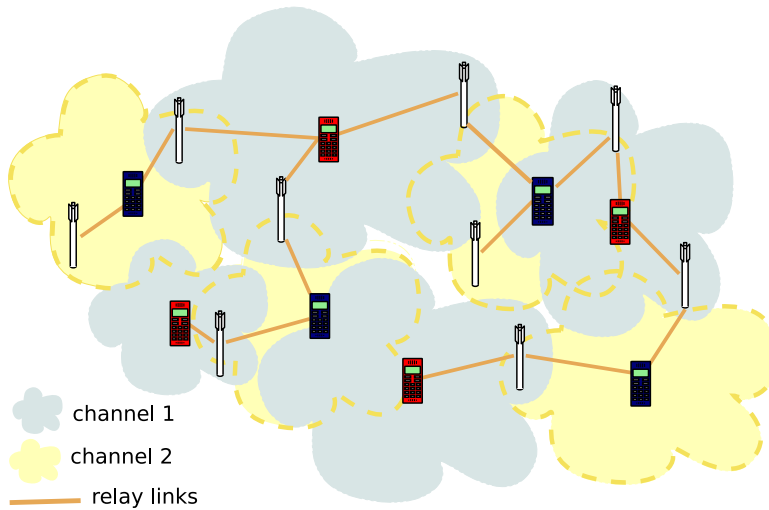


Fig. 1. Illustration of a DBS architecture exploiting 2 channels, and showing macrodiversity relay connections.

2. Challenges of the Distributed Base Station Network

2.1 Macrodiversity Potential

In a perfectly geometrical network with homogeneous traffic and symmetric propagation conditions, each DBS needs only to connect to the closest mobiles to maximize the provided quality of service. However, traffic is never homogeneous and varies across time and space in accordance with the users' schedules and patterns of usage. Moreover, propagation conditions are highly dependent on location, with varying availability of lines of sight and saturation of the frequency band due to heavy traffic. In such a context, there is a need for a simple, scalable, dynamic system to allocate relay links to mobiles and to continuously adapt the allocation pattern to changing conditions.

DBS can choose to relay mobiles far away from themselves in order to provide them with more macrodiversity, and thus better balance resource allocation. However, this choice involves a trade off. The exponentially-decaying link quality with the mobile-DBS distance could lead a remote mobile to consume many valuable relay links while deriving only marginal benefits, whereas closer mobiles would obtain much higher macrodiversity benefits from those same resources. Also, macrodiversity links provide not only enhanced overall quality links, but also reliability against network disconnection when undergoing severe fading, and it facilitates handover for mobiles moving outside from the range of some DBS to others. Therefore, a single criterion such as maximizing the minimum QoS for all mobiles would fail in certain conditions where enough resources would be available to provide the majority of mobiles with decent QoS, because of a few mobiles consuming much of these resources while deriving marginal benefits.

Such situations reveal the perils of pursuing a global solution based on a single perhaps overly simplistic quality criterion. In fact, many Pareto equilibrium solutions exist, in which no mobile can gain quality of service without stranding another user. And all these possible so-

lutions present multiple compromises on connection reliability and distribution of QoS. e.g. some solutions could favor maximizing overall signal strength for high transfer rate, others by distributing relaying links differently could prevent disconnections due to sudden strong fading or interference because mobiles would in general enjoy higher probability of being assigned multiple relay connections.

As such, it is not necessarily meaningful to define *a priori* goals for the search of a solution, as it is not known beforehand what are the benefits and drawbacks of each possible Pareto solution. This solution space is moreover hardly tractable due to the discrete nature of the problem, with a finite but large number of link resources to attribute. It is limited by physical conditions where some links may not be feasible due to the weakness of the considered signals. Also each link brings an increment of additional quality to the mobile's overall link quality, whose importance heavily depends on local propagation conditions that can vary continuously (with slow fading and changing mobile-DBS distances), or abruptly given the arrival of new connections in frequency allocation or strong fading situations. In this context, no analytically tractable mathematical framework exists leading to an optimum solution taking into account all the dimensions of the problem. More specifically, one must consider that an optimal solution, at a given state of the network and a given frame in time, could be too heavily specialized to that particular situation, such that a sudden change (strong fading, new mobiles joining the network) would make it ineffective. By analogy, it is known in biology that a species too well adapted to its environment is heavily endangered due to its limited capacity to adapt to environmental changes. Hence, a good solution is not an optimal one, but a good enough one that provides margins for adaptation in time to face changes.

There is a strong need for distributed techniques which are flexible enough to be tuned to the desired compromises while being able to handle unexpected events.

2.2 Channel allocation

Channel allocation faces the same propagation issues as connection management. To understand the implications of channel management, we introduce the concept of channel footprint. In a given situation (mobiles and DBS positions and relative densities, available channels, power allocations, etc.), a mobile's channel footprint can be understood as the space it occupies to maintain all its relaying links to DBS at a sufficient quality level. Hence, a second mobile, if emitting on the same channel inside this space, would affect some or all of the first mobile's connections.

In the cellular context, it is assumed that each mobile enjoys the same channel footprint which is controlled by the cell division of the space and an appropriate interference level threshold to allow or prevent the reuse of a channel across cells. In the 802.11 protocol, it is a handshake mechanism (the RTS/CTS exchange) which alerts neighboring transceivers that the channel will be in use, in order to control, to some extent, this channel footprint by preventing neighbors from reusing the channel in the vicinity, thus minimizing the hidden terminal effect (Ware et al., 2001).

In the DBS architecture, it would be appropriate that mobiles be offered varying channel footprints to adjust availability of channel resources and support the various needs of mobiles for macrodiversity. Indeed, mobiles needing more macrodiversity would require a larger footprint. Moreover, mobiles close to all of their relaying DBS should allow other mobiles to reuse the same channel at a closer range, compared to mobiles far from all DBS. This holds since these mobiles can support higher interference power and still maintain a good signal to interference plus noise ratio (SINR). This aspect of channel allocation was taken into consideration

in the dynamic allocation scheme known as the *umbrella cell system* (Furukawa & Akaiwa, 1994). However, channel reuse should not necessarily always be maximized, since it could lead to locally unnecessarily compacted channel allocation, even in low traffic conditions. Hence, channel allocation should be able to adapt to always balance resources to sustain the high throughput rate of modern communication services.

Another aspect to consider is the irregular geometry of the DBS network. The DBS density will necessarily change across space given some maximum or mean local load. And also, local mobile traffic varies across time and space (e.g. from residential neighborhood to office centers). This leads to constantly changing disparities in local loads of the network to which it must adapt the channel allocation. In the cellular world, solutions exist in the form of Dynamic Channel Allocation schemes (DCA), such as the well-known segregation scheme by Akaiwa & Andoh (1993), allowing cells with higher loads to use more channels.

Finally, faster channel allocation adaptation is required with high mobility, as mobiles move in and out of DBS' ranges. In cellular systems, this aspect is only tackled via a handover mechanism. However, in the DBS concept, DBS share channels via the macrodiversity relaying links. Hence, mobiles need not change their channel simply because they change a relaying link. This additional complexity implies channel segregation algorithms are not as efficient in the DBS concept.

2.3 Power allocation

Power level management can strongly affect the previously described aspects of connections and channel management. It can also offer a powerful means to leverage the possibilities in terms of resource management. Indeed, in the context of DBS, a higher power level implies modifying the mobile's channel footprint, and hence allowing it to reach perhaps more DBS for more macrodiversity. Or, on the contrary, a lower power level will leverage the channel's reuse possibilities by generating less interference. Power control can therefore help to provide much higher resource availability, and can provide synergistic behavior with the channel and connection allocation to leverage the possibilities that macrodiversity can offer to balance the resources across mobiles, throughout the network.

However, there potentially exists a maximum power level dynamic range which limits the effectiveness of power control. Indeed, decentralized DCA implies sensing the availability of channels. For example, the 802.11 protocol implements CSMA (Carrier-Sense Multiple Access) to prevent improper reuse of channels. In the cellular world with DCA, a maximum sensed interference power level threshold is considered to decide if the channel is available. Therefore, the dynamic range of power-level adjustments cannot exceed a certain range such that mobiles with low emission power needs are not disrupted by the interference caused by a new connection from another mobile, perhaps much further away, but with a much higher power level. It is noteworthy that this power level range is dependent on the space distribution of mobiles (and channel allocation). The range itself may not be constant at all scales such that it could be high throughout the entire network when compared to smaller areas of the network. Indeed, the SINR is a relative quantity which depends on the spatial distribution of access points and traffic.

Power level allocation is also a contention process where each mobile strives to maximize its QoS. Considering the viewpoint of one mobile, in order to maximize its QoS, it wants to be within reach of a maximum number of DBS to maximize macrodiversity, and it also wants to maximize its SINR for each of its links. Therefore, it wants to maximize its power level. However, if all mobiles did so, none would get any benefit. And since no mobile can obtain

any gains by reducing its power level, a non cooperative strategy is not a good choice for a game-theoretic approach to power level adaptation.

Necessarily, some mobiles will have to “accept” to reduce their power level in order to allow other mobiles in need to enjoy better QoS by reducing interference and enabling them to reach more DBS for macrodiversity. Yet, and due to the non linearity of the propagation environment, there necessarily is a point of diminishing returns for mobiles to reduce their power level. While any reduction necessarily implies a reduction in interference, the potential gain for other mobiles does not necessarily offset or compensate (given a compromise choice at a global scale) the loss in QoS for this mobile. It is to be understood here that there exist trade-offs for an infinity of Pareto solutions. Therefore, and again, postulating one global unidimensional criteria (e.g. as is done in traditional algorithms (Grandhi et al., 1993)) to derive a power allocation method would not allow assessment of the potential benefits of different trade-offs. Indeed, the results of the proposed design will show how the traditional approach to power control (which consists in maximizing the minimum SINR for all mobiles) in spite of offering interesting capabilities in some situations, also prevents most mobiles from achieving their QoS potential.

2.4 Complexity, Dynamics and Scalability

2.4.1 Complexity

In existing types of networks, the complexity is constrained by simplifying the hypotheses. For example, in cellular networks, channel allocation is simplified by segregating channels given an interference power threshold in order to guarantee a minimum SINR for all mobiles in a cell. This assumption simplifies the evaluation of provided QoS, as it guarantees a minimum QoS for connected mobiles, and avoids the hidden terminal effect, such that there only remains to evaluate the probability of a connection being blocked (when all channels in a cell or sector are occupied).

In the considered architecture, such assumptions are not made a priori as the purpose of the DBS architecture is to maximize flexibility. And considering the number of possible combinations of connections, or channels or even power levels, it is obvious that an exhaustive search to find all Pareto solutions is pointless. Even considering an exhaustive search in the case of a very simple scenario with only a few mobiles is pointless, since in such cases, the non-linear effects and interactions of large networks would not apply and the obtained results would be too limited to draw meaningful conclusions.

Also, postulating a unidimensional criterion and over-simplifying the non-linear effects involved, in order to provide a tractable mathematical framework would limit the solution space and therefore restrict the possibilities of such an architecture.

MA offer interesting properties to cope with complexity. The approach involves segmenting the problem into multiple subproblems where each is tackled by its own agent class. Heavy calculations for evaluating and selecting combinations are also avoided. Instead, specific combinations are attempted and modified by agents’ actions through local interactions.

2.4.2 Dynamics and Scalability

One particular aspect to consider is the fact that a given resource allocation solution must necessarily adapt to changes in a mobile wireless network. Such a solution must also adapt to unexpected events, such as the failure of a DBS. And finally it must scale, such that adding DBS locally will seamlessly, without any need for configuration, increase the capacity of the network in terms of either provided QoS or number of provided connections.

3. Multi-Agent Design

To solve the resource allocation problem, with the previously described considerations, multiple agents or bio-inspired optimization seems appropriate, as such approaches provide the most important sought-after characteristics, namely

- scalability ;
- dynamic adaptation ;
- auto-configuration ;
- reliability facing unexpected events.

Following Parunak's (Parunak, 1997) design principles, three main characteristics need to be provided in a MA design: coupling, auto-catalysis, and function. Coupling implies that each MA process is coupled directly or not to the others and their environment (e.g indirectly using pheromones via an environment). Auto-catalysis implies that the agents' actions taken in the right direction¹, by the nature of the agents' processes, favor similar actions leading the system to converge to a desirable state (positive feedback reinforcing the convergence towards the solution). And finally, the system must be such that a useful global function emerges out of the induced local interactions.

3.1 Coupling

To achieve coupling, Parunak explains that we first need an *active environment*. The radio propagation medium constitutes just such an environment, as each mobile emitting on a given channel influences the others due to interference. Hence, a mobile's movement changes the interference patterns for all others in its immediate vicinity. Additionally, mobiles are entities which strive to acquire connections and in so doing, they necessarily broadcast information to inform neighboring DBS of their presence and of their link quality. This forms an active environment in which information is exchanged to sustain coupled processes.

We emphasize the fact that DBS and mobiles do form appropriate entities to host agents that are *small in size and scope*. In particular, DBS, compared to central cellular base stations, are specifically meant to be small, and will necessarily have small scope as they can only relay a (smaller) limited number of mobiles in their vicinity.

As a final criterion related to coupling, agents should be mapped as *entities*, not *functions* since an agent does not implement a complete function. That is, the function optimizing resource allocation should be the result of the interaction of the agents and not be implemented as the output of one agent. Indeed, an ant (in ant colonies) does not find a shortest path alone.

In the proposed system, the agents are mapped to either mobiles or DBS. Their actions will then be to either allocate or deallocate a channel, or a connection, or modify a mobile's power level. Necessarily, all processes which modify resource allocation are all coupled since each agent's actions will not only influence the concerned mobile (changing channel, obtaining a new relay connection or changing its power level), but also influence the neighboring mobiles, modifying their own channel footprint, their QoS, hence influencing other agents, and coupling each agent's processes together indirectly.

¹ Since a priori goals are not explicitly defined, neither is the concept of a "right direction". Rather, a behavior is designed, tuned and retained because its auto-catalysis properties happen to converge to a solution which satisfies the needs of the system. Therefore, such a design allows wide exploration of the solution space rather than restricting to predefined goals by not including all the effects involved in the multidimensional problem. The design represents a certain creative process.

3.2 Auto-catalysis

3.2.1 Flows

For agents to maintain their interactions, they must be designed to let the process evolve continuously. Therefore, agents should not be designed based on discrete state transitions, leading to pauses in the processes because of unverified conditions. That is why we must favor *flows instead of transitions*. One way to achieve this is for agents to use volatile markers (i.e. permanent and non-obstructive source of information which dissipate in time as they become irrelevant –e.g. pheromones in ant colonies) to inform other agents on their particular state, so that the agents' processes continuously evolve rather than stop and wait for specific conditions.

It is a design choice that *no explicit information exchange* is performed concerning the positions of mobiles and DBS, available resources, etc. As mentioned, the available information stems from what DBS and mobiles can sense locally (mobiles' needs and QoS), which represents our volatile markers. These bits of information are by nature volatile, as they only stay in the environment as long as they are broadcasted by the mobiles, and hence are necessarily current.

Since agents should not wait for predefined conditions to take actions, it is a *comparative* basis that will trigger a corresponding action of:

1. allocating/deallocating a macrodiversity connection;
2. changing a mobile's channel (frequency hopping);
3. increasing/decreasing a mobile's power level to a certain amount.

3.2.2 Homeostasis

The notion stems directly from biology in which systems always strive to maintain an equilibrium or *homeostasis point*, e.g. the blood sugar concentration is maintained (mainly) by two different hormones which have opposite effects to balance the concentration.

This point of equilibrium must be sustained by an ongoing *flow* to ensure the system continuously explores the solution space and does not get stuck in a deadend. This flow is analogous to the variations of a stock market title whose value is influenced (at a macro level) by the traders' actions of selling and buying. In turn, at the micro level, the variations of the values influences the traders' decisions.

The corresponding aspect of our system is created by forcing DBS to continuously create and destroy connections, continuously change channels (via channel hopping), and continuously adjust power levels. Each of these actions — at the macro level of agents — influences the status of mobiles, and these changes are in turn sensed by surrounding mobiles and DBS.

In effect, the flow of actions makes the system converge to a homeostasis point. This point will be dependent on the the state of the network (traffic, available resources, etc.) due to the comparative basis that triggers actions. As long as there exists a bias observed by the agents that will trigger an action, the system will converge or oscillate to its homeostasis point. These variations are important, since without them, and if there is no other change in the system (e.g., induced by mobile motion), the sensed QoS of mobiles would never change, never trigger actions, and the system might simply stop short of an optimal state.

3.2.3 Amplification and limitation

Together, amplification and limitation constitute an other important aspect to generate the convergence to a homeostasis point. Amplification implies a positive feedback mechanism

such that convergence (to a solution) is favored. In other words, the actions of an agent which lead the system in a desirable global direction should be favored and should also influence the surrounding agents to act in the same direction.

In effect, an MA system is comparable to a Genetic Algorithm (Goldberg, 1989) preserving "genes" that seem to provide the best fitness and hence are part of an optimal solution. The difference is that there are no external observing entities that measure via a metric the fitness of candidate solutions. Rather, it is the interactions between agents and their environment — the propagation medium — that must provide the natural selection function.

Limitation also implies preventing the whole system from focusing on one point (exacerbating the convergence of actions to a local minimum) and thus miss a better solution. Moreover, limitation can favor convergence by dampening the effect of amplification to prevent the system from going past a solution or oscillating around it without converging.

3.3 Function

Coupling may be trivial to obtain and auto-catalysis somewhat more involved, but if the process as a whole does not realize a useful function, then it is irrelevant. Function implies that the homeostasis point described previously is useful for the system, e.g. in biology the homeostasis point for the blood sugar concentration is such that enough sugar is available to fuel the cells, but not too much to avoid excessive sugar loss through the kidneys.

In our system, the sought-after function consists in

- **maximizing the potential usage of the resources;**
- and **balancing them** to offer a good compromise of quality across all mobiles, while not hindering the overall system performance.

Most often, function is obtained through a *utility function* which translates the flow of variations (of QoS) sensed into rational decisions. That is, it converts a multi-dimensional problem into a one-dimensional quantity upon which decisions for actions are based.

In spite of the fact that many frameworks attempt to provide mathematical support to derive such utility functions (such as game theory (Mackenzie & Wicker, 2001) or COIN theory (Tumer & Wolpert, 2004)), these frameworks mostly consider intelligent agents having the ability to learn (eventually using reinforcement learning techniques), which is not the nature of the proposed design. Ultimately, defining simple agent behavior to obtain an intended global behavior still relies on intuition and art such as in Conway's "game of life" (Elwyn R. Berlekamp et al., 1982), or with Wolfram's cellular automata (Wolfram, 2002). Therefore, no systematic procedure is known which derives the locally-applicable utility function from the desired global behavior.

Function can also be sustained (especially if a utility function is not found) with

- *behavior diversity* and
- *randomness*.

Randomness can be helpful to introduce alternative solutions, that will or not be kept in time given how effective they are. Behavior diversity can be obtained by forcing neighboring agents to act differently so as to provide different reactions and experiments given identical stimulus. These properties support the function property by *breaking the symmetry* so as to prevent the system from entering any deterministic patterns which might hinder convergence. In the following section it is described how auto-catalysis and function are obtained for each class of agents.

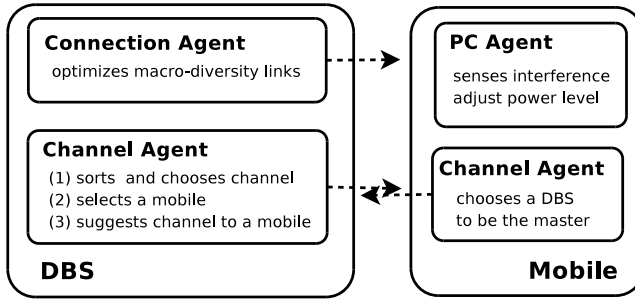


Fig. 2. Agents actions and mapping to entities.

4. Agents' Design

4.1 Agents Sensing Abilities

A mobile broadcasts its actual QoS designated $P_T(m)$, so that DBS in its vicinity can sense its needs. This QoS corresponds to the total BER after macrodiversity combining the mobile's received signals. Moreover, a DBS senses the potential additional link quality it can provide (or already provides) to a mobile. Considering mobile m and DBS B , this additional link quality is named $P_e(m, b)$, and represents the BER of the link from m to DBS b (Leroux et al., 2006). Also, to incorporate the notion of classes of QoS, the DBS knows that mobile m requires an overall quality of $P_d(m)$, i.e. $P_T(m) < P_d(m)$.

For convenience, these values ($P_e(m, b)$, $P_d(m)$ and $P_T(m)$) are expressed in a logarithmic scale of base 10 of the BER. It is shown in (Leroux et al., 2006) that the following holds in a Rice fading environment with different Rice fading parameters ($SINR, K$ factor) values for each link:

$$P_T(m) \approx a + \sum_b P_e(m, b), \quad (1)$$

where the relation (which implies that the combined BER is the sum in the logarithmic domain of the individual link BERs) is exact for certain types of modulation (such as DPSK) and approximate for other types (such as coherent QPSK), and a is a constant related to the modulation.

By definition, if mobile m is not relayed by DBS b , we have

$$P_e(m, b) \triangleq 0. \quad (2)$$

4.2 Connection Allocation

4.2.1 Coupling

Connection agents are mapped to DBS. Indeed, each DBS senses local information on mobiles' needs and can take decisions with regards to the allocation of connections. Mobiles can be relayed by many DBS offering more or less QoS. Therefore, a DBS decision (whether to relay or not a mobile) will influence what its neighboring DBS senses and therefore influence their actions. Hence, the whole network is interdependent and linked or coupled via macrodiversity links.

4.2.2 Flow of action

While attempting to maintain connectivity, DBS should *continuously* change their links to gradually converge to an optimal configuration which maximizes the benefits of macrodiversity. It is this flow of changes in connection allocation which sustains the properties of auto-catalysis and function.

In order to do so, two types of actions are defined : disconnection and reconnection. Each action is taken alternatively, given the number of active connections the DBS has. Suppose a DBS can provide at most N connections and has m active connections, it will perform

1. $N - n$ disconnections, if it has $m > N - n$ connections or,
2. $N - m$ connections, if it has $m \leq N - m$ connections.,

where n is a system parameter (typically $n = 1$) under the designer's control. For large values of N , increasing n helps accelerate the convergence of the system, yet high values of n will suppress the iterative selection mechanism such that the system may not converge any longer. Hence two opposing "forces" must be designed to link the information sensed to the choice that must be taken : which mobiles should be connected or disconnected. Finally, the balance between these two forces should lead connections to a state that represents an homeostasis point.

4.2.3 Function

Randomness is incorporated by having the connection agents activate (to perform a connection or disconnection action) randomly, following a Poisson law considering a discrete model of time. This should also help to prevent any periodic pattern from taking hold.

A *utility function* is designed to link the information sensed to the choice of actions : links are rated according to a continuous function so the DBS can compare the links and decide which to connect or to disconnect.

Two metrics are designed, providing information on:

1. a measure of how well mobile m is served with respect to its requested QoS, i.e.

$$F_{\text{need}}(m) = \frac{P'_T(m)}{P_d(m)}, \quad (3)$$

2. and a measure of how much diversity DBS b is providing to m with respect to the mobile's overall link quality, i.e,

$$F_{\text{div}}(m, b) = \frac{P_e(m, b)}{P'_T(m)}. \quad (4)$$

P'_T is understood in these definitions as the mobile's total link quality if DBS b is connected to the mobile (whether it is evaluating for connection or disconnection).

These two simple functions provide sufficient information for the DBS to compare the mobiles' links and take a decision based on:

- how much a mobile *needs* more macrodiversity links;
- and to which extent this DBS is the one which will provide the mobile with an efficient macrodiversity link (relative to the other DBS currently serving the mobile).

These information bits still need to be combined, and this is where an appropriate trade-off is induced by using different combination functions.

To design the combination function, different characteristics must be considered:

1. If a mobile has no connection, it must be favored, since basic connectivity should take precedence.
2. If a mobile has only one connection, the DBS should not disconnect it.

Furthermore, there are two complementary compromises involved in the DBS' decision process:

1. either to remove a link because the mobile already enjoys sufficient QoS,
2. or to maintain it because it is the main DBS providing it;

and,

1. either to connect a mobile because it is in need,
2. versus not connecting it because the additional diversity brought to this mobile would be low (compared to other possible connections).

Finally, the function must provide a natural ordering to classify the compromises in order to take a decision.

The following function addresses all the characteristics discussed above:

$$C(m, b) = F_{\text{need}}(m) \times \log(F_{\text{div}}(m, b)). \quad (5)$$

This function is necessarily positive or null. It is null if the mobile has no connection, since, if it were connected to the DBS, it would have $P_T(m) = P_e(m, b)$ which implies $F_{\text{div}} = 1$ giving a null value of the logarithm. Likewise, it is null if considered for disconnection and the mobile's only link is to the considered DBS. Hence, if this utility function is null, the agent will either privilege this mobile for connection or not disconnect it to keep the mobile's existing connection active.

The evaluation of the compromise is obtained by the multiplication of the two terms. Hence, the more the DBS provides diversity, or the higher is the current QoS enjoyed by the mobile, the higher is the function's value.

The choice of compromise itself comes derives from "shaping function" used prior to the multiplication of the two metrics. Simulation showed that the optimization happens most efficiently if the shaping function of the second term is concave (naturally, it should be strictly increasing), hence the use of the logarithm, which also provides the necessary null value for a mobile with a single link.

4.2.4 Limitation and amplification

Limitation and amplification is naturally obtained with the environment propagation properties. Indeed, a poor signal quality will favor multiple connections (amplification), but distance (mobile to DBS) and the infrastructure link capacity of DBS will restrict excessive connection growth (limitation).

Also, this amplification (or attraction of macrodiversity links) and limitation sustains the homeostatic behavior where mobiles in need get more links up to an equilibrium point where additional links to these mobiles would overwhelmingly affect an otherwise well-served mobile.

4.3 Channel allocation

4.3.1 Flow

The flow of actions in the channel allocation agents naturally consists of the changes in channel allocation, or *channel hopping* which modifies mobiles' QoS and interference patterns which in turn should trigger other changes.

For this flow to be generated properly, appropriate actions are specified in the following.

4.3.2 Coupling

Following Parunak's principles, the sought-after function (optimizing the allocation) is divided into independent actions whose interactions should lead to the other two properties (auto-catalysis and function).

First, given the macrodiversity context, a mobile will choose one of its relaying DBS to be its "master" connection, which implies one type of action and one agent (to select the master) mapped at each mobile.

Second, DBS will choose mobiles (from their master links) and change their channels as is done in cellular systems. Except that here, the change, or channel hopping, will not be triggered by specified conditions (e.g. a mobile SINR falling below a threshold, or a mobile changing cell). Instead, the flow of channel hopping will be sustained by having DBS choose a mobile at each agent activation and change its channel. Channel allocation agents will activate in the same way that the connection agents do. Two types of actions must be defined:

1. choosing a mobile, and
2. choosing a channel.

Mapping these actions at the DBS level, rather than letting the mobile decide when to change channel makes sense in that DBS can gather information most effectively on the different channels in use, thus preventing mobiles from having to continuously scan channels.

4.3.2.1 Sensing

In addition to the mobile's sensed link quality, DBS can sense

1. the received power of surrounding mobiles $p_r(m, b)$;
2. and the interference level on various channels $p_I(b, c)$ (for channel c at DBS b).

4.3.3 Function

Maximizing the channel usage constitutes, in a sense, an effort against the second law of thermodynamics. Indeed, the channel allocation, if optimal at some point in time, will necessarily deteriorate with mobility as two mobiles transmitting on the same channel get closer to a point where the interference will degrade the offered QoS, such that resources are not balanced anymore. Considering this aspect, and rather than trying to solve an NP-complete problem, load balancing is obtained by always attempting to change the channels of mobiles in need such that they enjoy better SINR.

Three utility functions need to be designed taking as input what the agents can sense, and yielding a chosen parameter value as output.

- a. Mobile m will choose a master DBS (among its relaying DBS) on activation (where its activation follows a Poisson law) based on the DBS from which it obtains the highest link quality:

$$b = \arg \max_b \{P_e(b, m)\}. \quad (6)$$

- b. DBS b will choose a mobile (among mobiles connected as master to b), that is the most in need, i.e.

$$m = \arg \min_m \{F_{\text{need}}(m)\}. \quad (7)$$

- c. Ideally, the DBS should try to use the channel with the lowest interference power level:

$$c = \arg \min_c \{p_I(b, c)\}. \quad (8)$$

However, it may be overwhelming for a DBS to systematically sense channels to maintain up-to-date information on interference levels on all channels, and this behavior (utility function c.) is therefore only used as a benchmark.

Akaiwa & Andoh (1993) suggested a selection mechanism which is used herein with some modifications. DBS b will scan channels in the order of a given priority list it maintains, and determine if a channel can be assigned according to

- whether the resulting SINR will be above an SINR threshold ;
- and (in addition to Akaiwa's method) whether it will also be above the actual SINR the mobile enjoys.

The SINR threshold represents a mean to control the hidden terminal effect. It is a studied parameter in order to observe to which extent it prevents HTE while not limiting the flexibility of the system.

For Akaiwa's segregation algorithm, the priority list is obtained dynamically given the ratio for each channel of previous assignments versus previous assignment attempts.

Finally, a random priority list is proposed as a simple, yet effective (as we will see) alternative to the segregation algorithm approach.

4.3.4 Limitation

DBS will only test a limited number of channels given by the Ch_{max} parameter, before giving up. Indeed, there is no guarantee that the DBS will find a channel that will suit the chosen mobile. Therefore, and instead of letting it scan all channels, it is forced to limit its search. Eventually, it will try again, or another DBS will, thus providing behavior diversity as well. In effect, the DBS are only trying to maintain channel assignments in a working state by "upgrading" the solution iteratively in an opportunistic fashion given the eventual availability of channels. It is the effect of a new channel allocation that will cause other DBS to also react and change channels for the mobiles that will see their QoS affected by the new neighboring interference. As this flow of action is sustained, the channel allocation remains functional and should adapt to changes.

4.3.5 Channel availability

An additional functionality is provided for channel availability. A few spare channels are reserved for the initial connection or reconnection of stranded mobiles (instead of using channels from the main pool). Then, a master DBS which has mobiles on these spare channels will attempt to change their channels as a priority instead of choosing another mobile. Such spare channels allow rapid network entry, providing higher availability as well as some time margin for the DBS to find free channels in the main pool. It therefore eases the process and the flow of channel hopping.

4.3.6 Alternatives

Some attempts were made in order to add additional functionality to better handle irregular topologies and classes of QoS. One such attempt was not only to change channels of mobiles in need to better ones, but to also change channels of over-served mobiles with worse ones. However, this approach proved fruitless. One reason is that it is not possible for DBS to differentiate between channels having high interference due to over-served mobiles or poorly served mobiles. And the change to a “worse” channel can have more cons (far too much degradation for other mobiles) than pros (compacting channels more efficiently to free resources). Indeed, the main challenge resides in the degradation of interference. And, in effect, better results are obtained by simple channel hopping with the poorly served mobiles until other mobiles are affected, react, and some balance is obtained.

The concept of classes of channels was also explored, where the threshold considered to assign a channel would be modulated given the class of the channel and the mobile’s need, in order to generate classes of channels with less interference and some with more, for mobiles which can sustain it (which is also inspired from the umbrella-cell mechanism). No significant increase of performance has yet been obtained with this approach.

4.4 Power Allocation

4.4.1 Flow

Necessarily, the flow in power allocation is the result of changes in power level, which in turn changes interference, and the QoS of surrounding mobiles, which in turn should trigger other power level changes. This striving for amplified/limited adjustments of power levels aims to converge to a point of equilibrium : an homeostasis point where the system actually exhibits its expected power control behavior.

4.4.2 Function

The previously defined $F_{\text{need}}(m)$ definition is considered. If it is above 1, the mobile is *over served* (it enjoys a BER better than requested) and should reduce its power level, or increase it if lower than 1. Yet, this in itself would be restrictive, as it would force all mobiles to the same mean quality $\overline{F_{\text{need}}}$ value, without taking into account the non linearity of the problem due in part to the limited dynamic range of mobiles’ power levels. Indeed, and as will be shown, the thermal noise at the receivers imposes a limit on the minimum power level a mobile can transmit without the resulting SINR being too low for the connection (however small the interference level is). And, of course, mobiles saturate at a maximal power level. Also, some mobiles might be able to obtain more quality while not restricting others to maximize their own, and if so, they should.

The Need variable is introduced to describe what could be the power level of the mobile given its $F_{\text{need}}(m)$ factor : it is the value to which the mobile’s power level should converge to if nothing else changes (which is not the case as other mobiles will adjust their power level). When, for all mobiles, the current Need_m value equals the current power level $p(m)$, then the homeostasis point is reached. The variable is defined

$$\text{Need}_m(F_{\text{need}}(m)) = 2e^{(-SF_{\text{need}}(m))} \times (SF_{\text{need}}(m) + 1) - 1, \quad (9)$$

where the value S is a scaling parameter for $F_{\text{need}}(m)$ in order to allow mobiles to potentially obtain QoS higher than their requested $P_d(m)$. Therefore, the Need function will not necessarily be smaller than the current power level if $F_{\text{need}}(m) > 1$. And mobiles will not be forced to restrict their obtained QoS to a global mean value. In the current simulations, this QoS is

maximized with $S = 0.8$, and this has been shown to hold in many different conditions of traffic, mobile speeds in (9) and available resources. .

The exponential in (9) is a shaping function which also naturally affects the dynamics of the system. In effect, it affects mobiles' convergence speed differently given their needs, and this translates into behavior diversity as no mobile will react in a precisely proportional manner. The proposed function is of course not the only possible choice, but it has proved stable and effective. Again, for MAS, effectiveness does not lie in the mathematical exactness of the function, but in the interactions it will generate.

4.4.3 Homeostasis

Finally, this Need factor must be converted to a delta (step) value to adjust the power level. *Homeostasis* is obtained by comparing the Need value to the current power level the mobile uses to transmit. Hence, the delta value is in the form of $\text{Need}_m - p_m$. The mobile will then try to converge to a Need_m value which depends on local interactions given itself and neighboring mobiles' F_{need} values (given that these are indirectly linked via interference). Eventually, a non-linear concave function helps convergence so that with Need_m and p_m close, the generated delta is kept small to slow down variations and help stabilize the convergence. We postulate

$$\Delta_m = \beta \text{sign}(\text{Need}_m - p_m) (|\text{Need}_m - p_m|)^{1.5}, \quad (10)$$

where the β factor is used to modify the dynamics of the system to attain the proper compromise between convergence speed and stability.

4.4.4 Limitation

Experience shows that this function is too unstable with high values of β . Still, it can be stabilized with additional scaling parameters, while maintaining fast adaptation in time with large values of $\beta \geq 5$, which is important for mobility ($\beta = 5$ is used in the presented simulations). Therefore, it is proposed that Δ be scaled according to the current power level and also the desired power level (the Need value). That way, if these values are small, Δ is also kept small to prevent strong changes in the system that would otherwise suddenly generate exaggerated interference. Indeed, such changes would lead to complications such as breaking existing links or simply propagating exaggerated reactions throughout the system. Building upon (10), the following function is used :

$$\Delta_m = p_m \times |\text{Need}_m| \times \beta \text{sign}(\text{Need}_m - p_m) (|\text{Need}_m - p_m|)^{1.5}. \quad (11)$$

Finally, the delta value is constrained to not exceed the power level range:

$$\Delta_m < 0 \Rightarrow \Delta'_m = \max\left\{\Delta_m, \frac{-p_m}{2}\right\} \quad (12)$$

$$\Delta_m > 0 \Rightarrow \Delta'_m = \min\left\{\Delta_m, \frac{1}{2}(1 - p_m)\right\}. \quad (13)$$

As the mobile's PC agent activates, its power level is adjusted as follows:

$$p_m^{(v+1)} = p_m^{(v)} + \Delta'_m. \quad (14)$$

5. Evaluation

5.1 Simulation platform

For the channel and power agents, simulations are based on the following platform. The results shown for the connection agents are based on a simpler scenario (detailed in the appropriate subsection), in order to isolate the effect of connection management and observe its convergence, while not confusing it with the effect of interference and power-level management.

Physical parameters

A square field of 25 square kilometers is considered, in which 1000 mobiles evolve and 100 DBS are scattered randomly. Hence, the traffic's and network resources' geometry are not uniform, thus generating good and bad coverage of different areas. A mobile moves in a random direction at a random speed taken (at the start of a scenario) out of a uniform distribution over $[0, V_{\max}]$. DBS can relay 25 mobiles each, such that the mean number of macrodiversity links per mobile is 2.5. A mobile's maximum transmit power is 1W at 1 meter of its antenna, and the propagation exponent is 4 ($g_{ij} \sim 1/d^{-4}$). Rayleigh fading is considered, except near a DBS (closer than 100m) where a line of sight component is added with Rice factor $K = 5$ dB. Thermal noise at the receiver is considered for a bandwidth of 30kHz at a temperature of 20°C , hence $N_0 = -129$ dBW. The number of available channels is denoted Ch.

Agents' emulation

Simulations are run for 1000 seconds and repeated 10 times with different initializations of the geometry (DBS positions and mobiles' initial position, directions and speeds). Time is discretized with a time step of 1 second. At each time step, physical parameters are evaluated (mobile's position, propagation, interference, BER, connection outage). Agents activate randomly given a Poisson distribution to estimate the next activation time with parameter $\lambda = 3$ time steps. At each time step, the agents which activate evaluate their local state and take actions accordingly (adjust the power level, hop to a new channel, change connections of the concerned mobile, etc.).

Results

At each time step, the set of QoS indexes (total BER level given on a logarithmic scale $P_T(m) = \log_{10}(\text{BER}(m))$) for each mobile are sorted, thus providing a snapshot in time of the distribution of the network's resources across all mobiles. These sorted distributions are then averaged for all the time steps of the simulation. Given this information, it is then possible to compare how each algorithm distributes resources. The same is done for the power level allocation. Also, to verify the stability in time (considering the dynamic properties) of the algorithm, two factors are interesting to observe to understand how the system handles outage :

1. the mean number $\overline{N_d}$ of mobiles that loose all connections to the network per second, and
2. the mean time $\overline{t_r}$ it takes for the network to reconnect a mobile after it has been disconnected.

The latter also provides insight on how well the system is able to provide resources to mobiles with high availability.

5.2 Connection agents

In order to show that the connection agents are indeed optimizing the connections to balance resources, a simple centralized algorithm based on heuristics is proposed.

Algorithm 1 Centralized connection allocation algorithm

Considering initially that all DBS provide connections to all mobiles, and as long as there exist DBS with more than N maximum connections :

- A1 eliminate the connections with smallest $P_e(m, b)$ as long as m has $P_T(m) < P_d(m)$ and provided that DBS b has more than N connections ;
 - A2 (compromise on QoS) remove the ones with smallest $P_e(m, b)$ as long as m remains connected (another DBS is providing it a connection);
 - A3 (compromise on connectivity) finally remove connections with the smallest $P_e(m, b)$ until b has N maximum connections.
-

This algorithm is optimum at maximizing the sum of QoS $\sum_m P_T(m)$, given it only removes the smallest values. However, and given its limited ability to make compromises, it will not be efficient at balancing resources for mobiles and preventing disconnections of some mobiles if the network is resources-constrained. Necessarily, it offers a different trade-off than the agent algorithm provides.

Three cases are observed:

1. there are not enough resources (Fig. 3(a)),
2. there are enough resources for connections, but not enough headroom / margin and the agent system is not able to achieve swarming and converge (Fig. 3(b)),
3. there are enough resources to connect all mobiles and provide sufficient QoS, i.e. the connection agent is efficient (Fig. 3(c)).

In practice, only the third case should be relevant provided that the network is appropriately scaled for the needs of the users.

For the results shown in Figs. 3 and 4, a trellis of 19 DBS is used with 200 mobiles. Channel management is not considered and each mobile has its own channel.

In the third case (Fig. 3(c)), three successive phases in time can be observed :

1. a connection stage, where connections are established to the closest mobiles;
2. a connection optimization stage, where connectivity is maximized, and
3. a connection rearrangement stage, where QoS is maximized.

Figure 4 depicts a sort of the QoS $P_T(m)$ of mobiles to provide insight on how well resources are balanced. In this simulation, two classes of QoS are created, each comprising 100 mobiles. Compared to the centralized algorithm, it is obvious that some load balancing is performed by the agent system.

Due to lack of space, figures for the dynamic behavior are not shown herein. However, it is important to note that, without requiring any information centralization or excessive signaling (which would generate delays), and based only on local interactions induced by connection and disconnection actions, the agent system is able to keep up (maintain the connection allocation in a relatively optimal state) fast enough to sustain mobiles moving at speeds of 50 km/h with connection agents activating in the mean only once every 3 seconds. Above that

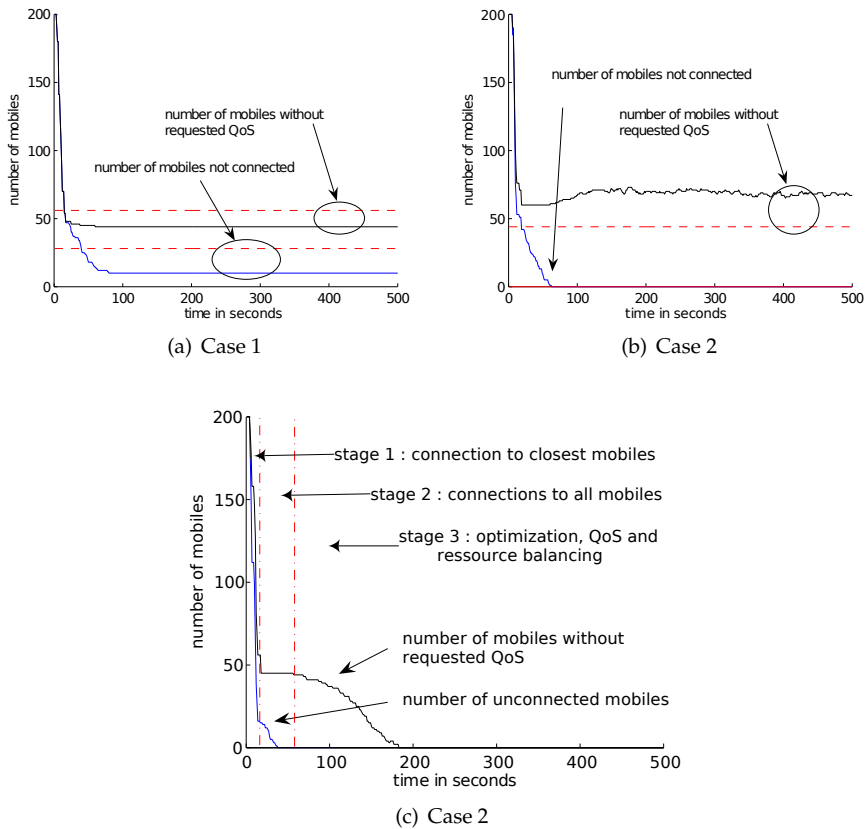


Fig. 3. Convergence of the connection management agents (horizontal dashed lines show results of the centralized heuristic algorithm).

speed, performance in terms of QoS degrades smoothly as the agents are not able to converge fast enough to the optimal state. The system still provides much headroom as activation of agents could be much faster.

Given that the proposed design is bio-inspired, it is most interesting to observe “health parameters” (analogous to e.g. blood pressure in the human body) which give us insight on the capacity of the agents to achieve their function. For the connection agents, the mean number of connections should be close to the mean number of disconnections. This indicates that the agents have sufficient headroom (when faced with changes in the network) to actually swap connections for optimization. If there are more connections than disconnections, it means that the system is not able to keep up with changes so that some of the relay links are disconnected for physical reasons (e.g. loss of signal quality) instead of explicit decisions by the agents.

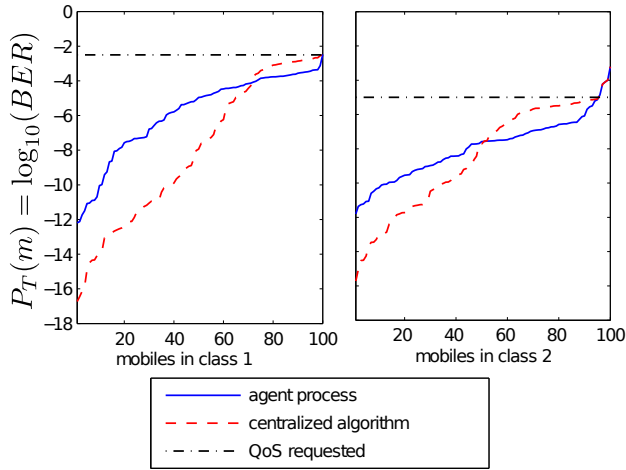


Fig. 4. Sorted profiles with 100 mobiles in class 1, 100 mobiles in class 2, and $N = 25$

5.3 Channel allocation

Channel management in the current context is hard to analyze and compare. Indeed, there exist many parameters which are dependent and which render objective comparison of designs and algorithms rather difficult because of the many trade-offs involved.

For example, the hidden terminal effect (N_d) can be minimized at the expense of the capacity to reconnect quickly (t_c). Also, QoS can be enhanced at the expense of connectivity. Indeed, the existence of a handful of unconnected mobiles implies more headroom (due to less interference) for either QoS or faster reconnection.

The notion of block rate must, as it stands in a cellular context, be revisited, since in this distributed system, a mobile is not necessarily blocked because an attempt at connection fails. It might fail simply because the DBS suggesting the connection is not well positioned, and another DBS will try and succeed hopefully fast enough. However, the block rate (of blocked allocation attempts) still represents an interesting bit of information as it reveals how effective the agents are at trying to connect. It therefore translates the notion of fluidity. Yet, it is not directly comparable to the notion of block rate used in cellular networks.

In Figure 5(a), the effect of the SINR threshold is shown by comparing the number of lost connections per second (for 1000 mobiles) and the mean time it takes to reconnect them. There is an obvious ideal trade-off point. It is interesting to note that the best compromise, which also maximizes the mean number of connected mobiles (Fig. 5(b)) is just above 0dB. Indeed, such a threshold would be much higher in a cellular system. However, the DBS architecture allows much more flexibility and sustains lower SINR with macrodiversity.

Figure 6 shows the effect of using reserved channels for reconnection on a log-log scale. For this scenario, time has been discretized at one tenth of a second and agents activate in the mean once every second (or ten time units). In the mean, mobiles are reconnected up to ten times faster, without any loss in QoS. Numerically, slightly more mobiles get disconnected but a much larger number of mobiles are connected in the mean. Indeed, since more mobiles are in the mean connected, there is more interference to deal with, which makes it more difficult

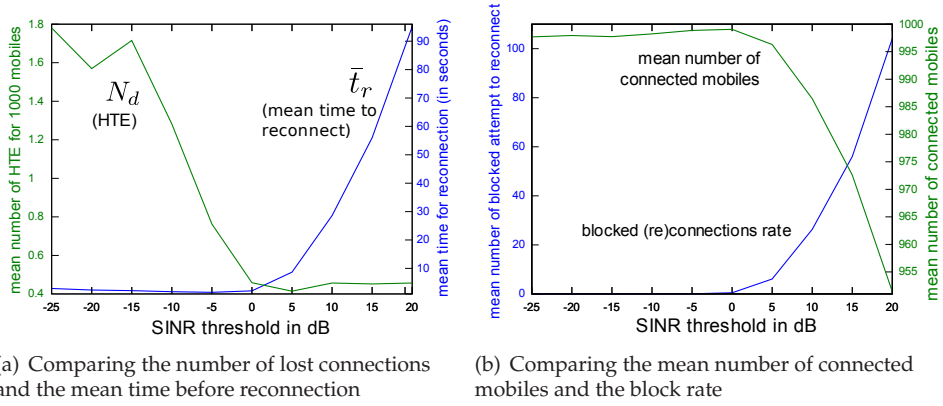


Fig. 5. Effect of the SINR threshold ($V_{\max} = 5$ m/s, Ch = 40, $N = -129$ dBW).

to prevent momentary disconnection. But the efficiency of reconnection is so much improved that in the mean, disconnection only occurs for less than 1 second (.8 seconds), and given the exponential distribution of reconnection time, 90% of the disconnections have smaller reconnection times. On the other hand, disconnection time is on the order of 5 seconds with no reserved channels.

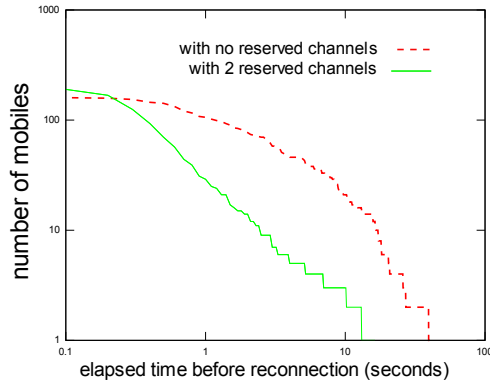


Fig. 6. Mean reconnection time with and without reserved channels. The Y axis represents the number of mobiles remaining unconnected since their disconnection (for 1000 mobiles during a 1000 seconds scenario length ($V_{\max} = 5$ m/s, Ch = 40, $N = -129$ dBW)).

Comparing the three different methods to select channels, little difference was observed in terms of the mean number of connected mobiles. Differences appear as trade-offs between mean number of disconnected mobiles and mean time of reconnection. The segregation method seems more efficient at minimizing disconnection, but takes more time to reconnect, compared to the random method. Differences appear most explicitly when looking at the

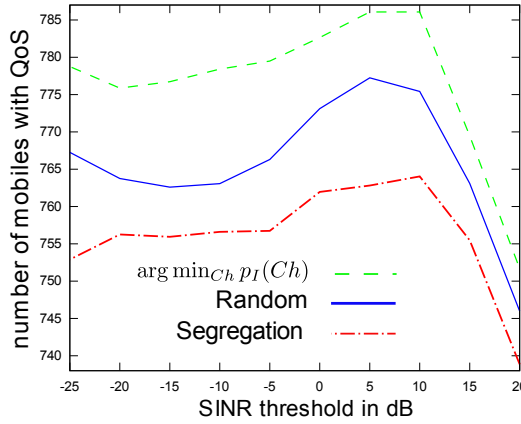


Fig. 7. Comparison of the performance of the three different methods for choosing channels ($V_{\max} = 5$ m/s, $Ch = 40$, $N = -129$ dBW).

provided QoS (Fig. 7). Choosing the channel with lowest interference power yields the best results, followed by the random choice and the segregation method. We notice an increased number of mobiles with their QoS demand met when the SINR threshold increases slightly. This is simply due to the fact that fewer mobiles are connected, generating less interference and higher QoS. However, this fact does not hold true for long, since a further increase in the threshold leads to a rapid increase of the agents' block rate, showing that they are unable to keep up with changes, and are not finding free channels to adapt the allocation pattern. This leads to a fall in QoS (above 5-7 dB).

Finally, the maximum number of channel scans per allocation attempt Ch_{\max} , without power management, reveals an exponential gain which saturates at around 5-6 channels scanned per allocation attempt. However, combined with the effect of power management, no significant gain is observed. It appears that with power management, mobiles adjust their footprint thus offering more channel availability, such that only one channel test per allocation attempt is sufficient. And, optimization of the channel allocation occurs naturally through the many different attempts from all surrounding DBS in time. Therefore, complexity is kept to a minimum by having DBS only test and eventually allocate one channel per agent activation.

Connection management remains efficient as long as there are sufficient channel resources to provide a large enough channel footprint for the macrodiversity links. However, this minimum number of channels is low as macrodiversity links only require a small SINR to provide sufficient QoS after macrodiversity combination.

In the current simulated scenarios, 25 channels² shared by a 100 DBS for 1000 mobiles, including 2 reserved channels (for (re)connection), are enough to provide sufficient flexibility to the connection agents for them to be able to swap links for optimization (this is with the synergistic effect of power level management). Below this threshold, there is too much interference for the connectivity potential of DBS to be fully exploited for macrodiversity, and those resources are left unused.

² A cellular system with a channel reuse pattern of 7 hexagonal cells, would require 70 channels for 100 pico cells, without offering the flexibility the current architecture provides.

5.4 Power Control

This section begins with a description of two known PC algorithms adapted for the DBS context and used for benchmarking purpose. These are then compared through numerical experiments with the multi-agent-based power control (MAPC) method described in 4.4.

5.4.1 Centralized Power Control (CPC)

Grandhi's centralized power control (CPC) algorithm (Grandhi et al., 1993) is applied in the DBS architecture by considering the M master DBS for the M mobiles on a given channel with g_{ij} ($1 \leq i \leq M, 1 \leq j \leq M$) denoting the gain of the link from mobile i to DBS j , with DBS i being mobile i 's master connection. Matrix \mathbf{A} is defined as

$$A_{ij} = g_{ij}/g_{ii} \text{ if } i \neq j, \quad (15)$$

$$A_{ii} = 0. \quad (16)$$

And the SIR at the master DBS is defined as

$$\gamma_i = \frac{p_i}{\sum_{j=1}^M A_{ij} p_{ij}}. \quad (17)$$

The power level for each mobile is then given by the eigenvector associated with the largest positive eigenvalue of \mathbf{A} .

Note that this algorithm is not trying to maximize each mobile's SIR. Rather, it finds a set of power levels which maximizes the lowest SIR, thus leading to each mobile's SIR being equal to the minimum (maximized) SIR. Also, the obtained power levels are proportional to at least the eigenvector and thus need to be scaled to fit inside the mobiles' power level range. This is where the instability of this algorithm becomes apparent, since under certain interference conditions, if a mobile is very close to its master DBS, its power level will be very low. Yet, since its SIR is forced to be equal to the other mobiles' SIR, proportionally, the noise at the receiver will have a much stronger impact leading to very poor SINR. Supposing a mobile faces 1W interference power and 0.1W noise power, and emits 10W to obtain a SIR of 10dB, it has an SINR of 9.6dB. In contrast, consider a mobile faced with .1W of interference; it emits 1W to obtain the same SIR of 10dB, but has an SINR of 7dB, hence an effective penalty of half. In order to minimize this effect, the minimum power level should be high enough so that noise remains as much as possible negligible. Hence, the power levels will be scaled such that the maximum power level evaluated is set to the maximum mobile's range.

On the other hand, a more complex evaluation of such effects would make it possible to lower the maximum power level, keeping it as low as possible, and hence, maximizing the efficiency of the link quality versus the power used per mobile.

5.4.2 SIR-balanced macro power control (SBMPC)

Yanikomeroglu's SIR-balanced macro power control (SBMPC) (Yanikomeroglu & Sousa, 1998) proposes an interesting algorithm for CDMA distributed antennas using macrodiversity. The algorithm aims to balance, over all mobiles, each mobile's aggregate (sum) SIR over all antennas. This is a valid approach in the context of Rayleigh fading. However, as mentioned in the introduction, in Rice fading, two similar average SIR values can lead to two different BER figures given each may not have the same K factor (i.e. fading impact is more or less severe). As such, balancing SIR does not balance BER with the relative importance of line of sight components varying depending on mobiles' locations.

As the SBMPC article suggests, it is straightforward to adapt the algorithm to a general cellular system. For all mobiles on one channel, we define the matrix \mathbf{B} to describe the connections of mobile i (out of M mobiles) to DBS j (out of L DBS) such that

$$B_{ij} = 1 \text{ if mobile } i \text{ is relayed by DBS } j, \quad (18)$$

$$B_{ii} = 0 \text{ otherwise.} \quad (19)$$

The global SIR for mobile i is then

$$\gamma_{i,SBMPC} = \sum_{j=1}^L B_{ij} \frac{g_{ij} p_i}{\left(\sum_{k=1}^M g_{kj} p_i k \right) - g_{ij} p_i}. \quad (20)$$

This equation is rearranged to obtain the power level of mobiles $i = \{2, \dots, M\}$ given the mobile $i = 1$ in an iterative manner:

$$\mathbf{P}^{(0)} = \{p_i^{(0)}\} = \left\{ \left(\sum_{j=1}^L B_{ij} g_{ij} \right)^{-1} \right\}, \forall i, \quad (21)$$

$$\gamma_1^{(v)} = \sum_{j=1}^L B_{1j} \frac{g_{1j} p_1 1^{(v)}}{\left(\sum_{k=1}^M g_{kj} p_i k^{(v)} \right) - g_{1j} p_1 1^{(v)}}. \quad (22)$$

$$p_i 1^{(v+1)} = p_i 1^{(v)}, \quad (23)$$

$$p_i^{(v+1)} = \frac{\gamma_1^{(v)}}{\sum_{j=1}^L \frac{B_{ij} g_{ij}}{\left(\sum_{k=1}^M g_{kj} p_i k^{(v)} \right) - g_{ij} p_i i^{(v)}}}, \quad i \in \{2, \dots, M\}. \quad (24)$$

Just like in the previous algorithm, the obtained power level vector needs to be scaled to minimize the noise effect.

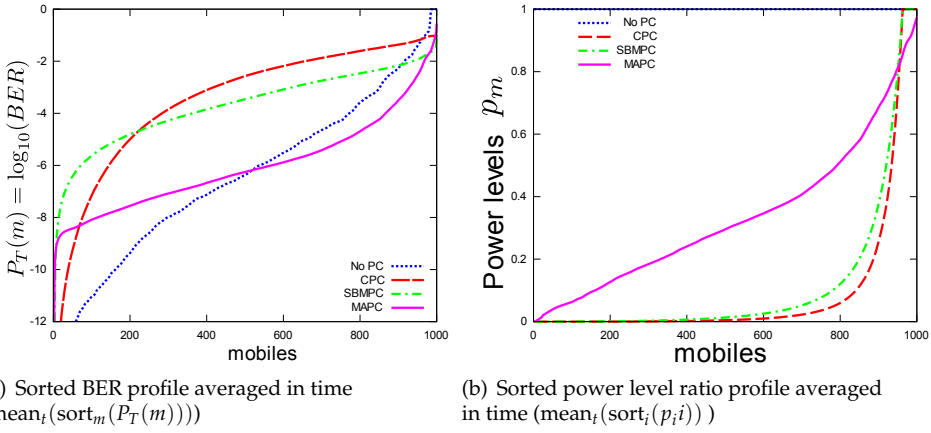
Given that this is an iterative solution and the purpose is not to evaluate its convergence, the algorithm is run for 20 iterations at each simulation time step of a simulation. It has been verified that this is enough to ensure convergence.

5.4.3 Results

Figure 8(a) shows the base results, that is, with a static simulation scenario where fading is otherwise accounted for as if mobiles were moving, but mobility is not considered (to observe a nominal capacity without taking into account dynamic adaptation of the algorithms). Noise is also not considered in this case.

The plot reveals different aspects. First, with no PC, the QoS is clearly not balanced, but more importantly, not all mobiles can be connected as is seen from the right hand side of the graph. With the centralized algorithms, we can clearly see that QoS is balanced, and all mobiles are connected. On the other hand, MAPC is able to provide much more QoS to almost all mobiles, while only impeding (compared to SBMPC) very few mobiles.

What is most interesting in Table 1 is how the CPC handles outage extremely well. It takes only 1 second (1 iteration of the simulation) to reconnect a lost mobile, and the probability that a mobile is disconnected is extremely low. Even SBMPC is not as good, but remains excellent compared to no PC. MAPC is only doing slightly worse, but with a much enhanced QoS provided to all mobiles.

Fig. 8. Ch = 40, $N = 0$, $V_{\max} = 0$.

	No PC	CPC	SBMPC	MAPC
$\bar{N}_d (\times 10^{-6})$	570	8.9	130	190
\bar{t}_r (seconds)	24.2	1	2.0	2.65

Table 1. Outage behavior without mobility and noise with Ch = 40.

Figure 8(b) reveals how both CPC and SBMPC offer similar distributions of the power levels. On the contrary, the MAPC power level allocation is radically different.

Faced with higher interference levels (Ch = 25), it can be seen that the centralized algorithms break down (Fig. 9(a)). Indeed, in high interference levels, maximizing the minimum SIR leads to very poor SIRs for all mobiles. In turn, this generates many disconnections. Figure 9(a) clearly shows that the traditional algorithms are here inefficient and even worse than without PC. Still the MAPC algorithm manages to provide acceptable levels of QoS, while still connecting more mobiles.

The situation deteriorates even more when noise is introduced. Figure 9(b) reveals how noise, as explained previously, renders the traditional algorithms unstable, generating lots of disconnections. Indeed, the centralized algorithms, by maximizing the minimum SIR, force most mobile power levels to be extremely low (c.f. Fig. 8(b)) supposing thermal noise is not an important factor. This may be valid for a regular hexagonal cell geometry with homogeneous traffic and high guaranteed SIR. However, it is not the case here, leading to very poor SINR, and also significantly exacerbating the HTE as such mobiles' presence on channels will not be sensed by other DBS, rendering the PC algorithm completely inefficient.

Also, facing important mobility (Figure 9(c)), the CPC algorithm loses its strength (of minimizing the outage probability) as Table 2 reveals. Indeed, with mobility, more interference is present because mobiles do not obtain an optimal reallocation of channels at each iteration. This implies far too many very low power levels with the CPC. This conflicts with the channel agents trying to reorganize the channel allocation as it generates important hidden terminal effects. This also shows that the CPC algorithm loses much of its capacity with even small

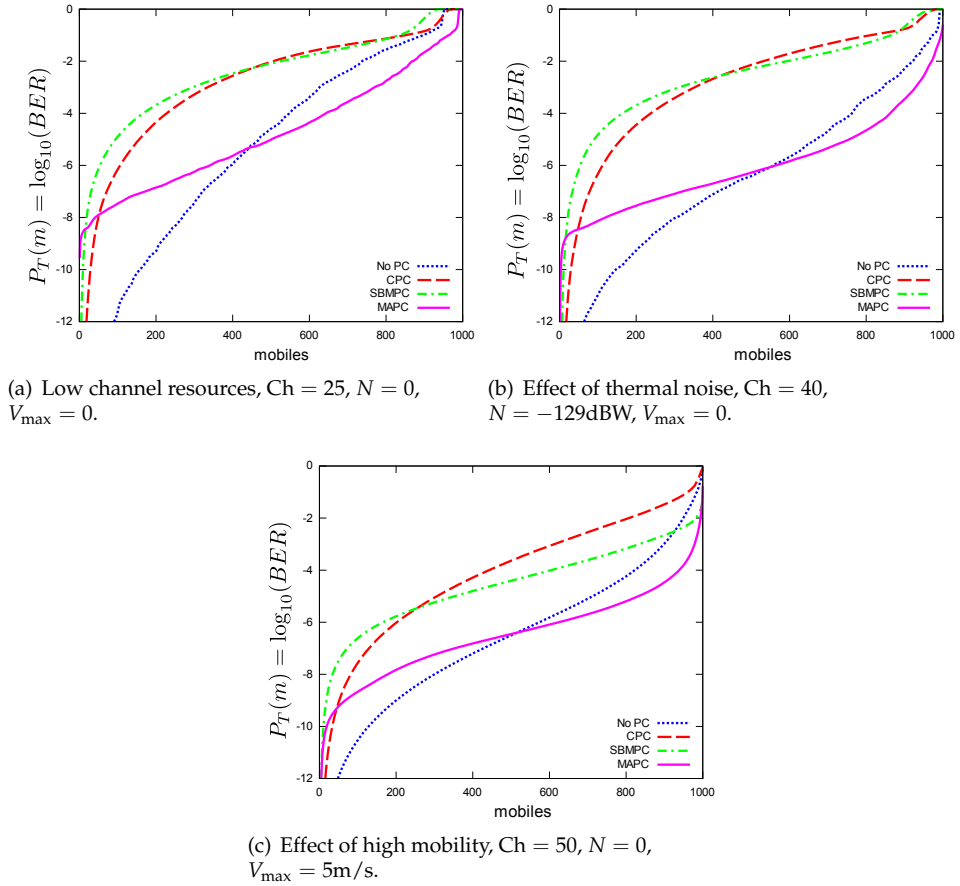
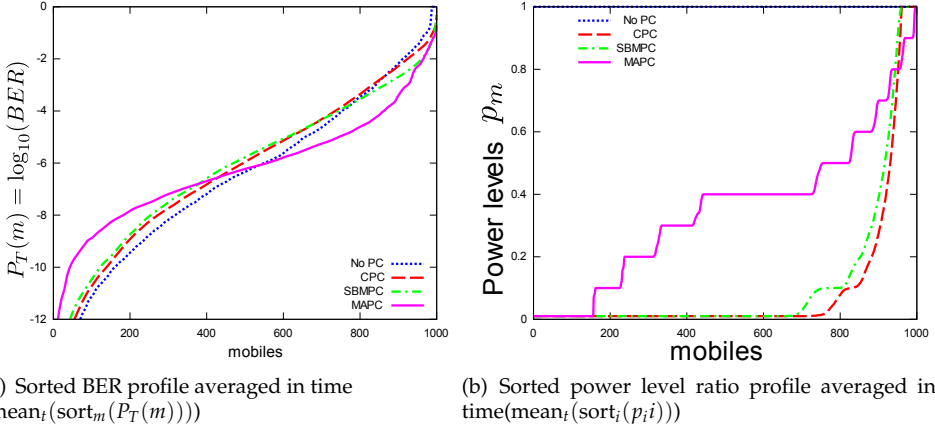


Fig. 9. Sorted BER profile averaged in time ($\text{mean}_t(\text{sort}_m(P_T(m)))$).

changes of the efficiency of the channel allocation. In addition, we are not even considering the burden of calculating power levels at each iteration while first centralizing the data bits, which necessarily takes time and would prevent the algorithm from rapidly adapting to the changes in the system. SBMPC is doing much better, yet not as well as MAPC.

Figure 10(a) shows the behavior of the algorithms when power levels are restricted to a set of discrete values. The power level range is uniformly divided into 10 discrete levels (ratios of 0.1 to 1.0). While the MAPC is not as efficient as with continuous power range, it still remains the most efficient at uniformly balancing the QoS. The CPC and SBMPC algorithms lose their ability to balance QoS across all mobiles, since in order for them to work properly they need an important dynamic range of power levels (which is not the case here with only a 10x range). Therefore, these schemes only manage to obtain marginal benefit, minimizing disconnections (compared to no PC), and visually seem better than with no discretization as more mobiles obtain more QoS (compared to Fig. 8(a)). In a way, discretization helps the

	No PC	CPC	SBMPC	MAPC
$\bar{N}_d (\times 10^{-4})$	9.8	19.1	3.8	1.8
\bar{t}_r (seconds)	5.13	2.4	2.2	1.38

Table 2. Outage behavior with mobility ($V_{\max} = 5$ m/s) Ch = 50.Fig. 10. Ch = 40, $N = 0$, $V_{\max} = 0$, discrete power level.

centralized schemes by preventing them from assigning excessively low power levels, yet also preventing them from achieving any proper resource balancing as the MAPC does.

Figure 10(b) shows the allocated power level over all mobiles in the discrete case.

Considering dynamics, it was found that tuning the different parameters mentioned in the design (most specifically β and S) makes the agent system stable enough to prevent erratic changes of the power levels, while still allowing fast adaptation in the wake of abrupt (discrete) changes in interference levels.

It is also noteworthy that the power management proposed provides additional benefits to the channel agents (lowering connection loss rate, block rate and mean time for reconnection) and connection agents (offering them more headroom to swap links for optimization). Despite being designed independently, the various described agents remain stable working together in a synergistic way, such that one change induced by one type of agent (e.g. a connection change) is correctly compensated by the other types of agents (e.g. an adaptation of the power level) and not overcompensated, which would otherwise lead to erratic behaviors.

One of the great advantages of this approach is that each dimension of the problem can be tackled independently, by its own class of agents, without affecting the performance of the other classes. No explicit interaction mechanism between the agent classes needs to be designed, yet they synergistically cooperate to achieve desirable results with a surprisingly low implementation complexity.

6. Conclusion

The proposed proof-of-concept design described herein demonstrates that minimalist multi-agent systems do provide all expected qualities: scalability, dynamic properties, efficiency, simplicity, adaptability and auto-configuration. Moreover, it represents a novel and surprisingly simple solution to resource allocation. It is most obvious with the power allocation scheme which results in drastically lower spatial power distributions when compared with traditional algorithms.

The multi agent design approach, based on heuristics appears effective. It requires an understanding of the underlying mechanisms and compromises within the context of the problem at hand, from which insights and intuition can be drawn and used to design the agents.

While it is unclear a priori to which end result the system will converge to, it should be noted that it is also unclear a priori to which it should. Indeed, the current context is far different from formal frameworks such as information theory which are often characterized by a single uni-dimension criterion, e.g. the channel capacity. In our multi-dimensional context, information theory remains too limited at the time to model and grasp the many possibilities and compromises facing a multitude of mobiles and DBS with macrodiversity where limited resources lead to interference. And in such a context, the proposed MA approach has the virtue of demonstrating via simulation that some novel allocation solutions (which should be understood as compromises) can lead to much higher efficiency of resource usage (where efficiency is necessarily also a notion of compromise).

The design in itself is not so complicated, and one should keep in mind the fuzziness of such an approach. Indeed, the utility functions proposed could have a variety of alternatives. What matters is not their exactness, but that they provide certain properties that will sustain the interactions of agents. These properties remain to be understood and studied to provide insights on the inner workings of the agent system. For example, the shaping function in the utility function of the connection agents uses a logarithm which could be replaced by a first degree approximation $(x - 1)$ and still converge, but a concave function with similar properties (e.g. $(x - 1)^2$, null for $x = 1$ and strictly increasing for $x > 1$) would, despite providing some degree of convergence, fall short of a more balanced solution. The shaping function is therefore crucial to converge to certain Pareto solutions, and this remains to be studied in detail.

Concerning MA design, we considered more specifically the notion of homeostasis which is not explicitly mentioned in Parunak's methodology. The proposed design shows how the search for such an equilibrium helps in designing and tuning the properties of the agents' behavior to obtain the desired global function.

Considering future work, the proposed MA design and DBS architecture offers malleability and vast margins for tuning, enhancing, or providing additional functionality. It was studied in (Leroux et al., 2008) that the reuse of channels could be enhanced by pairing mobiles to cooperate in exploiting a single channel while multiplying their diversity gain. Interesting results have been found in this study. Yet, coupled with power-level control, the management of cooperation between mobiles revealed counter synergistic effects. To date, finding a way to have the cooperation and power-control agents interoperate in a synergetic manner remains an open problem.

Another additional functionality to be studied is beamforming. Channel allocation agents would need to be improved to account for dynamically-created directive beams and provide network-wide gains by minimizing interference.

Channelization also needs to be further studied, including a model to implement the IEEE 802.11 shared random access mechanism (CSMA/CA on conjunction with the so-called dis-

tributed coordination function) which is now globally deployed, rather than relying only on orthogonal channels (whether it be time/frequency/ or code division) as is the case in this study. Finally, practical implementations should be tested, as discussed in Leroux (2008) where it was shown that macrodiversity could be obtained through minimal terminals synchronized at the packet level. It would then be possible to implement the proposed MA strategies using consumer Wi-Fi terminals and perhaps connect such terminals to a wired network, make them work in synergy and thus offer much more reliable and efficient connections. The proposed system is therefore not a simple exercise for MA design. It represents a meaningful starting point for a new design paradigm of mobile wireless networking. It offers vast potential for improvements, new designs and additional functionality.

7. References

- Akaiwa, Y. & Andoh, H. (1993). Channel segregation-a self-organized dynamic channel allocation method: application to TDMA/FDMA microcellular system, *IEEE J. Select. Areas Commun.* **11**(6): 949–954.
- Beongku, A., Dohyeon, K. & Innho, J. (2003). A modeling framework for supporting QoS in mobile ad-hoc networks, *Vehicular Technology Conference, 2003. VTC 2003-Spring. The 57th IEEE Semiannual* **2**: 935–939.
- Brueckner, S. & Parunak, H. V. D. (2003). Self-organizing MANET management, in G. D. M. Serugendo, A. Karageorgos, O. F. Rana & F. Zambonelli (eds), *Engineering Self-Organising Systems*, Vol. 2977 of *Lecture Notes in Computer Science*, Springer, pp. 20–35.
- Elwyn R. Berlekamp, John H. Conway & Richard K. Guy (1982). *Winning Ways for your Mathematical Plays*, New York: Academic Press.
- Furukawa, H. & Akaiwa, Y. (1994). A microcell overlaid with umbrella cell system, *Vehicular Technology Conference, 1994 IEEE 44th*, Stockholm, pp. 1455–1459.
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*, Kluwer Academic Publishers, Boston, MA.
- Grandhi, S., Vijayan, R., Goodman, D. & Zander, J. (1993). Centralized power control in cellular radio systems, *Vehicular Technology, IEEE Transactions on* **42**(4): 466–468.
- Leroux, P. (2008). *Architecture d'un système de stations de base distribuées*, PhD thesis, Université Laval.
- Leroux, P., Roy, S. & Chouinard, J.-Y. (2006). The Performance of Soft Macrodiversity Based on Maximal-Ratio Combining in Uncorrelated Rician Fading, *17th annual IEEE international symposium PIMRC'06*, Helsinki, Finland.
- Leroux, P., Roy, S. & Chouinard, J.-Y. (2008). Synergetic cooperation in a distributed base station system, *Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium on*, pp. 1–6.
- Mackenzie, A. & Wicker, S. (2001). Game Theory and the Design of Self Configuring, Adaptive Wireless Networks, *IEEE Commun. Mag.* .
- Muraleedharan, R. & Osadciw, L. A. (2003). Balancing the Performance of a Sensor Network Using an Ant System, *37th Annual Conference on Information Sciences and Systems (CISS 2003)*, Baltimore, MD. .
- Parunak, H. V. D. (1997). “Go to the Ant” : Engineering Principles from Natural Agent Systems, *Annals of Operations Research* .
- Tumer, K. & Wolpert, D. (2004). *Collectives and the Design of Complex Systems*, Springer-Verlag, NY.

Inter-RAT Handover Between UMTS And WiMAX

Bin LIU and Philippe Martins

*Télécom ParisTech - École Nationale Supérieure des Télécommunications
France*

Philippe Bertin and Abed Ellatif Samhat

*France Telecom Research and Development
France*

1. Introduction

The future beyond third generation (B3G) or fourth generation (4G) systems will consist of different radio access technologies, such as GSM/GPRS, UMTS, WiFi, and WiMAX. Many intensive efforts have been made to identify the unsolved issues about the future mobile systems, and one important issue is what the future vertical handover management solution will be. A variety of mobility management solutions have been proposed, such as MIPv6/FMIPv6 (D. Johnson, et al., 2004; R. Koodli, 2005), SCTP (M. Afif, et al., 2006), inter-RAT (Radio Access Technologies) handover of 3GPP (3GPP TS 43.129; 3GPP TR 25.931). Among these solutions, the layer 2 inter-RAT handover solution of 3GPP is a promising way for its high reliable handover procedure. Unfortunately, the 3GPP inter-RAT solutions only support inter-RAT handover between cellular networks, and do not support inter-RAT handover between WiMAX (Worldwide Interoperability for Microwave Access) and UMTS (Universal Mobile Telecommunications System).

Another important issue is the interworking architecture and the coupling scenario that are used to provide an efficient inter-RAT handover management. Depending on where is the coupling point, there are several interworking architectures: no coupling, loose coupling, tight coupling, very tight coupling (integrated coupling) (G. Lanpropoulos, et al., 2005). The loose coupling and tight coupling architectures often use Mobile IP or part of Mobile IP as the handover management protocol. So these two kinds of coupling architectures require less complicated modifications to the existing protocol stacks and are more flexible than integrated coupling. However, they often suffer from longer handover latency varying from some hundreds of milliseconds to some seconds. The integrated coupling generally achieves better handover performance at expense of adding complex modification to existing network protocol stacks.

In recent years, the 3GPP and IEEE organizations have proposed their respective interworking solutions for convergence of heterogeneous networks. For instance, the ongoing 3GPP standard for interworking between UMTS and WiFi (3GPP TS 23.234) only focuses on the control plane, and defines the interworking topologies, access gateways,

AAA, charging, interfaces and so on. It does not provide a scheme to resolve the handover problems in the user plane, like packet loss, long handover latency. Another promising vertical handover solution is IEEE 802.21 (IEEE 802.21, 2006). This standard defines a generic link layer to mask the heterogeneities of various RATs as well as three kinds of services. How to resolve handover problems is still manufacturers' task. For these reasons, in this chapter, we will only focus on the user plane instead of control plane, and propose an inter-RAT handover solution to resolve some typical handover problems. Besides, our inter-RAT handover solution can be applied to a variety of interworking scenarios in addition to the integrated and tight coupling architectures. This solution is a novel solution for interworking between UMTS and WiMAX, which has not been stated by other references or standards.

In section 2, we firstly summarize the features of 3GPP Packet Switched (PS) network/cell switch procedures, i.e. reselection and handover, and then we get some guides to the design of inter-RAT handover mechanism between UMTS and WiMAX. The PS handover is introduced in order to support real-time packet-switched traffics with strict QoS requirements on low latency and packet loss. For one thing, the handover reduces the service interruption of the user plane information when cell changes compared to the cell reselection; for another, it enables buffer handling of user plane data in order to reduce packet loss when cell changes. For unreal-time services with loose QoS requirements on packet loss, or Mobile Station (MS) is not in dedicated state, the PS cell reselection is introduced. Compared to PS handover, the cell reselection suffers from uncertain packet loss, but benefits from the reduced signaling and resource overhead between cells or RATs. Next, in section 3 and 4 respectively, based on the requirements of inter-RAT handover between UMTS and WiMAX, we propose a novel layer 2 inter-RAT handover scheme by introducing a novel common sublayer named IW (InterWorking) sublayer and SR ARQ (Selective Repeat ARQ) mechanism in the integrated and tight coupling architectures to resolve several typical inter-RAT handover problems, such as packet loss, long handover latency. The better handover performance is validated by the simulation results carried out on the NS2 emulator. In addition, this novel IW sublayer scheme also eliminates the false fast retransmission, which is due to packet loss or out-of-order packet arrivals during a handover period.

Finally, we come to our conclusions in section 5.

2. 3GPP Handover Features

This section mainly describes procedures that are used by the MS to select a suitable cell to be connected to in the 3GPP cellular networks. For details, please refer to references (3GPP TS 25.331; 3GPP TS 43.129; 3GPP TS 44.060; 3GPP TR 25.931; J.P Romero, et al., 2005).

When a MS is switched on, it must first select a PLMN (Public Land Mobile Network) and a RAT (Radio Access Technology) automatically or manually. It searches for the most adequate cells so as to camp on a suitable cell. Then the MS will register its presence in the registration area of the chosen cell if necessary. As long as the MS remains in idle mode, it will continuously execute the cell reselection procedures in order to choose and camp on the most suitable cell of the selected PLMN. The events triggering the cell reselection may due to high path loss, downlink signaling failure and so on. If the new cell is in a different registration area, a Location Registration (LR) request is performed (3GPP TS 23.122).

When the MS is in the packet transfer or dedicated mode, the network use PS Handover to command a MS to move from its source cell to a new cell, and continue the ongoing PS service operation in the new cell. The handover trigger conditions could be serving cell resource limitation, measurement reports from a MS, and the cell change notification from a MS. This handover procedure between cells of the same RAT is also referred to as *intra-RAT handover*.

Depending on the PLMN availability and network configuration, it is possible that cell reselection and handover procedures involve a cell switch from one RAT to another RAT (e.g., from 2G GSM/GPRS/EDGE to 3G UMTS networks). The handover procedure between cells of different RATs is referred to as *inter-RAT handover* in 3GPP standards or *vertical handover* in IETF protocols.

In general, in contrast with the cell reselection, the intra-RAT handover and inter-RAT handover of 3GPP have the following distinct features:

- The network makes the handover decision depending on the measurement reports, network states and negotiation with target base station or cell.
- The network controls the whole handover procedure, including message transfer, handover timing, handover target, resource allocation, and context transfer.
- Only packet transfer mode or dedicated mode needs the handover procedure due to handover overheads.
- The handover can be considered as a kind of specific QoS-guaranteed cell reselection procedure.

It should be stressed that Mobile IPv6 and its extensions (D. Johnson, et. al., 2004; R.Koodli, 2005) support both network-initiated and terminal-initiated handover procedures. In this chapter, we only consider 3GPP inter-RAT handover procedures initiated by the network.

The conventional 3GPP inter-RAT handover procedure must involve the SGSN, whether for handover from GSM/GPRS to UMTS or from UMTS to GSM/GPRS. This is because the Link Control Sublayer (LLC) terminates at SGSN, which is in charge of making lossless packets forwarding during the MS mobility thanks to its retransmission mechanism. This is not a problem for GSM/GPRS or UMTS core network, because the SGSN is their common network entity. But for inter-RAT handover between 3GPP cellular network and IEEE wireless IP network like WiFi/WiMAX, it becomes a challenge for the packet lossless RAT switch procedure due to the lack of SGSN in IEEE wireless network. In the following sections, we utilize the 3GPP inter-RAT handover procedure, messages and signaling to resolve this problem for two typical coupling network architectures: integrated coupling and tight coupling.

3. Inter-RAT Handover between UMTS and WiMAX in Integrated Coupling Architecture

In order to realize a seamless inter-RAT handover for future B3G or 4G mobile networks, a variety of interworking architectures and inter-RAT handover mobility managements have been proposed. Based on the integrated architecture, in this section, a novel common interworking sublayer (IW sublayer) is proposed at Layer 2 on RNC and MS to provide a seamless PS inter-RAT handover between UMTS and WiMAX systems. This IW sublayer scheme focuses on eliminating packet loss and reducing handover latency that are common problems for most inter-RAT handover scenarios. Compared with other context transfer

schemes, the simulation results show the IW sublayer with ARQ mechanism can achieve a lossless and prompt handover procedure. In addition, this IW sublayer scheme can eliminate false fast retransmission of TCP traffics that is usually caused by packet losses or out-of-order packet arrivals. In what follows, the IW sublayers in integrated and tight coupling architectures are specified in this section and section 4 respectively.

3.1 Context Transfer

The problems during the inter-RAT handover period have been extensively studied in (G. Lanpropoulos, et al., 2005; S.L. Tsao, et al., 2002; N. Dailly, et al., 2006; J. Sachs, et al., 2006; H. Inaura, et al., 2003; H. Rutagemwa, et al., 2007), such as long handover latency, BDP (Bandwidth Delay Product) mismatch, delay spikes, packet losses, premature timeout, false fast retransmission. Among these problems, the packet losses and long handover latency are in particular not desirable for real-time and throughput-sensitive traffics. The most common solution is applying the context transfer mechanism (J. Sachs, et al., 2006; R. Koodli, 2005) or retransmission (3GPP TS 25.323; N. Dailly, et al., 2006) to accelerate the handover process or reduce the amount of lost packets. There exist the following typical context transfer and retransmission schemes: PDCP Synchronization, buffering-and-forwarding (B&F), SDU Reconstruction, R-LLC.

PDCP Synchronization: In 3GPP UMTS network (3GPP TS 25.323), the PDCP sublayer is applied to guarantee reliable data transmission service during Service Radio Network Subsystem (SRNS) relocation. For this purpose, PDCP maintains PDCP sequence numbers to avoid any data losses during SRNS relocation. After the successful relocation, the data transmission starts from the (first) unconfirmed SDU having a sequence number equal to the next expected sequence number by the PDCP entity. For instance, in the uplink, if some transmitted SDUs are still left unacknowledged, the data transmission is resumed by retransmission of the SDU with the “uplink send” sequence number equal to the uplink receive sequence number. Otherwise, the data transmission is resumed with the transmission of the first unsent SDU. Moreover, when the RLC entities mapped to a lossless PDCP entity are reset or reestablished for reasons other than SRNS relocation, then it is the PDCP’s responsibility that the peer lossless PDCP entities do not go out of synchronization by the means of “PDCP sequence number synchronization procedure” (3GPP TS 25.323).

The retransmission mechanism of PDCP works well when a MS performs SRNS relocation during data transmission in the domain of UMTS. Unfortunately, in the scenario of inter-RAT handover, the PDCP will not take effect any more because:

- The other heterogeneous network system usually does not have the similar mechanism, especially IEEE 802 RATs such as WiMAX or WiFi.
- In addition, the WiMAX system has its own IP packet header compression mechanism rather than ROHC (3GPP TS 25.323) in UMTS. If the packets or frames stored in the source system with their particular headers and control signaling parts are forwarded to the target systems directly, the target system may discard these unreadable packets, which will induce sequence number asynchronization and break down current communication connection.

In a word, the sequence number synchronization, header compression and retransmission mechanism of respective RAT complicate the inter-RAT handover procedure instead.

Buffering-and-Forwarding: R. Koodli (2005) propose to utilize buffering-and-forwarding mechanism (B&F) to forward unsent data packets from previous access router to new access

router in FMIPv6 to eliminate packet losses during a handover period. In this IP solution the packets stored at link layer of one RAT usually cannot be retrieved to the IP layer. So, generally, there usually exist packet losses if an inter-RAT handover management protocol with B&F is realized only at IP layer or above.

SDU Reconstruction: J. Sachs (2006) proposes the SDU (Service Data Unit) reconstruction scheme. In order to make a lossless handover, the segments stored in the PDU (Packet Data Unit) buffer of source link are first reconstructed back to a SDU and then forwarded to the target link as well as the SDUs from the SDU buffer. When this proposal is applied to the inter-RAT handover from UMTS to WiMAX or from WiMAX to UMTS for TCP traffics, the handover performance may still degrade, for the following reasons:

- If a PDCP (Packet Data Convergence Protocol) PDU sequence number asynchronization takes place just before an inter-RAT handover from UMTS to WiMAX, the asynchronization problem cannot be resolved by WiMAX system. This leads to an unreliable handover procedure.
- A RLC (Radio Link Control) SDU whose corresponding RLC PDUs have not be successfully transmitted in total, cannot be reconstructed and will be discarded locally, because the successfully transmitted RLC PDUs have been already removed before.
- WiMAX and UMTS systems support different header compression algorithms. That means it needs a deliberate loop back setting in respective system.

From above discussion we can see that the PDCP sequence number synchronization mechanism, which is used to assure lossless handover in UMTS system, becomes an obstacle for inter-RAT handover when the context transfer mechanism is SDU reconstruction. We may suppose the PDCP sublayer is configured in transparent mode (i.e., not attach any PDCP header, no header compression and no sequence number synchronization) when the inter-RAT handover is based on this SDU reconstruction mechanism. But, what the point of using an UMTS network without PDCP sublayer?

R-LLC: N. Dailly (2006) introduce a novel sublayer called R-LLC (Link Layer Control) locating on the BTS for handover between GPRS and WiFi is proposed. This R-LLC sublayer takes the role of conventional LLC, and retransmits packets lost during inter-RAT handover when the retransmission timer expires. The simulation results in this reference (N. Dailly, et al., 2006) demonstrate zero packet loss for handover and cell reselection procedures. However, the packet loss is only indicated by retransmission timer expiration, which usually is set to 5 sec.. Obviously, such a long period is unfavorable to keep TCP congestion window from shrinking. In addition, the configuration of retransmission window is not specified.

Since these typical context transfer or retransmission schemes do not satisfy the inter-RAT handover requirements for interworking UMTS and WiMAX, in what follows, a novel context transfer scheme is proposed at a novel common sublayer – InterWorking (IW) sublayer.

3.2 InterWorking (IW) Sublayer

3.2.1 InterWorking (IW) Sublayer Description

In order to propose a feasible way to provide a seamless handover procedure, a promising solution is designed in terms of the following principles:

- Minimize altering the existing standards and protocol stacks as far as possible.

- Consider the UMTS as the center of the integrated system and preserve its signaling and control procedures as many as possible.
- In WiMAX access network, additional network components and new signaling and primitives could be added.
- The two integrated systems should guarantee seamless service continuity, and execute mobility management processes (e.g., connection establishment, handover) as fast as possible in order to maintain the required QoS.

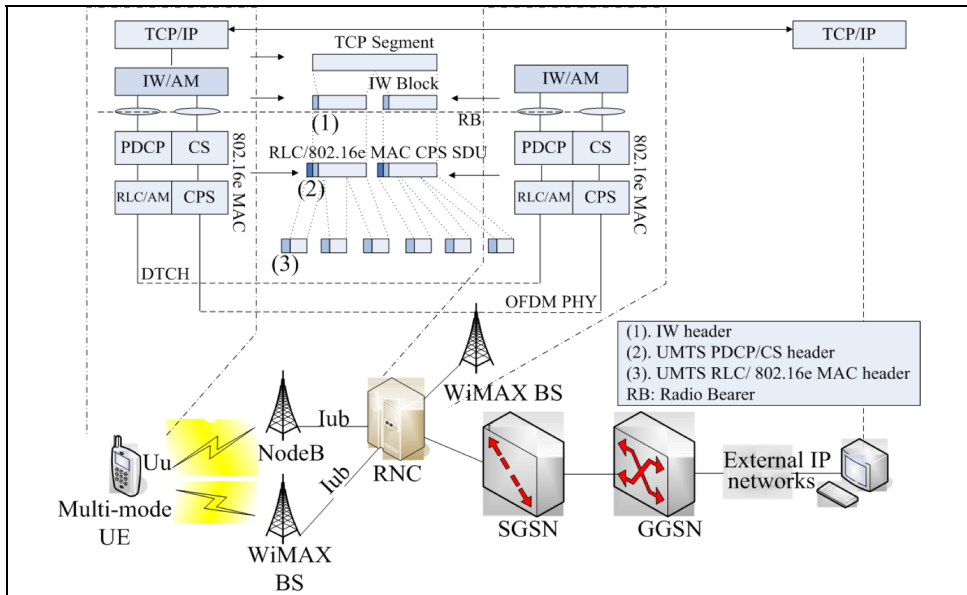


Fig. 1. IW sublayer working mechanism of integrated coupling

As stated above, our inter-RAT scheme is first based on the integrated coupling architecture. We assume UMTS to be the master home network with roaming privileges to WiMAX network. A novel common network entity named interworking sublayer (IW) is introduced on the top of PDCP (Packet Data Convergence Protocol) sublayer of UMTS and the Medium Access Control (MAC) CS sublayer of 802.16e on the RNC and MS, as shown in Fig.1. The WiMAX BS is integrated with the RNC (Radio Network Controller) through Iub interface. The IW takes the role of LLC sublayer of conventional cellular networks, such as retransmission mechanism and handover support. The main functions of IW sublayer are:

- Determination of a suitable target network.
- Primitive mapping between the IW and the UMTS network, or between the IW and the WiMAX network in case of inter-RAT handover.
- Support SR ARQ (Selective Repeat ARQ) mechanism, including packet segmentation and re-sequencing, retransmission, and retransmission window size adjustment.

In Fig. 2 and Fig. 3, the user and control planes of the proposed architectures are illustrated. It should be stressed that the SR ARQ retransmission mechanism is realized in the user plane. While in the control plane, IW sublayer shall translate handover related signaling

between source and target networks. When an inter-RAT handover is made, the IW sublayer is activated according to the QoS requirements of a PDP (Packet Data Protocol). In the control plane, we prefer to reuse the RRC protocol functionality in the MS and RNC respectively, instead of building them from scratch. What is actually needed is to enhance RRC protocol entities in order to forward inter-RAT handover primitives to IW sublayer. In order to minimize the modifications to respective systems, IW sublayer also realizes some essential WiMAX-related primitives and makes RNC act as another WiMAX BS to the WiMAX network.

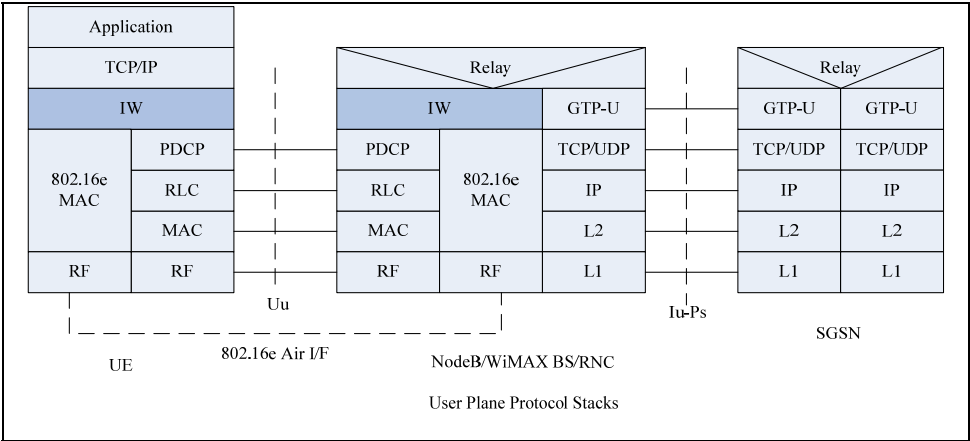


Fig. 2. User plane protocol stacks in integrated coupling architecture

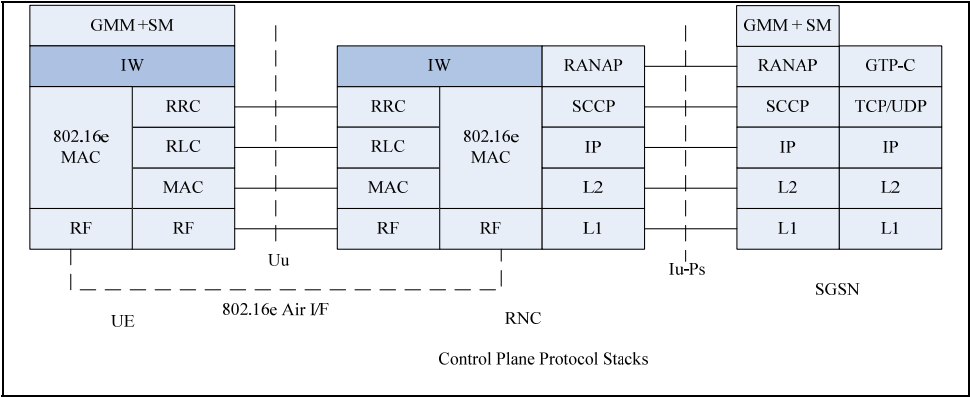


Fig. 3. Control plane protocol stacks in integrated coupling architecture

3.2.2 Signaling and Primitives

3.2.2.1 Overview

In order to have insight into IW sublayer working mechanism, this sub-clause describes the inter-RAT handover signaling procedures and primitives among IW, PDCP, RRC (Radio Resource Control) and WiMAX MAC. Some newly added cross-layer primitives are complemented to the conventional inter-RAT handover signaling procedures of 3GPP (3GPP TS 43.129; 3GPP TR 25.931). We suggest the future WiMAX and UMTS standards should support these primitives and parameters for the seamless and smooth inter-RAT handover.

Generally, the inter-RAT handover consists of handover preparation phase and handover execution phase. In the case of a handover from UMTS to WiMAX, when the inter-RAT handover conditions e.g. low RSSI or load increase, are met, the MS is instructed by the RNC to switch on its WiMAX transceiver. Then MS seeks and monitors the neighbor WiMAX BSs given in System Information Block (SIB) on BCCH of serving cell. After the WiMAX scanning intervals (IEEE 802.16e, 2005), the MS provides the network with its measurement results of the target networks using Measurement Reports message. Meanwhile, other important wireless link parameters, such as round trip time (RTT), BDP are also calculated by the RNC. After that, the inter-RAT handover will enter into execution phase if the RNC makes a positive handover decision.

3.2.2.2 Handover from UMTS to WiMAX

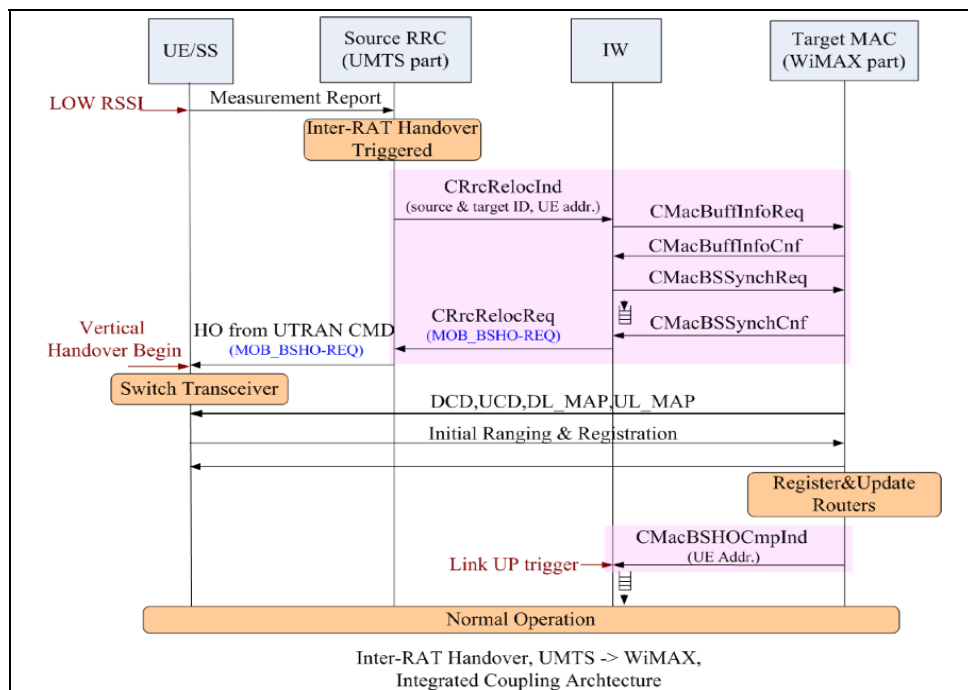


Fig. 4. Handover signaling procedure from UMTS to WiMAX

Fig. 4 describes the inter-RAT handover from UMTS to WiMAX and shows the exchanged messages and primitives.

- 1) Based on measurement reports and knowledge of the RAN topology, the RNC, more precisely source RRC decides to initiate an inter-RAT PS handover.
- 2) The source RRC sends the CRrcRelocInd primitive (contains target WiMAX cell id) to the IW sublayer.
- 3) Then the IW sends the CMacBuffInfoReq primitive to the target WiMAX MAC to request the buffer characteristics. The WiMAX MAC shall return the CMacBuffInfoCnf primitive to inform the IW of the buffer size in its MAC sublayer. According to this information, the IW adjusts its retransmission window size. (Note that current WiMAX MAC does not support this interface, so the IW may adjust its retransmission window size to a default value).
- 4) At this stage, the IW sends the CMacBSSynchReq primitive to the WiMAX MAC to negotiate the location of the dedicated initial ranging transmission opportunity for the MS. This information is returned by primitive CMacBSSynchCnf.
- 5) After that, the IW begins to buffer data packets that require delivery order and sends a CRrcRelocReq primitive (including Transparent Container (MOB_BSHO-REQ)) to the source RRC.
- 6) The RRC sends the Handover from UTRAN Command message to the MS, which includes a MOB_BSHO-REQ.
- 7) The MS performs hard handover and normal WiMAX network entry procedure.
- 8) After the provisioned service flow is activated (IEEE 802.16e, 2005), the target WiMAX MAC sends CMacBSHOCmpInd primitive as a Link_Up (LU) trigger to the IW sublayer. On this trigger, the IW shall restart data packet forwarding.

3.2.2.3 Handover from WiMAX to UMTS

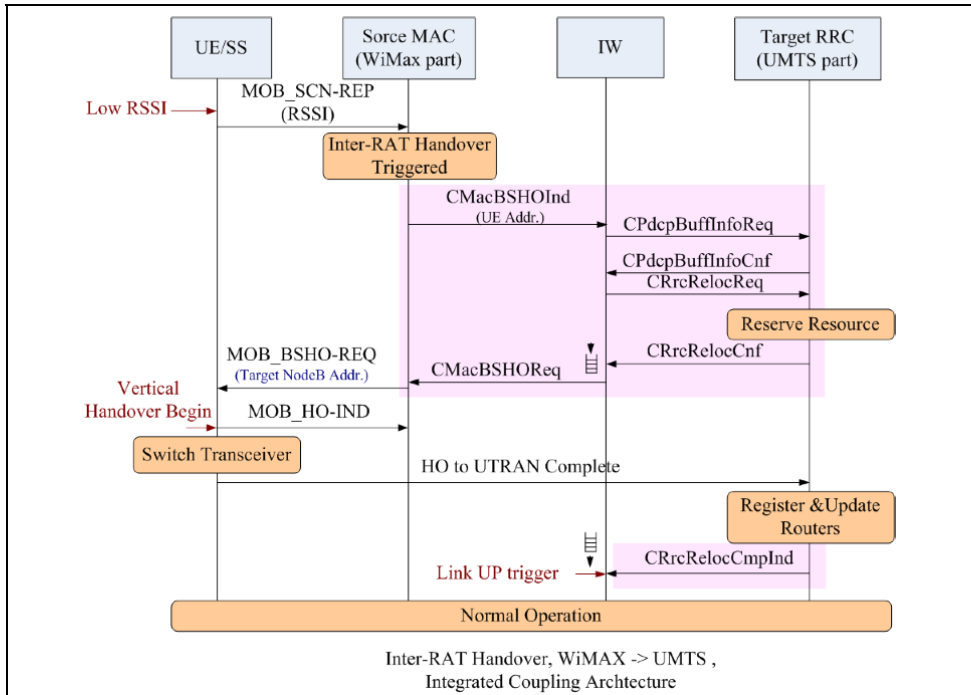


Fig. 5. Handover signalling procedure from WiMAX to UMTS

The inter-RAT handover from WiMAX to UMTS is described in Fig. 5.

- 1) After the scanning interval, the MS sends scanning report to WiMAX serving BS by message MON_SCN-REP that contains physical information such as mean RSSI.
- 2) The source WiMAX MAC sends CMacBSHOInd primitive to inform the IW sublayer of target cell id. The IW then sends CPdcpBuffInfoReq primitive to the target RRC of the UMTS network. The RRC shall return the CPdcpBuffInfoCnf primitive to inform the IW sublayer of buffer size and buffer occupation. According to this information, the IW adjusts its retransmission window size.
- 3) The IW sublayer sends a CRrcRelocReq primitive to the target RRC to apply for resource allocation. The result is returned in CRrcRelocCnf primitive by the target RRC.
- 4) Upon receipt of the CRrcRelocCnf, the IW suspends and buffers data packets that require delivery order.
- 5) IW sends CMacBSHOReq primitive to inform source MAC that the target network is ready.
- 6) The MS performs handover to one of BSs specified in MOB_BSHO-REQ and responds with a MOB_HO-IND message.
- 7) MS performs normal UMTS hard handover.
- 8) After the MS successfully finishes UMTS radio link setup, the target RRC shall send the CRrcRelocCmpInd primitive to the IW, and the IW restarts data packet forwarding.

Note that primitive CMacBSHOCmpInd and primitive CRrcRelocCmpInd are defined as the Link_Up (LU) triggers for handover from UMTS to WiMAX and for handover from WiMAX to UMTS respectively.

3.2.2.4 IW ARQ Mechanism

For the sake of achieving lossless inter-RAT handover, a modified Selective Repeat ARQ (SR ARQ) mechanism is applied to the IW sublayer during the handover period, which is renamed IW ARQ. The IW ARQ is an error control mechanism that involves error detection and retransmission of lost or corrupted packets. When a packet is accepted from upper layer, it is segmented into smaller IW blocks, each of which is assigned a sequence number (see Fig. 6). This new IW sub-header is used for block loss detection and block re-sequencing in the receiver to guarantee in-sequence delivery. Afterward, each IW block is transmitted through the UMTS or the WiMAX interface. These IW blocks are also queued in the retransmission buffer in order to be scheduled for retransmission. The IW ARQ transmitter maintains an adaptive window size that is set to target network buffer size. When an IW block is received by the receiver, a positive or negative acknowledgement (ACK/NACK) is sent back immediately for the purpose of reducing handover latency. In addition, in order to avoid dead lock due to IW ACK/NACK losses during a handover period, a status report timer is set when the receiver sends an ACK/NACK. When this timer expires, the receiver sends back a status report (ARQ feedback bitmap) providing the receipt status. This status report is map of the acknowledgement (ACK) or negative acknowledge (NACK) of each IW block within the window. Compared with conventional SR ARQ mechanism of RLC, the IW ARQ has the following features:

- **Receiver-Driven scheme:** the received status and ACK/NACK are sent back on receipt of an IW block initiatively without transmitter's polling message.
- **Support Link_Up (LU) trigger:** when a handover is finished, the target network will signal the IW sublayer with a Link_UP trigger. On receipt of this trigger, the IW sublayer will retransmit blocks in retransmission buffer to avoid unnecessary waiting for an expiration of the status report timer.
- **Adaptive Window Size:** In order to avoid any buffer overflow in the target network when the packets are retransmitted by the IW sublayer after a handover, the IW ARQ window size is adaptively set to buffer size of the target network.

In Fig. 6, an example of the IW ARQ mechanism when the window size is 12 is depicted. The right parts are two retransmission mechanisms: IW ARQ and R-LLC. In this figure, the difference between them is in that the lost blocks are retransmitted when status report timer expires in R-LLC scheme, while IW ARQ retransmits unacknowledged blocks not only on the expiration of this timer but also on the Link_Up trigger.

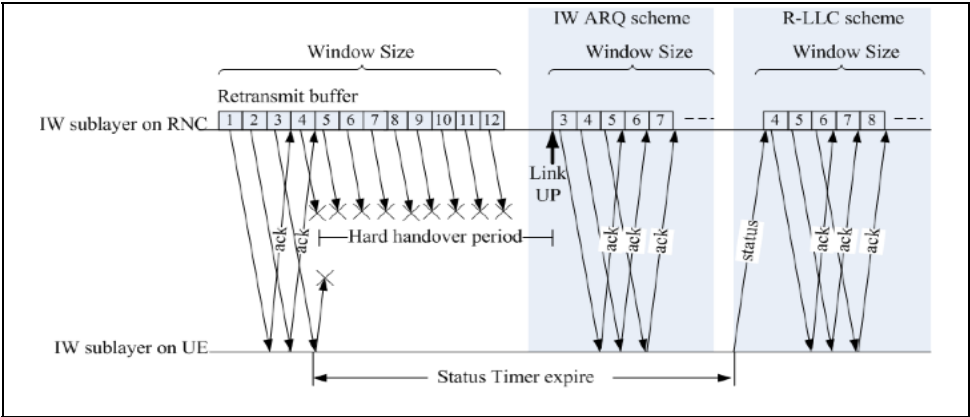


Fig. 6. IW ARQ and R-LLC protocol: an example of time evolution

3.3 Simulation Environment

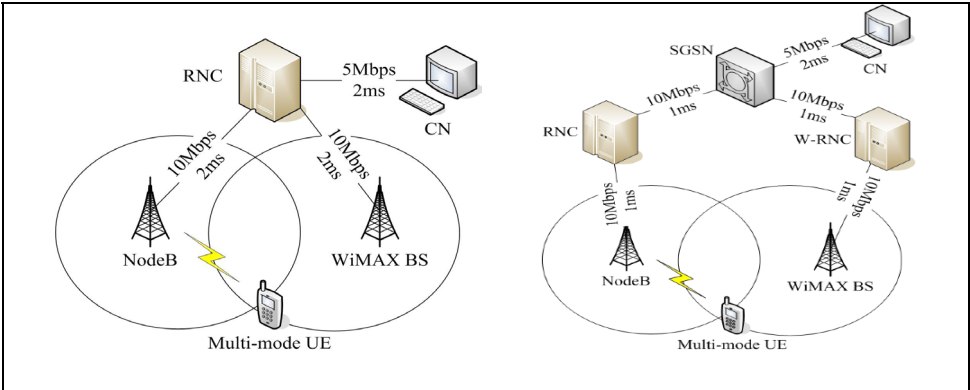


Fig. 7. Simulation topologies: integrated coupling (left) and tight coupling (right)

In order to analyze the performance of the IW sublayer during inter-RAT handover between UMTS and WiMAX, network-level simulations are carried out using NS2. Several extensions are made to this simulator, UMTS and WiMAX models, IW sublayer, multi-channel model, IW ARQ mechanism and new signaling and primitives. The simulation topologies for integrated and tight coupling scenarios are illustrated in Fig. 7. There is only one MS with two transceivers and no other background traffics in this “clean” scenario. The MS always has enough bandwidth to send packet whether it is in WiMAX region or in UMTS region. Note that in this topology, the transmission delay in the wired network is set very small deliberately to minimize its influence to handover procedure. An FTP session is examined, with the CN designated as the sender and the MS designated as the receiver. In UMTS module, a drop-tail policy is applied to radio network queues in PDCP and this queue length is set to 25 IW blocks. As to the WiMAX module, the queue length is set to 50 IW blocks, which considers the fact that generally the bandwidth of WiMAX is much higher. Other important simulation parameters are summarized in Table 1.

	Parameter	Value		Parameter	Value
IW	Fragment Switch	OFF	UMTS PHY	TTI (ms)	10
	Max retransmit count	10		Frame Duration(ms)	10
				BLER	1e-6
	Default Windows size (block)	30	WiMAX MAC	Allocated data rate	unlimited
	Status Report Timer (s)	2.5		Queue length	50
PDCP	TCP/IP Header compression, and Retransmission	no		Payload Header Suppression	no
	Allocated data rate	64kb/s		Frame duration (ms)	4
	Queue length	25	WiMAX PHY	Modulation	OFDM
RLC	RLC Mode	AM		Interleaving interval (frames)	50
	Windows size (Blocks)	500		FFT	256
	Block size (Bytes)	20		Number of subcarrier used	200
	maxDAT	20	TCP/IP	Variant	Reno
	Ack timerout period (ms)	50		MSS (bytes)	512
				Default cwnd	32

Table 1. Simulation Parameters

3.4 Simulation Results in the Integrated Coupling Scenario

3.4.1 Handover from UMTS to WiMAX

For the simulation of inter-RAT handover from UMTS network to the WiMAX, an FTP session starts at 0.4 sec., and the MS starts to perform handover at about 4 sec. after it enters into the coverage region of WiMAX. The handover type is hard handover. At about 4.035 sec. the WiMAX network entry procedure is finished and the IW sublayer on the RNC receives a Link_Up trigger. Fig. 8 shows the packet flows of three kinds of context transfer schemes: R-LLC, SDU Reconstruction and IW ARQ.

The R-LLC scheme does not support Link_Up trigger, so it retransmits the last unacknowledged data packets on the expiration of status report timer. During this period, the TCP timer expires and the congestion window shrinks to one, as shown in Fig. 9. There is a retransmitted TCP segment at about 5.7 sec.

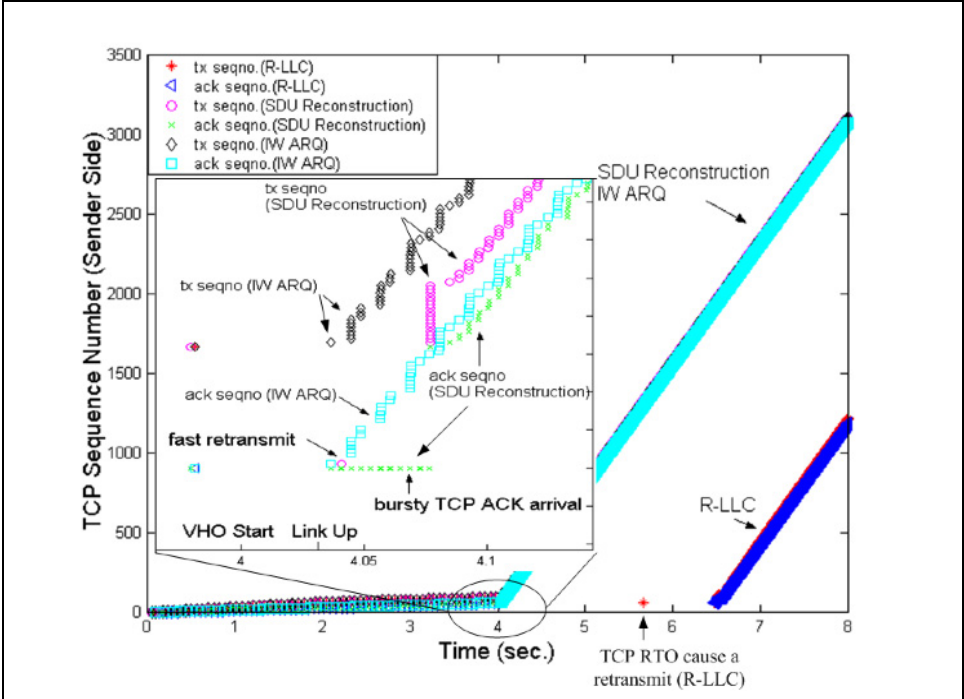


Fig. 8. TCP segment number comparison (umts->wimax, sender side)

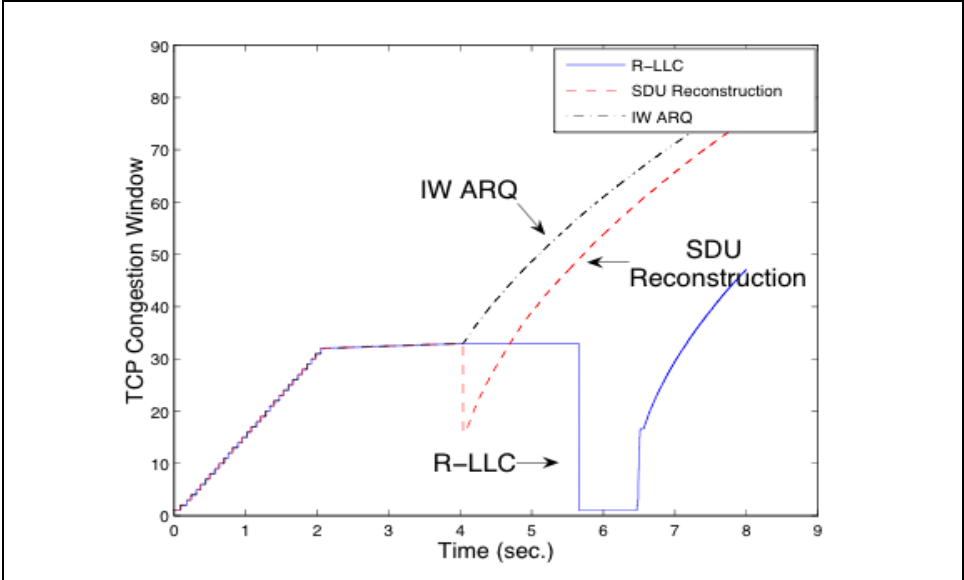


Fig. 9. TCP congestion window (umts->wimax)

The SDU Reconstruction scheme reconstructs the RLC PDUs stored in the RLC retransmission buffer. However, if one PDU of a SDU is successfully transmitted, this PDU is deleted from retransmission buffer and the remaining PDUs of this SDU cannot be reconstructed and are discarded locally. The remaining RLC SDUs (TCP packets here) are forwarded to WiMAX network after handover on RNC. These arrivals of out-of-order packets generate several duplicate ACK and trigger TCP fast retransmission process. The TCP congestion window size shrinks to half of congestion window of steady state, and the average throughput is also reduced.

The IW ARQ scheme adjusts its retransmission window according to the target network's queue size and forwards the IW blocks in its retransmission buffer on receipt of Link_Up trigger. After handover is over, there are no packet losses and the TCP ACK arrivals are not as bursty as those of SDU Reconstruction scheme thanks to the IW ARQ window mechanism (see Fig. 8 between 4 and 4.1 sec.).

3.4.2 Handover from WiMAX to UMTS

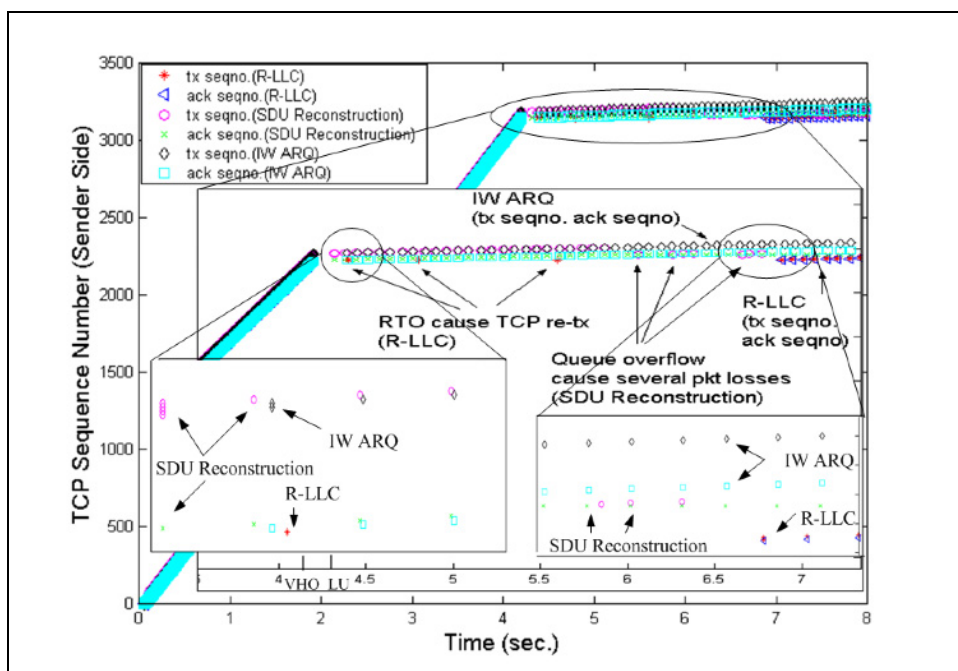


Fig. 10. TCP segment number comparison (wimax->umts, sender side)

A typical problem during handover process from high bandwidth data network WiMAX to relative low bandwidth network UMTS is buffer overflow, which is caused by BDP mismatch between these two networks. The UMTS network is likely to undergo buffer overflow when a TCP congestion window for WiMAX is much larger than the buffer allocation per MS in UMTS RNC.

For SDU Reconstruction scheme, even though TCP congestion window is not larger than buffer size of UMTS, the buffered packet forwarded from WiMAX to UMTS still may have the probability to overflow the UMTS queue, because queue in WiMAX may buffer more packets than the queue size of UMTS due to inflated transmission time. For SDU Reconstruction scheme, in Fig.10, the buffer overflow in UMTS after handover leads to TCP retransmission starting at about 6.0 sec. The corresponding TCP window shrinks, as shown in Fig. 11.

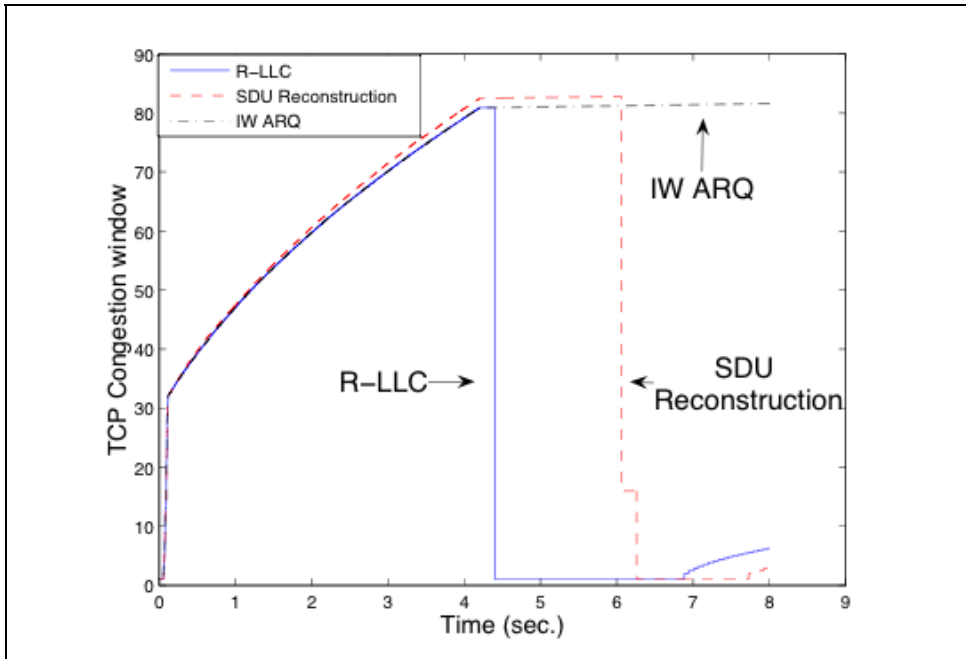


Fig. 11. TCP congestion window (wimax->umts)

For R-LLC scheme, the long status report period leads to TCP RTO and a segment is retransmitted by the TCP sender three times before a status report timer expires. The period of status report timer is set to 2.5 sec. in this scenario.

Whereas for IW ARQ scheme, the support of Link_Up trigger accelerates handover response time, and the adaptive IW ARQ window size effectively eliminates buffer overflow in the target UMTS network. It can be seen that the lossless handover of IW ARQ mechanism has a “side effect”: eliminate the false fast retransmission caused by packet losses or out-of-order packet arrivals during a handover.

4. Inter-RAT Handover between UMTS and WiMAX in Tight Coupling Architecture

4.1 The IW Sublayer in the Tight Coupling Architecture

4.1.1 IW Sublayer Description

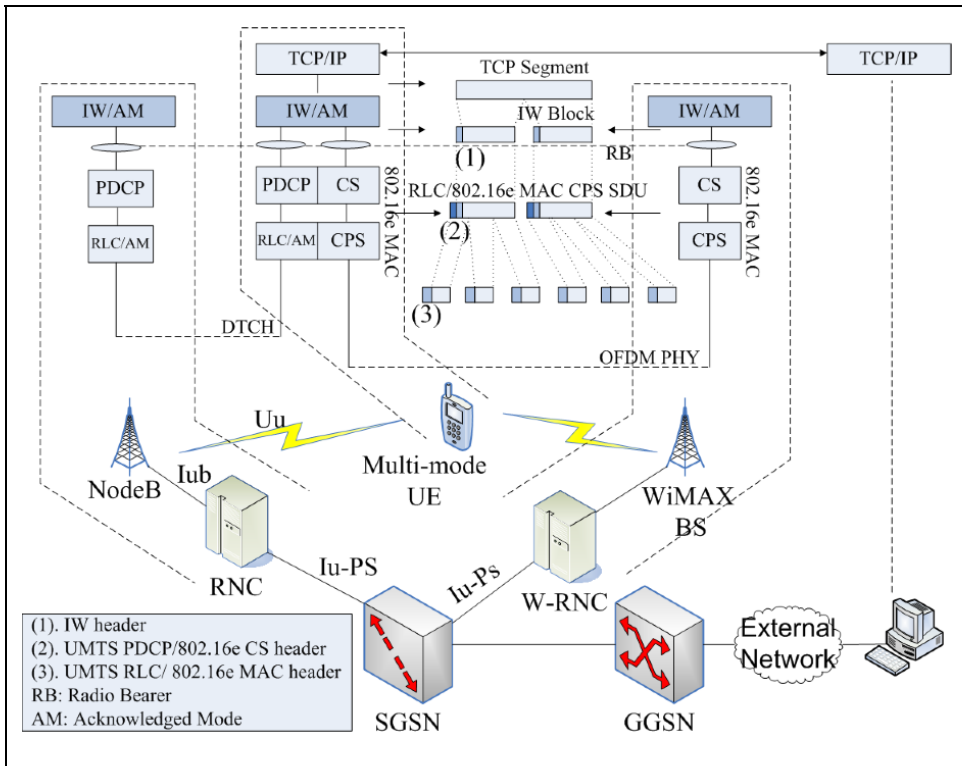


Fig. 12. IW sublayer working mechanism of tight coupling

In the tight coupling scenario, the WiMAX network may emulate a RNC (Radio Network Controller) or a SGSN (Serving GPRS Support Node). We only consider RNC emulation in this chapter. Thus, we introduce a new network component called RNC emulator for WiMAX (W-RNC) in the WiMAX access network, which connects with the UMTS CN (Core Network) at the Iu-PS interface, as shown in Fig. 12. Actually, the W-RNC is an enhanced WiMAX BS with a novel sublayer named IW sublayer, which lies on the top of WiMAX MAC (Medium Access Control) sublayer. The W-RNC with the IW sublayer has the following functions:

- Realize Iu-PS interface.
- Primitive mapping between the IW and the UTRAN network or between the IW and the WiMAX network in case of an inter-RAT handover.

- When an inter-RAT handover takes place, the IW sublayer functions as the LLC sublayer of conventional cellular networks by enabling the SR ARQ (Selective Repeat ARQ) mechanism that includes packet segmentation, re-sequencing, retransmission, and retransmission window size adjustment.
- When a handover takes place, the IW sublayer transfers context to target RNC or W-RNC where the counterpart sublayer locates. In order to provide a seamless inter-RAT handover between UMTS and WiMAX, a peer IW sublayer shall also be realized on the top of the PDCP sublayer on the conventional RNC. While on the MS, the IW sublayer is a common sublayer on the top of the PDCP sublayer of UMTS and the MAC sublayer of WiMAX.

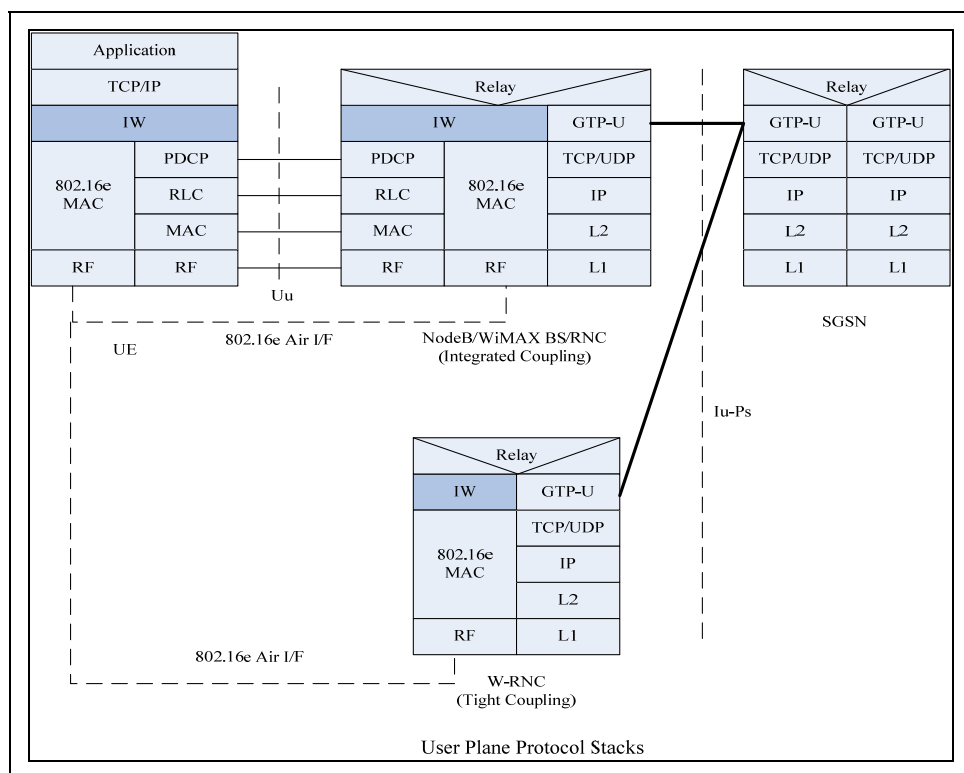


Fig. 13. User plane protocol stacks of tight coupling architecture

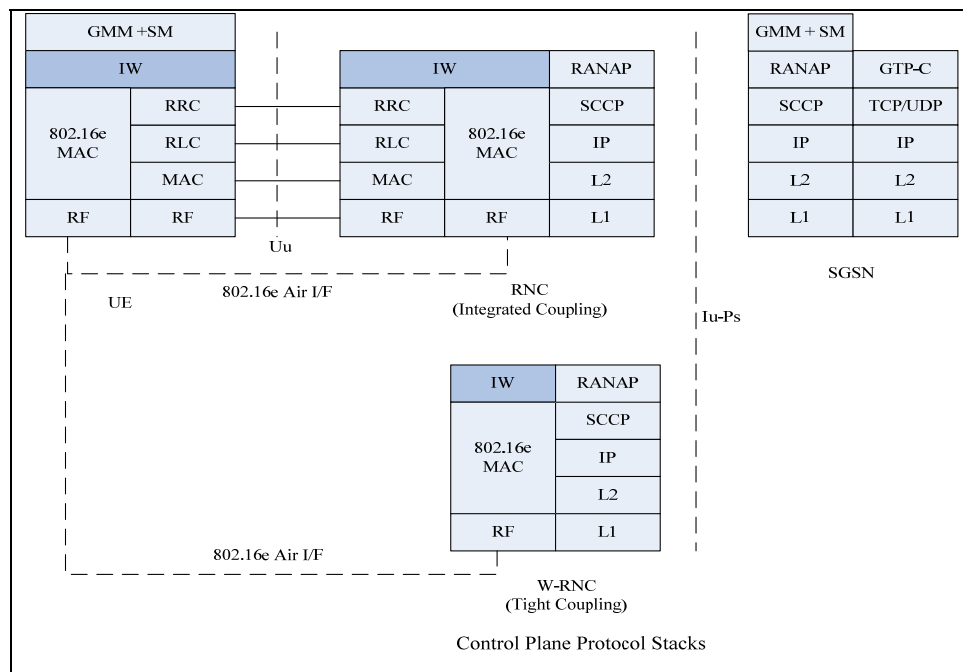


Fig. 14. Control plane protocol stacks of tight coupling architecture

In Fig. 13 and Fig. 14, the user and control planes of the proposed tight coupling architecture are illustrated. W-RNC is assumed to cover the same Routing Area (RA) like the RNC. The IW sublayer on the W-RNC communicates with its counterpart entity on the RNC in order to execute inter-RAT handover to/from its control area. The main contents of the communication between them are as follows:

- GTP-U sequence numbers as well as GTP packets that need to be forwarded by the W-RNC for PDP contexts requiring delivery order.
- IW ARQ parameters, such as windows size, queue length, retransmission timer period, and retransmission count.
- The IW blocks stored in local retransmission buffer.

There are two reasons why add IW ARQ mechanism to W-RNC in addition to SRNS (Serving RNS) context transfer of conventional RNC:

- When an inter-RAT handover takes place, there may exist packet sequence number asynchronization between the source RNC and the target WiMAX BS. It is necessarily that there exists a common context transfer mechanism for these two systems to assure a lossless handover.
- The second reason is that the WiMAX supports cell reselection initiated by MS for active traffics (like dedicated mode in UMTS), which is not the case in UMTS. Hence, the packets that are lost during the cell reselection from WiMAX to UMTS, cannot be retransmitted by the target network.

4.1.2 Signaling and Primitives

4.1.2.1 Handover from UMTS to WiMAX

This sub-clause describes the inter-RAT handover signaling procedures and primitives among IW, PDCP, RRC (Radio Resource Control) and WiMAX MAC in the tight coupling architecture. In the following figures, IW/RNC means the function combination of IW sublayer and RNC, so are the IW/W-RNC and MAC/W-RNC. Some newly added cross-layer primitives are augmented to the conventional inter-RAT handover signaling procedures of 3GPP (3GPP TS 43.129; 3GPP TR 25.931). We suggest the future WiMAX and UMTS standards should support these primitives and parameters for the smooth and seamless inter-RAT handover. The handover preparation period is similar to that of integrated coupling architecture and is omitted in this sub-clause.

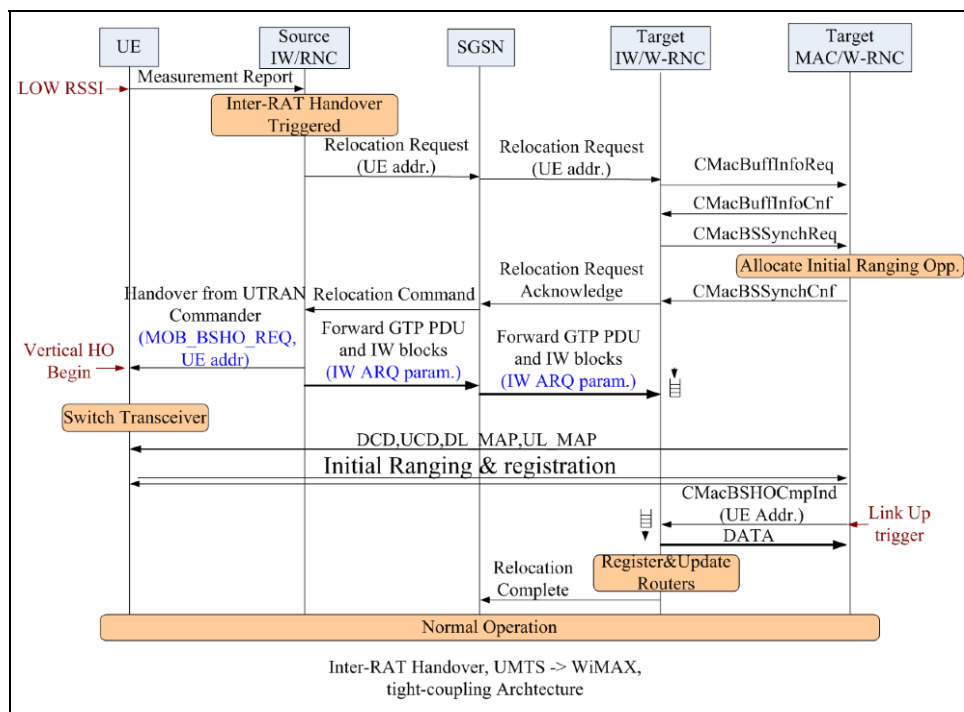


Fig. 15. Signaling procedure of the handover from UMTS to WiMAX

Fig. 15 describes the inter-RAT handover from UMTS to WiMAX and shows the exchanged messages.

- 1) Based on measurement reports and knowledge of the RAN topology, the RNC, more precisely source RRC decides to initiate an inter-RAT PS handover.
- 2) The source RNC sends a Relocation Request (contains target WiMAX cell id) message to the SGSN. The SGSN forwards Relocation Request message to target W-RNC.
- 3) Then the IW of target W-RNC sends the CMacBuffInfoReq primitive to the WiMAX MAC to request the buffer characteristics. The WiMAX MAC returns the

- CMacBuffInfoCnf primitive to inform the IW of the buffer size in its MAC sublayer. According to this information, the target IW adjusts its retransmission window size. It should be mentioned at this point that current WiMAX MAC does not support this interface, so the IW may adjust its retransmission window size to a default value.
- 4) At this stage, the target IW sends the CMacBSSynchReq primitive to the WiMAX MAC to negotiate the location of the dedicated initial ranging transmission opportunity for the MS. This information is returned by primitive CMacBSSynchCnf.
 - 5) The target W-RNC sends the Relocation Request Acknowledge message to SGSN, and the SGSN continues the handover by sending a Relocation Command to source RNC (including Transparent Container (MOB_BSHO-REQ)).
 - 6) Upon receipt of Relocation Command message, the IW of source RNC will forward IW context to target IW of W-RNC. The IW context consists of IW ARQ parameters, received IW ACK and remaining IW blocks that have not been transmitted successfully.
 - 7) The RRC of the source RNC sends the Handover from UTRAN Command message to the MS.
 - 8) The MS performs hard handover and normal network entry procedure.
 - 9) After the provisioned service flow is activated, the target WiMAX MAC sends CMacBSHOCmpInd primitive as a Link_UP (LU) trigger to the IW sublayer. On this trigger, the IW starts data packet forwarding.

4.1.2.2 Handover from WiMAX to UMTS

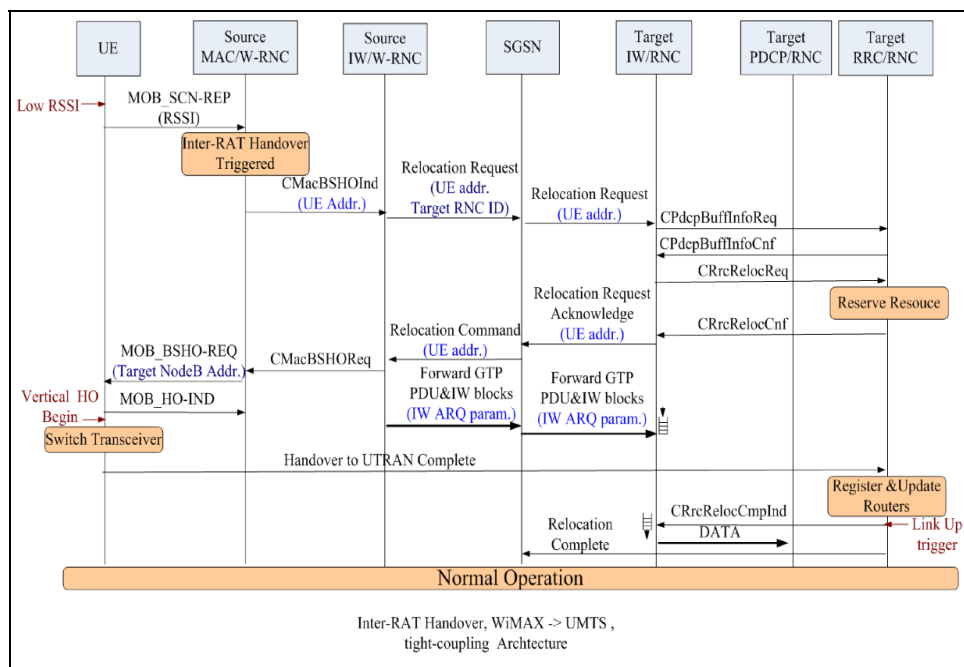


Fig. 16. Signaling procedure of the handover from WiMAX to UMTS

The inter-RAT handover from WiMAX to UMTS is described in Fig. 16.

- 1) After the scanning interval, the MS sends scanning report to WiMAX serving BS by message MON_SCN-REP that contains physical information such as mean RSSI.
- 2) The source WiMAX MAC sends CMacBSHOInd primitive to inform the IW sublayer of handover and target cell id. Then, the source W-RNC sends a Relocation Request (contains target cell id) message to the SGSN. The SGSN forwards Relocation Request message to target RNC.
- 3) The IW of target RNC sends CPdcpBuffInfoReq primitive to the RRC sublayer to request the buffer characteristics of the PDCP sublayer, and RRC returns the CPdcpBuffInfoCnf primitive to inform the IW of buffer size. According to this information, the IW adjusts its retransmission window size.
- 4) The target IW sends a CRrcRelocReq primitive to the target RRC to apply for resource allocation. The result is returned in CRrcRelocCnf primitive by the target RRC.
- 5) The target RNC sends the Relocation Request Acknowledge message (contains target RNC to source W-RNC transparent Container) to SGSN. The SGSN continues the handover by sending a Relocation Command to source W-RNC.
- 6) On receipt of Relocation Command message, the IW of source W-RNC will forward IW context to the IW of target RNC. The IW context consists of IW ARQ parameters, received IW ACK, and remaining IW blocks that have not been transmitted successfully.
- 7) The source IW sends CMacBSHOREq primitive to inform MAC that the target network is ready.
- 8) The MS performs handover to one of BSs specified in MOB_BSHO-REQ and responds with a MOB_HO-IND message.
- 9) MS performs normal UMTS hard handover.
- 10) After the MS successfully finishes UMTS radio link setup, the target RRC shall send the CRrcRelocCmpInd primitive to the IW, and the IW starts data packet forwarding.

Note that primitive CMacBSHOCmpInd and primitive CRrcRelocCmpInd are defined as the Link_Up (LU) triggers for handover from UMTS to WiMAX, and for handover from WiMAX to UMTS respectively.

4.1.3 Buffering-and-Forwarding (B&F) in Tight Coupling Architecture

We have mentioned that: in FMIPv6 protocol (R. Koodli, 2005), in order to make a handover lossless, previous access router (PAR) will forward buffered packets destined for the MS during the handover period to the new access router (NAR) through an established tunnel, after receiving the new care-of-address (NCoA) of the MS from the NAR. One essence of this IP layer handover solution is the utilization of buffering-and-forwarding (B&F) context transfer scheme. In our IW sublayer solution, the B&F scheme is also applied to forward GTP-U packets, IW blocks and so on from source IW sublayer to target IW sublayer. It becomes meaningful and interesting to compare IW sublayer handover solution with FMIPv6 in tight coupling architecture. For fairly comparing the inter-RAT handover performance of IW sublayer solution with that of FMIPv6, we also only realize the B&F scheme in the IW sublayer in our inter-RAT handover scenario, where the RNC or W-RNC takes the responsibility of buffering and forwarding. This kind of Layer 2 realization considers the fact that: the conventional IP sublayer terminates on the GGSN in the UMTS network, which suffers from longer transmission delay between the MS and the GGSN at IP

layer. It must be stressed that this Layer 2 realization of B&F has better performance than IP layer realization like FMIPv6 thanks to the ability to directly operate Layer 2 data packets stored in one RAT. In the simulation sub-clause, we will inspect these two kinds of context transfer schemes- IW ARQ and B&F in the tight coupling architecture. If the handover performance of IW ARQ is better than that of B&F at Layer 2, we can say that the IW sublayer solution is more suitable for inter-RAT handover than FMIPv6.

4.2 Simulation Results in the Tight Coupling Scenario

4.2.1 Handover from UMTS to WiMAX

For the simulation of inter-RAT handover from UMTS network to the WiMAX, an FTP session starts at 0.4 sec., and the MS starts to perform handover at about 4 sec. after it enters into the coverage region of WiMAX. The handover type is hard handover. At about 4.035 sec, the WiMAX network entry procedure is finished and the IW sublayer on the RNC receives a Link_Up trigger. Fig. 17 shows the packet flows of two kinds of context transfer schemes: buffering-and-forwarding (B&F) and IW ARQ.

During the handover period, there are no new TCP segment arrivals and consequently no segments are forwarded through the tunnel between RNC and W-RNC for both context transfer schemes (see Fig.18). One can see from Fig. 18 that, in B&F scheme, the TCP sender retransmits the last unacknowledged segments on the timeout of TCP retransmission timer (RTO) at about 5.7 sec. During this period, the congestion window shrinks to one, and throughput reduces significantly.

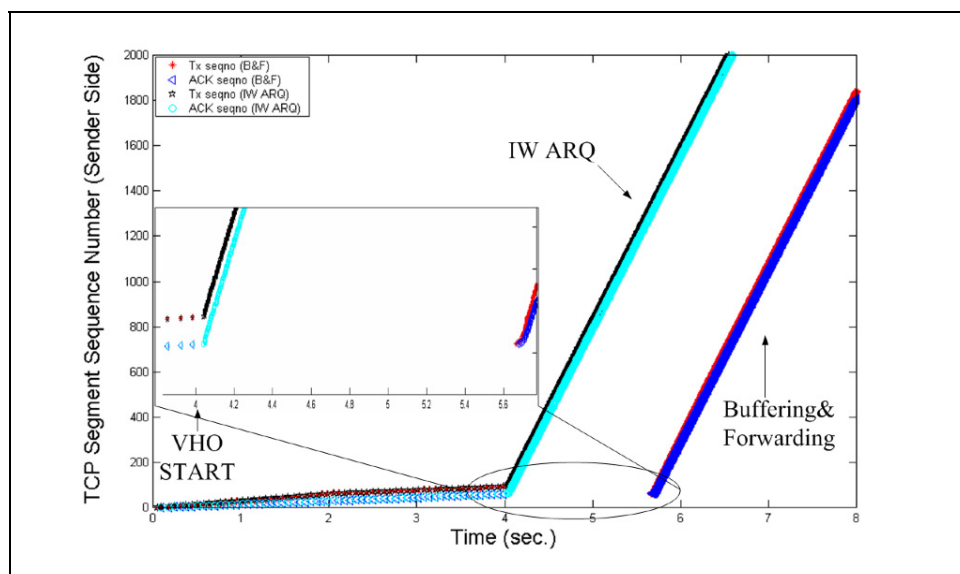


Fig. 17. TCP segment number comparison (umts->wimax, sender side)

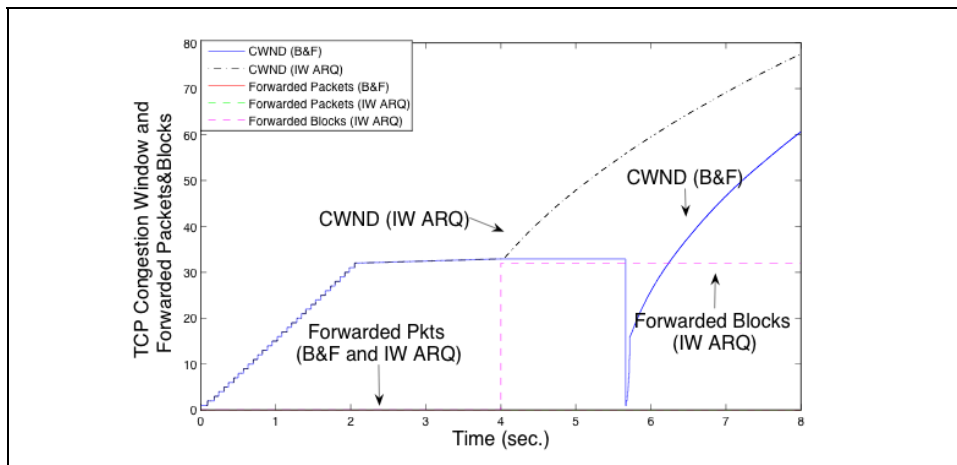


Fig. 18. TCP congestion window (umts->wimax)

The IW ARQ scheme adjusts its retransmission window according to the target network's queue size, and sends the IW blocks that are forwarded from the source IW on receipt of Link_Up trigger. After the handover, there will be no packet losses and TCP congestion window does not shrink thanks to the retransmission mechanism.

4.2.2 Handover from WiMAX to UMTS

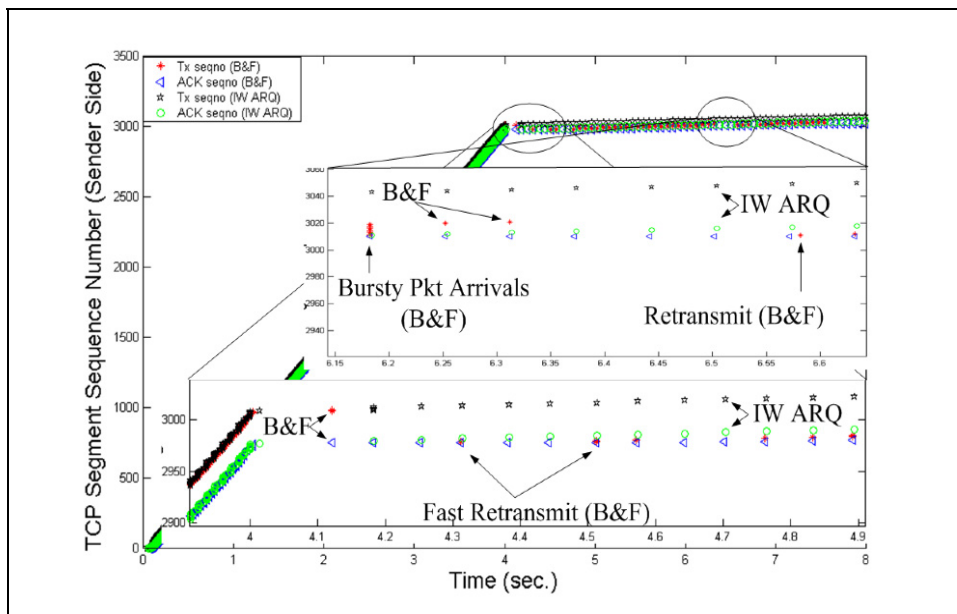


Fig. 19. TCP segment number comparison (wimax->umts, sender side)

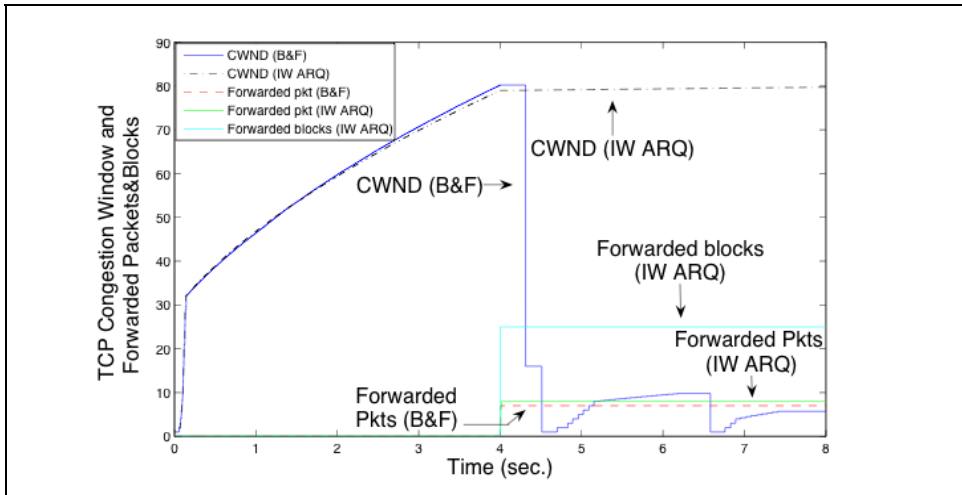


Fig. 20. TCP congestion window (wimax->umts)

When a handover from WiMAX to UMTS happens, there exist some TCP segments and IW blocks, which are forwarded from W-RNC to RNC through the tunnel for two schemes, as shown in Fig.20. For B&F scheme, the arrivals of tunneled segments (about 8 segments) trigger the fast retransmissions twice at time 4.31 sec. and 4.51 sec. for the lost segments during the handover (see Fig. 19), and then congestion window size reduces significantly. From then on, the TCP sender retransmits the segments numbering from the first lost segment to the tunneled segments. The receiver will acknowledge again those segments that have been tunneled before at about 6.18 sec, which herein trigger the bursty segment arrivals. Furthermore, those retransmitted segments that have been tunneled during handover procedure delay the ACK feedback of new segments, and in consequence lead to a retransmission caused by RTO at 6.58 sec.. We can see that, the B&F scheme degrades the handover performance instead of improving it for TCP traffics due to the lack of a mechanism that recovers the lost packets.

For IW ARQ scheme, there is no packet loss during the handover. The support of Link_Up trigger accelerates handover response time, and the adaptive IW ARQ window size effectively eliminates buffer overflow in the target UMTS network. The only price for this lossless handover procedure is that the IW sender may retransmit a couple of IW blocks that possibly have been received by IW receiver but the corresponding ACKs are lost in the air during the handover period.

5. Conclusion

This chapter focuses on introduction of proposed inter-RAT handover solution for interworking UMTS with WiMAX. First, the 3GPP cell reselection and handover mechanism are outlined in the first section. This section gives us main guideline for designing a new mechanism to deal with typical problems of inter-RAT handover between UMTS and WiMAX. Then, a novel Layer 2 inter-RAT handover scheme on basis of the integrated coupling and tight coupling architectures for the seamless roaming between UMTS and

WiMAX networks are elaborated. For instance, in integrated coupling architecture, a new common sublayer named IW sublayer that lies on the RNC and MS is added on the top of PDCP (UMTS) and MAC (WiMAX) sublayer. At this novel sublayer, a new retransmission scheme called IW ARQ is also proposed to eliminate packet losses during handover procedure and accelerate handover procedure. Compared with other context transfer mechanisms, such as R-LLC, SDU Reconstruction and buffering-and-forwarding, IW ARQ can achieve lossless and prompt handover procedure for TCP traffics thanks to the introduction of SR ARQ mechanism. The better handover performance is validated by the simulation results carried out on NS2 emulator. In addition, this novel IW sublayer scheme also can eliminate the false fast retransmission due to packet loss or out-of-order packet arrivals during a handover period. It also provides a suitable framework to solve the TCP problem of BDP mismatch, premature RTO and so on.

6. References

- 3GPP TS 04.18. Technical Specification Group GSM/EDGE Radio Access Network; mobile radio interface layer 3 specification; radio resource control protocol
- 3GPP TS 05.08. Technical Specification Group GSM/EDGE Radio Access Network; Radio subsystem link control (release 99)
- 3GPP TS 23.060. Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS); Service description, stage 2 (Release 7), v7.4.0
- 3GPP TS 23.122, "Technical Specification Group Core Network and Terminals; Non-Access-Stratum (NAS) functions related to Mobile Station (MS) in idle mode", (Release 7), V7.2.0
- 3GPP TS 23.234. 3GPP system to Wireless Local Area Network (WLAN), interworking; System description (Release 6), V6.5.0
- 3GPP TS 25.304. User Equipment (UE) procedures in Idle Mode and Procedures for Cell Reselection in Connected Mode
- 3GPP TS 25.323. Technical Specification Group Radio Access Network; Packet Data Convergence Protocol (PDCP) specification, (Release 7), V7.5.0
- 3GPP TS 25.331. Technical Specification Group Radio Access Network; Radio Resource Control (RRC); Protocol Specification (Release 7), v7.5
- 3GPP TR 25.922. Technical Specification Group Radio Access Network; Radio resource management strategies (Release 7), V7.10
- 3GPP TR 25.931. Technical Specification Group RAN; UTRAN functions, examples on signaling procedures, (Release 7), V7.4.0
- 3GPP TS 43.022, "Technical Specification Group GSM/EDGE; Radio Access Network; Functions related to Mobile Station (MS) in idle mode and group receive mode", (Release 7), V7.2.0
- 3GPP TS 43.129. Technical Specification Group GSM/EDGE" Radio Access Network; Packet-switched handover for GERAN A/Gb mode; Stage 2 (Release 7), V7.2.0
- 3GPP TS 44.060. Technical Specification Group GSM/EDGE Radio Access Network; General Packet Radio Service (GPRS), Radio Link Control/Medium Access Control (RLC/MAC) protocol (Release 7), V7.9.0
- 3GPP TS 45.008. Technical Specification Group GSM/EDGE; Radio Access Network; Radio subsystem link control (Release 7), V7.8.0

- C. Johnson, R. Cuny & N. Wimolpitayarat. (2005). Inter-System Handover for Packet Switched Services", 2005 6th IEE International Conference on 3G and Beyond, pp: 1-5, 7-9 Nov. 2005
- D. Johnson, C. Perkins, & J. Arkko. (2004). IP Mobility Support in IPv6, <http://www.ietf.org/rfc/rfc3775.txt>, IETF, June 2004
- E. Seurre, P. Savelli & P.J. Pietri. (2003). *GPRS for Mobile Internet*, Artech House Ltd, 2003
- G. Lanpropoulos, N. Passas, L. Merakos & A. Kaloxylas. (2005). Handover Management Architectures in Integrated WLAN/Cellular Networks. *IEEE Communication Survey & Tutorials*. Vol.7, No.4, pp: 30-44, Fourth Quarter 2005
- H. Inaura, G. Montenegro, R. Ludwig, A. Gurtov & F. Khafizov. (2003). TCP over Second (2.5) and Third (3G) Generation Wireless Networks, IETF, RFC 3481
- H. Rutagemwa, S. Park, X.M. Shen & J.W. Mark. (2007). Robust Cross-layer Design of Wireless Profiled TCP Mobile Receiver for Vertical Handover, *IEEE Tans. On Vehicular Technology*, Vol.56, No. 6, pp: 3899-3911, Nov. 2007
<Http://www.isi.edu/nsnam/ns>
- IEEE 802.16e (2005). IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands, 2005
- IEEE 802.21 (2006). Standard and Metropolitan Area Networks: Media Independent Handover Services, Draft P802.21/D00.05, January 2006
- J.P Romero, O. Sallent, R. Agusti & A.D.G. Miguel. (2005). *Radio Resource Management Strategies in UMTS*, John Wiley & Sons, Ltd, 2005
- J. Sachs, B. S. Khurana & P. Mahonen. (2006). Evaluation of Handover Performance for TCP Traffic Based on Generic Link Layer Context Transfer, *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2006 (PIMRC'06)*, pp: 1-5, Sept. 2006
- M. Afif, P. Martins, S. Tabbane & P. Godlewski. (2006). SCTP Extension for EGPRS/WLAN Handover Data, *31st IEEE Conference on Local Computer Networks*, pp: 746-750, Nov. 2006
- N. Dailly, P. Martins & P. Godlewski. (2006). Performance evaluation of L2 handover Mechanisms for Inter-Radio Access Networks, *IEEE Vehicular Technology Conference, 2006 (VTC2006-Spring)*, pp: 491-495, May 7-10, 2006
- N. Vulic, I. Niemegeers & S.H de Groot. (2004). Architectural Options for the WLAN Integration at the UMTS Radio Access level, *IEEE Vehicular Technology Conference, 2004 (VTC 2004-Spring)*, pp: 3009-3013, May 17-19, 2004
- R. Koodli. (2005). Fast Handovers for Mobile IPv6, <http://www.ietf.org/rfc/rfc4068.txt>, IETF, July 2005
- S.L. Tsao & C.C. Lin. (2002). Design and Evaluation of UMTS/WLAN interworking Strategies, *IEEE Vehicular Technology Conference, 2002 (VTC2002-Fall)*, pp: 777-781, 2002

MHD-CAR: A Distributed Cross-Layer Solution for Augmenting Seamless Mobility Management Protocols

Faqir Zarrar Yousaf, Christian Müller and Christian Wietfeld
*Communication Networks Institute,
 Dortmund University of Technology (TU Dortmund),
 Germany*

1. Introduction

The Next Generation Network (NGN) architecture is evolving into a highly heterogeneous infrastructure composed of a variety of Wireless Access Technologies (WAT). In such a technologically diverse network; one of the key challenges would be to ensure ubiquitous communication services to mobile entities, with varying mobility patterns and speed, irrespective of their location and/or the underlying WAT. This calls for devising efficient mobility management solutions that would provide *location management* and *handover management* services to mobile entities. The location management service is responsible for keeping track of the location of the mobile entities whereas the handover management service enables the mobile entity to change its point of connection in the Internet. The essential mandate of any efficient mobility management solution is to provide effective and fast location monitoring and updating services while enabling seamless inter-WAT handover. The notion of seamless handover implies handovers with minimum latency and packet losses.

Providing seamless handover is an imposing challenge because the location update takes place after the successful execution of the handover. The handover process is executed at both the data link layer (L2) and at the network layer (L3) based on prescribed rules at these respective layers. The *L2 handover* process enables the mobile node (MN) to switch its link connectivity from its serving access point (AP) or base station to the new one. After successfully establishing link connectivity, the MN will then need to perform *L3 handover* process to make the MN IP capable on the new link. Till the completion of the handover, the MN is practically disconnected from the network resulting in loss of data. The amount of data lost depends on the handover latency, which in turn is a sum of delay incurred by handover procedure prescribed at L2 and L3 respectively. This implies that the latency of the L3 handover process is directly dependent on the latency of the L2 handover process and also on the timing of the provisioning of L2 triggers that will initiate the L3 handover process.

Thus to develop seamless handover methodology, it is important to take into account the effect of L2 handover process on the L3 specified handover operations. This calls for devising *cross-layer mobility management solution* that will enable inter-layer communication

(i.e., between L2 and L3) of critical information that will enhance the performance of the overall handover process.

In this chapter we present the details of one such cross-layer solution called *Multi-Hop Discovery of Candidate Access Router (MHD-CAR)* that not only optimises the standard Candidate Access Router Discovery (CARD) protocol (Liebsch et al., 2005) but it also offers inherent cross-layer capabilities that can potentially contribute towards achieving low latency handovers and hence enhance the operational efficiency of seamless mobility management protocols in general.

The details of this novel solution will be presented in the context of Fast Mobile IPv6 protocol. A portion of the work presented in this chapter is based on our previous efforts that has been recorded and published in (Yousaf et al., 2008(b)).

2. Technical Background

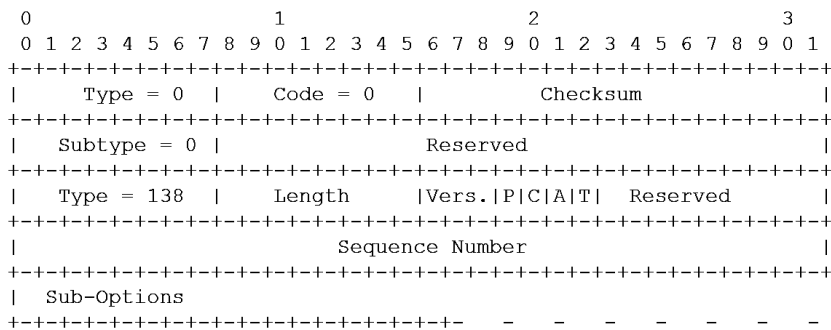
To provide mobility management to MNs, IETF has specified Mobile IPv6 (MIPv6) protocol at L3 (Perkins et al., 2004). However, it is an established fact that the handover performance of MIPv6 is not seamless, in that it incurs a high handover latency and packet delay. This is because the MIPv6 operation is based on a *break-before-make* operation in which the MN will initiate the MIPv6 protocol after it has disconnected from its serving AP as it moves out of its coverage range. After the disconnection, the MN will search and connect to the appropriate new in-range wireless AP. After establishing link connectivity (or performing L2 handover), the MIPv6 handover process will be initiated which is based on a sequential execution of a series of sub-processes; namely Care-of-Address configuration, Duplicate Address Detection (DAD) test, Home Registration, Return Routability Test and Correspondent Registration. Each of the sub-process will incur a finite amount of delay (DAD test alone incurs a delay of 1 sec) contributing thereby to the total handover latency. During the execution of the MIPv6 protocol, the MN remains disconnected from the Internet and is unable to transmit or receive packets resulting in data losses. Hence, the MIPv6 handover latency, which is in excess of 1.5 seconds (Yousaf et al., 2008(c)), is unsuitable for delay sensitive and throughput sensitive applications.

To provide seamless handover services, IETF has specified *Fast Mobile IPv6 (FMIPv6)* protocol in RFC 5268 (Koodli, 2008) which extends the standard MIPv6 protocol. The main operational concept of FMIPv6 is based on the ability of the MN to detect and negotiate a handover with the New Access Router (NAR) in advance while the MN is still connected to its Present Access Router (PAR). During handover negotiation with NAR, a bi-directional tunnel is established between the PAR and NAR so that packets arriving at the PAR (and destined towards the MN) are tunnelled towards NAR where they will get buffered. These buffered packets will get forwarded to the MN soon after it establishes link connectivity with the AP associated with the NAR and becomes IP capable on the NAR's link. In other words, FMIPv6 is based on a *make-before-break* concept which not only reduces handover latency but also the packet loss by virtue of the tunnelling and buffering of packets.

However, the key to the success for the FMIPv6 handover operation is the ability of the MN to detect and identify the presence of in-range Candidate Access Routers (CARs) and then select an appropriate NAR from amongst the identified CARs. The FMIPv6 protocol specification only specifies the handover operation by assuming that the MN has already identified and selected a suitable NAR and hence does not provide any specific mechanism

The summary of the CARD protocol and its interaction with the protocol operation of FMIPv6 is given in the following sub-section.

The CARD protocol is a standard IETF solution the operational details of which are specified in RFC 4066. The protocol provides a generic mechanism that allows a MN to acquire the necessary and relevant information about the ARs that are potential candidates for the MN's next handover. It is based on the exchange of a series of *request* and *reply* messages between the MN and its serving AR and also amongst ARs as well. The messages exchanged between the MN and its serving AR (i.e., PAR) is designated as *MN-AR CARD Request* and *MN-AR CARD Reply* message, whereas those exchanged between the ARs are termed as *AR-AR CARD Request* and *AR-AR CARD Reply* message. These messages are transported as options inside the ICMPv6 message. The format of the CARD Request and CARD Reply message is illustrated in Figure 1 and 2 respectively.



0										1										2										3										
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1									
Type = 0										Code = 0										Checksum																				
Subtype = 0										Reserved																														
Type = 139										Length										Vers. P U L										Reserved										
Sequence Number																																								
Sub-Options																																								

Fig. 2. Format of the CARD Reply Message Carried as an Option in ICMPv6 Message

The CARD protocol is designed to perform the following two functions namely;

1. Reverse Address Translation (RAT)
2. Discovery of CAR Capabilities (DCC)

The *RAT function* enables a MN to map the L2 identifiers (L2-IDs) (e.g., a MAC address for 802.11 networks) of one or more in-range APs to the IP address (L3-IDs) of the associated CAR connected to it. The L2-IDs are typically discovered during the scan operation initiated by the MN as a reaction to the link condition going below a certain specified threshold value of SNR or RSSI.

The *DCC function* on the other hand, allows the MN to acquire the capability's information of the discovered CARs. The notion of *capabilities* implies of the various QoS aspects offered by a CAR that would then be used as input to the MN's target AR (TAR) selection algorithm to make optimal handover decisions. The DCC function will prevent the MN to make inaccurate network selections and hence connections to the wrong one in case of many available CARs.

Central to the CARD operation is a L2-L3 address mapping Table called a *CAR Table*, which is managed and maintained inside each AR. The information content of the *CAR Table* is used to resolve the L2-IDs of a *Candidate AP (CAP)* to the IP address and capabilities of the associated CAR.

RFC 4066 suggests the use of a central entity called a *CARD Server* as one of the strategy to populate the *CAR Table*. During boot up time all ARs within the administrative domain of the *CARD Server* will register their IP addresses and the L2-IDs of the associated CAPs with the *CARD Server*. The *CARD Server* will then be queried during the RAT process.

The functionality of the CARD protocol extends many benefits which are outlined in (Trossen et al., 2002). For example, the information that the MN acquires due to the CARD operation will enable it to select and connect to an appropriate network that would provide the necessary service to the MN based on its application requirements. This can be beneficial to multi-homed MNs as it may also enable the MN to select the least-cost network and enable inter-WAT handoff. It can also be used to perform load balancing by enabling the MN to switch its connection from a heavily loaded AR to a comparatively lightly loaded one.

2.1.1 Operational Summary

The CARD protocol operation, depicted in Figure 3, is typically initiated upon receiving appropriate L2-Triggers or when the MN discovers L2-ID(s) of in-range CAP(s) as part of a scan operation. The MN will be able to discover the L2-ID(s) when it is in the overlapping coverage region of two or more neighbouring AP(s). The MN will then request its current AR to resolve the identity (*RAT function*) and discover the capabilities (*DCC function*) of the CAR(s) associated with the CAP(s) whose L2 ID(s) is carried in the *MN-AR CARD Request* message. The current AR will check in its local *CAR Table* for a corresponding entry against the L2 ID(s). If the current AR fails to resolve the identity of the associated CAR(s), it will send an *AR-AR CARD Request* message (depicted as *AR-Server Req* in Figure 3) to a central CARD Server, to which all the ARs would have registered their identity information (IP address, address prefix or prefix length) during boot up time. The CARD Server will return the IP(v6) address(es) of the CAR(s) via an *AR-AR CARD Reply* message (depicted as *AR-Server Rep* in Figure 3), and the current AR will update its local *CAR Table* with the resolved CAR information. The initial idea of performing RAT function using a centralised server

was first presented in (Funato et al., 2002) which has been adopted by RFC 4066 as one of the probable method.

The current AR, depending on the status of the C-flag (capabilities request flag) in the MN-AR CARD *Request* message, will then directly contact the resolved CAR(s) and perform capabilities discovery via AR-AR CARD *Request/Reply* message pair and then send the resolved identities and capabilities of the CAR(s) to the MN via a MN-AR CARD *Reply* message. It may be mentioned that the identity and capabilities information are carried in specified message sub-options and containers the details and format of which is given in (Liebsch et al., 2005). Based on the capabilities information and preset criteria, the MN will select an appropriate TAR to which it will perform a handover with. This process will ensue every time the handover is imminent.

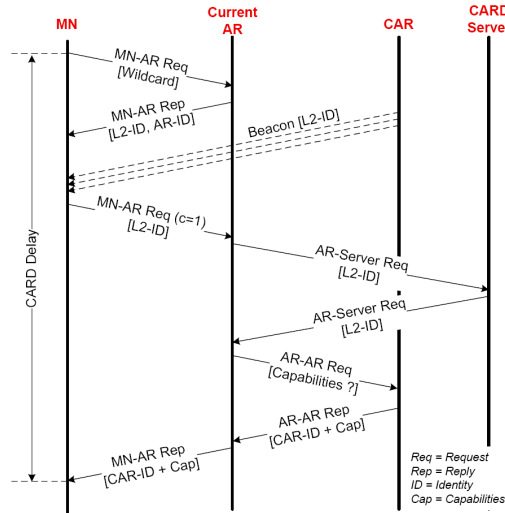


Fig. 3. The CARD Protocol Operation

Upon connecting to the target AR the MN will send a *wildcard* MN-AR CARD *Request* to obtain the information of its CAR Table in order to improve the prospects of the next CAR discovery. It is observed that the maintenance and management of local CAR Tables is critical to the effectiveness of CARD protocol in terms of assisting seamless and fast handover by way of quick address resolution and capabilities discovery.

2.2 FMIPv6 Operation with CARD Protocol

Figure 4 illustrates the FMIPv6 protocol operation in conjunction with the CARD protocol. The only difference is that the MN-AR CARD *Request* and MN-AR CARD *Reply* message are piggybacked on the FMIPv6 protocol specified *Router Solicitation for Proxy* (RtSolPr) and *Proxy Router Advertisement* (PrRtAdv) messages respectively (Liebsch et al., 2005). The RtSolPr and PrRtrAdv are ICMPv6 type messages the format of which is specified in (Koodli, 2008).

The MN will start scanning for in-range APs in response to deteriorating link conditions. The MN will then send the L2-ID(s) of the scanned AP(s) to the upper layer (i.e., L3) in the form of L2

trigger which will initiate the CARD operation as described in Section 2.1.1. The PAR after acquiring the identities and capabilities of the available CAR(s) against the L2-ID(s) provided by the MN in the *RtSolPr* message will send this information as [L2-ID, AR-Info]¹ tuple appended to the *PrRtAdv* message. The MN will then select a suitable NAR from amongst the discovered CAR(s) based on some TAR selection criteria which is beyond the scope of this chapter.

Based on the identity information of NAR, the MN will auto-configure (Thomson et al., 2007) a prospective New Care of Address (pNCoA) and will send this in a *Fast Binding Update Message (FBU)* message to the PAR indicating the NAR to which the MN wishes to handover its connection. The PAR will then notify the NAR of the pCoA and request for a handover by sending a *Handover Initiate (HI)* message. The NAR in response will acknowledge the handover request by transmitting a *Handover Acknowledge (HACK)* message towards the PAR. The PAR upon receiving the *HACK* will immediately set up a forwarding tunnel with the NAR (referred to as PAR-NAR tunnel) and will start tunnelling subsequent packets destined for the MN towards NAR where they will get buffered. In the meantime the PAR, upon processing the *HACK*, will inform the MN of the NAR's decision via a *Fast Binding Acknowledgement (FBACK)* message. It should be noted that all this operation is executed while the MN is still connected to the PAR.

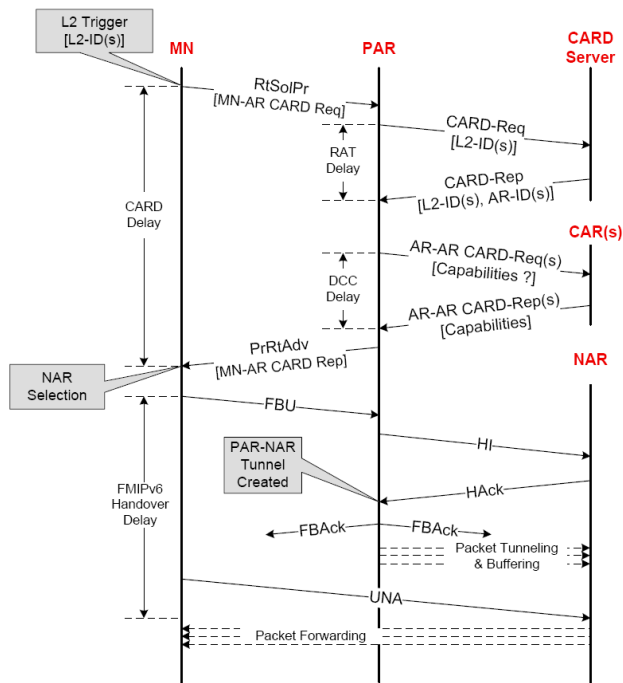


Fig. 4. The FMIPv6 Predictive Handover Operation Utilizing the CARD Protocol for NAR Discovery

¹ The AR-Info corresponds to the identity and capabilities of an AR.

After the MN moves out of the communication range of PAR, it will establish link connectivity with the New Access Point (NAP) and will announce its presence to the NAR by sending an *Unsolicited Neighbour Advertisement (UNA)* message (Narten et al., 2007) to the NAR. The NAR will immediately forward all the buffered packets, and the subsequent packets arriving via the PAR-NAR tunnel, towards the MN. The MN will then inform the HA and the CN(s) of its new location as per the MIPv6 protocol rules (Perkins et al., 2004), after the completion of which the handover is said to be complete.

It should be noted that the FMIPv6 protocol specifies two handover modes namely *Predictive Handover Mode* and *Reactive Handover Mode*, depending on whether the MN receives the *FBack* on the PAR's link or not. Of the two modes, the Predictive Handover Mode is more seamless and hence is the preferred and default mode.

Besides FMIPv6, the CARD protocol functionality can also be used by the MIPv6 protocol to facilitate the MN's decision to select the appropriate AR for handover that would best serve its QoS requirements. For further details see (Liebsch et al., 2005).

3. Problem Statement

As described previously, central to the success of the FMIPv6 protocol is the ability of the MN to discover and select NAR using the CARD protocol facilities. However the process to discover and select NAR depends on two discovery processes namely:

1. CAP discovery, and
2. CAR discovery.

Both of these processes will influence the overall discovery process in consideration of the inherent process limitations described below.

Candidate Access Point (CAP) Discovery Delay:

The CAP discovery process is undertaken by the technique specified for the underlying WAT, whereas the CAR discovery process is carried out by the CARD protocol after the L2 provides it with the identities of the discovered CAP(s) using specific L2 constructs such as triggers, events, hints etc.

Since the CARD protocol, and hence the FMIPv6 protocol, rely on the *timely* and *accurate* provisioning of these L2 constructs, therefore the performance inadequacies of the L2 specified operations will have adverse consequences on the performance of the CARD and thus the FMIPv6 protocol. For example, in reference to the IEEE 802.11 WLAN, the MN is required to perform scan operations (active or passive) to determine the presence of in-range CAP(s). However, as pointed out in (Mishra et al., 2003), the scan operation accounts for almost 90% of the L2 handover delay and can be approximately up to 400 ms. It should also be noted that during the scan operation, the MN is unable to transmit or receive packets resulting in data loss. Besides packet loss, the CAP discovery delay will translate into the delayed provisioning of L2 triggers which in turn will delay the initiation of the CARD protocol.

Thus in order to reduce the packet loss and ensure the timely initiation of the CARD protocol, the duration of the scan operation must be reduced.

Candidate Access Router (CAR) Discovery Delay:

The CAR discovery process is undertaken by the L3 specified CARD protocol after receiving L2 triggers. As described, the CARD protocol performs RAT and DCC function by the exchange of *CARD Request/Reply* messages. This incurs a high signalling cost especially over the error prone radio links. Besides signalling cost, each function incurs a finite amount of delay which is also influenced by the location of the CARD Server and its topological distance from the PAR. Besides influencing the delay, the CARD Server is a central network entity that must be managed and maintained thereby increasing the cost of network management. The CARD Server also introduces a potential single-point-of-failure, the failure of which will result in a failed handover.

In consideration of the above performance issues, a unified solution approach is desired that must incorporate the following recommendations;

1. Minimize the duration of the L2 scan operation, and hence the CAP discovery delay.
2. Ensure the timely delivery of L2 triggers to initiate the CARD process.
3. Remove the dependence of the CARD protocol on a central CARD Server.
4. Reduce the signalling load of the CARD protocol, especially over the error prone wireless link.

This implies a tightly coupled liaison between L2 and L3 operations and calls for a cross-layer management solution that must incorporate the above recommendations to enable a MN to discover NAR in the shortest possible time with low signalling latency and high probability of success.

4. Related Work

Two approaches for discovering CARs worth mentioning is Push-Mode-Multicast based Candidate Access Router Discovery (PMM CARD) (Dario et al., 2006(a)) and Access Router Information Protocol (ARIP) (Kwon et al., 2005). PMM CARD introduces added complexity of maintaining and managing multicast groups and addresses and its performance is restricted within a single operator's domain, whereas ARIP does not scale to complex network architectures and is not dynamic. Beside introducing additional signalling messages, ARIP requires the ARs to maintain identity information of the adjacent AR's but, similar to (Ono et al., 2003), suggests manual set up by the network administrator or by the automatic learning of the AR's from the handover information offered by the MNs. This makes ARIP unscalable to complex network architectures. Also both the above proposals do not provide any cross-layer management capabilities and have not been designed keeping in view the requirements and dynamics of a fast moving MN in the context of NGN.

In this chapter we present an enhanced and scalable mechanism for discovering CARs that can enable a fast moving MN to discover the identity and capabilities of the CARs which may not be geographically adjacent and/or directly linked to the PAR. We term this new approach as *Multi-hop Discovery of Candidate Access Routers (MHD-CAR)*, which is a simple approach that does away with the complexity and limitations of both PMM CARD and ARIP. The MHD-CAR provides an inherent scalable cross-layer mobility management solution that eliminates the performance issues discussed earlier by incorporating the solution recommendations.

The protocol details of the MHD-CAR protocol is submitted to the IETF as an Internet Draft (Yousaf, Wietfeld, 2008(a)) and the proof of concept presented in (Yousaf et al., 2008(b)).

5. MHD-CAR Operation Summary

MHD-CAR is a distributed mechanism proposed to enhance the operational reliability and robustness of seamless and fast handover protocols without introducing any additional message(s) and/or relying on any central server.

In MHD-CAR ARs dynamically update their local CAR Table with the identity information of not only the neighbouring ARs but also of CARs located multiple wireless-hops away through an iterative exchange of *unsolicited AR-AR CARD Reply* message and without relying on a CARD Server.

The CAR Table information of the current AR is then transferred to a MN on the fly where it updates/refreshes a local cache called *New Access Network (NAN) Cache* allowing the MN to resolve the CAR(s) locally with minimum exchange of Request/Reply messages over the error prone radio link.

The MHD-CAR protocol operation is depicted in Figure 5 and the functional details are discussed in the subsequent sub-sections.

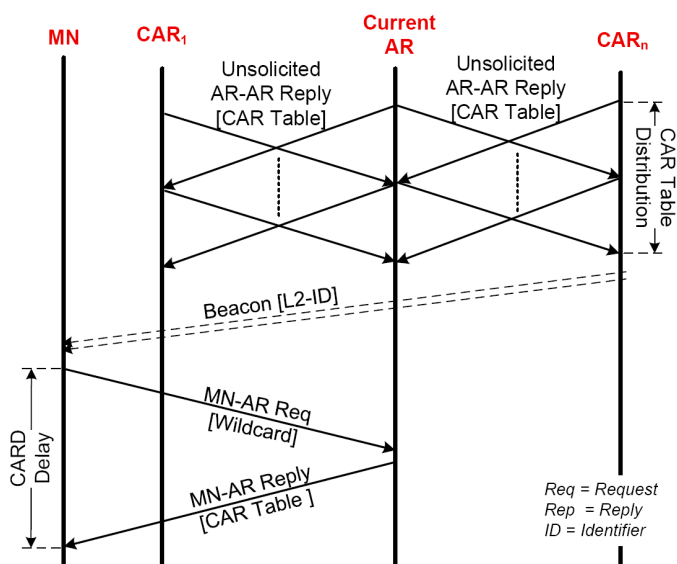


Fig. 5. MHD-CAR Protocol Message Sequence Diagram

5.1 CAR Table Initialization

The composition of the CAR Table in the context of MHD-CAR is different from the one proposed in (Liebsch et al., 2005) in that it offers more detailed information content. Table 1 shows the conceptual design of CAR Table in the context of MHD-CAR that contains not only the identities of the CAP/CAR but also information regarding the type of wireless access technology and the channel number in use by a CAP. It also informs about the capabilities of the CAP in terms of supported bit-rate and SSID (in case of 802.11), and most importantly the 'Distance' parameter, which is a measure of the distance of a CAR, in terms of the number of wireless-hops, with reference to the local AR maintaining the CAR Table.

At initialization, each AR will populate its CAR Table with its own (and associated AP(s)) identity and capabilities information and set the '*Distance*' parameter to zero, indicating local AR information.

For the MHD-CAR operation, it is imperative that each AR should be aware of the *identity* of the neighbouring AR and the associated AP(s) and store this information in its local CAR Table with a *Distance* value set to 1. The neighbour association can be established dynamically using the handover information of a bootstrapping MN as an input to establish a neighbour relationship. The first handover between any two neighbouring ARs will serve as a bootstrapping handover that would invoke the discovery process between the two ARs. This idea was first presented in (Shim, Gitlin, 2000), (Trossen et al., 2003) and also endorsed by the official CARD protocol standard (Liebsch et al., 2005) , which is adopted by the MHD-CAR operation for CAR Table initialisation and described as follows.

CAR Table		
AP Information	AR Information	Capabilities
<i>macaddr</i> MAC Address <i>int</i> L2 Type <i>int</i> Channel Number	<i>ipaddr</i> IP Address <i>prefix</i> Network Prefix <i>int</i> Prefix Length	<i>double</i> Bitrate AP <i>string</i> SSID <i>int</i> Distance <i>avpair</i> User Defined QoS AV Pair

Table 1. Conceptual Design of the CAR Table

When some MN performs an inter-AR handover, it will inform its current AR about the identity of the previous AR using a Router Identity message option appended to the *MN-AR CARD Request* message (Liebsch et al., 2005). The serving AR will acknowledge this with a *MN-AR CARD Reply* message and will store the identity of the MN's previous AR in its CAR Table and indicate it as its immediate neighbour. The serving AR will then send an *unsolicited AR-AR CARD Reply* message to the previous AR informing it of its own identity and identifying itself as its neighbour. The previous AR will thus store the identity information of the MN's current AR in its local CAR Table as an immediate neighbour indicated by the *Distance* value of 1. In this way, all the ARs along the motion path of the MN will bootstrap their local CAR Tables with the identity of the neighbouring ARs. Besides the identity information, the two ARs must also exchange the capabilities information with each other. This process will also eliminate the reliance on maintaining a CARD Server. It may be noted that the identity information contains both the IP address of the AR and the L2-Id(s) of the associated AP(s).

5.2 CAR Table Distribution

After the ARs are bootstrapped with the identity and capabilities of the neighbouring ARs, each AR will exchange its local CAR Table information with the neighbouring AR(s) through the iterative exchange of *unsolicited AR-AR CARD Reply Message*, where the number of iterations is equal to the specified maximum distance, in wireless-hops, corresponding to the maximum entries an AR will maintain in its CAR Table.

In the first iteration, the ARs will exchange their local CAR Table information (see Table 1) with their neighbouring ARs, which will add this new information to their local CAR Tables and increment the *Distance* value by 1. Now each AR will have information about the AR(s)

two wireless-hops away and this will be indicated by the *Distance* value of 2. It should be noted that the ARs do not forward the received inter-AR messages in order to prevent network flooding.

During the second iteration, the above process will be repeated and the receiving ARs will compare the new information with their present CAR Table entries and if no match is found, it will add the new CAR information to its local CAR Table by incrementing the distance parameter by 1.

Distribution Iteration	CAR Table Contents in ARs [Access Network Identifier (ID) – Distance (D)]															
	AR _A		AR _B		AR _C		AR _D		AR _E		AR _F		AR _G		AR _H	
	Id	D	Id	D	Id	D	Id	D	Id	D	Id	D	Id	D	Id	D
0	A	0	B	0	C	0	D	0	E	0	F	0	G	0	H	0
	B	1	A	1	B	1	C	1	D	1	E	1	F	1	I	1
1	A	0	B	0	C	0	D	0	E	0	F	0	G	0	H	0
	B	1	A	1	B	1	C	1	D	1	E	1	F	1	I	1
	C	2	D	2	A	2	E	2	F	2	D	2	E	2	J	2
2	A	0	B	0	C	0	D	0	E	0	F	0	G	0	H	0
	B	1	A	1	B	1	C	1	D	1	E	1	F	1	I	1
	C	2	D	2	A	2	E	2	F	2	D	2	E	2	J	2
	D	3	E	3	F	3	A	3	G	3	C	3	D	3	K	3
3	A	0	B	0	C	0	D	0	E	0	F	0	G	0	H	0
	B	1	A	1	B	1	C	1	D	1	E	1	F	1	I	1
	C	2	D	2	A	2	E	2	F	2	D	2	E	2	J	2
	D	3	E	3	F	3	A	3	G	3	C	3	D	3	K	3
	E	4	F	4	G	4	H	4	I	4	J	4	K	4	L	4

Fig. 6. Conceptual Representation of the Iterative CAR Table Distribution Process in MHD-CAR

This new entry will thus be marked with the *Distance* value of 3, indicating that the corresponding CAR is at distance of three wireless-hops.

The above inter-exchange of unsolicited inter-AR Reply messages will continue until each AR has the information about CAR, which is at the specified maximum distance from the current AR, after which the CAR Tables are said to have *converged*. The maximum distance for which an AR is supposed to maintain CAR information is a constant that depends on the network topology and can be specified by the administrator. The ARs will then periodically exchange their local CAR Table information with their neighbouring AR(s) after every 60 seconds, or when there is some change in the contents of it (for example, change in capabilities information). *During CAR Table initialisation and distribution process, the ARs will transmit the unsolicited inter-AR messages at a uniformly distributed random time between 0 and 100 msec.*

The iterative CAR Table distribution process is illustrated in Figure 6, which shows the contents of the CAR Table of an AR for each iteration of the distribution process. The distribution takes place for a *Distance* value of 4, i.e., the iterative distribution process will continue till each AR has information about ARs which are up to 4 wireless hops away. Figure 6 illustrates the process for the reference topology shown in Figure 7 composed of 12 Access Networks (AN) (from A to L), where the domain of each AN is defined by an AR and an associated AP. For the sake of demonstration and simplicity only the AN *Identity* and *Distance* parameters of a CAR Table are considered in Figure 6. The *identity* (*Id*) is characterised by the [L3-ID, L2-ID] pair and is denoted by the AN identifier, whereas the *Distance* (*D*) signifies the topological distance of a CAR/CAP from the local AR in terms of wireless hops. The entries indicated in red signify the new entries that get stored in the CAR Table during the particular distribution iteration. Iteration # 0 signifies the contents of the CAR Table during the initialisation process as explained above.

Figure 7 clearly shows how the ARs (e.g., F) acquire the information of CARs (for instance B & J) that are not immediate neighbours through the iterative distribution of the CAR Tables.

NAN Cache	
CAP Information	CAR Information
<i>bool</i> Reachability	
<i>macaddr</i> MAC Address	
<i>int</i> L2 Type	<i>ipaddr</i> IP Address
<i>int</i> Channel Number	<i>prefix</i> Network Prefix
<i>double</i> Bitrate	<i>int</i> Prefix Length
<i>double</i> Last Received Beacon time	<i>int</i> Distance
<i>double</i> RSSI	
<i>string</i> SSID	

Table 2. Conceptual Design of New Access Network (NAN) Cache

5.3 Mobile Node Operation

The CARD protocol specification (Liebsch et al., 2005) suggests a MN to maintain address and capability information of CAR(s) discovered during *previous* CARD operation in a local cache to avoid requesting the same information repeatedly and to select an appropriate TAR as quickly as possible when a handover is imminent, but it does not specify the conceptual design of such a cache. Besides, this proposal will only improve the CAR selection if the MN

is *revisiting* some previously visited CAR domain, a situation not very much likely in case of high speed MNs.

MHD-CAR proposes a MN to maintain a local cache called *New Access Network (NAN) cache* (see Table 2) that maintains the identity and capabilities information of not only the neighbouring CAR(s) and associated CAP(s) but those located multiple wireless-hops away. This will allow the MN to perform RAT and DCC functions locally without the exchange of any MN-AR *CARD Request/Reply* message pair with its current AR, or without involving a CARD server, every time handover is imminent.

The information content of the NAN cache, that is expected to enhance the MN's TAR selection process and handover related decision tasks, is derived mostly from the CAR Table that is usually *pulled* by the MN on the fly from its current AR (via a wildcard MN-AR *CARD Request* Message) when the MN senses that it is moving away from its current AR. The MN will then add and/or refresh the relevant entries of its cache. The NAN cache also derives some of the AP related information from the periodic beacon signals received from the in-range wireless APs.

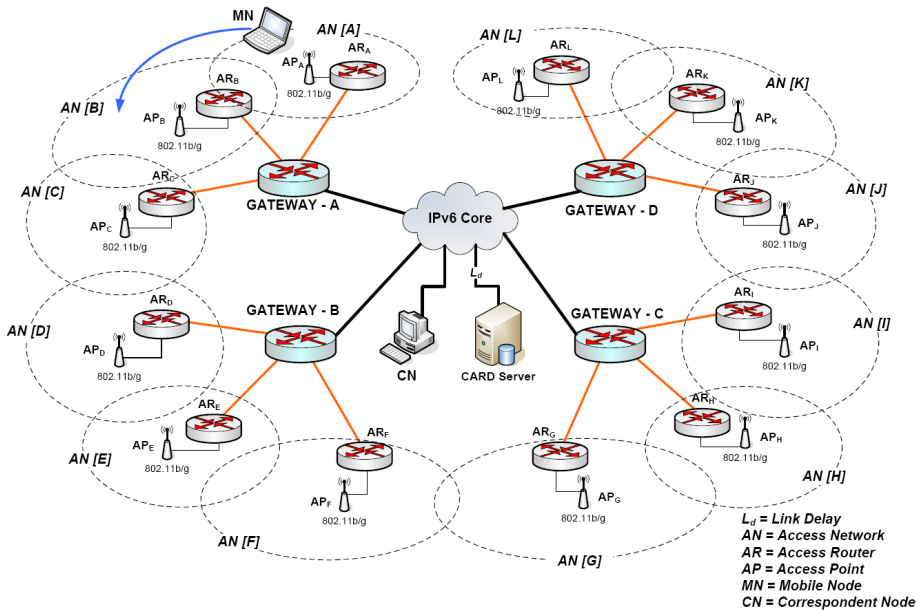


Fig. 7. Simulation Topology

6. Performance Analysis

In this Section we present the results of the simulation experiments comparing and analyzing the performance of the proposed MHD-CAR mechanism to that of the IETF's CARD protocol using the CARD server. Both the protocols are modeled in our mobility management framework (Yousaf et al., 2008(c)) developed in OMNeT++ (OMNeT++) and using realistic message structures and timer implementations.

The simulation experiments and its analysis is similar to the one presented by us in (Yousaf et al., 2008(b)) with the difference that in this chapter we have extended our simulation model to realise a more realistic *hierarchical network topology* instead of a *flat network topology* in which all the ARs were directly connected to their immediate neighbours via Ethernet links. In a flat network topology, the ARs will be able to derive the identity of their neighbouring ARs by simply sending relevant messages directly and hence populate their local CAR Tables. In contrast, to realise the MHD-CAR protocols in the hierarchical network topology, we have extended our simulation framework with the bootstrapping mechanism (see Section 4.1) that would enable the ARs to discover the IP address of the neighbouring ARs. With this major difference, the same experiments were repeated and found the results to match those presented previously in (Yousaf et al., 2008(b)).

The simulation network topology is shown in Figure 7 and the experiments are carried in a homogeneous 802.11b wireless environment using a free space propagation model at 2.4 GHz for a radio channel over a total coverage area of 800m x 800m. To initialise the CAR Tables, in ARs a bootstrapping MN is moved across the reference network at the beginning of the simulation enabling each AR to become aware of the IP address of the previous neighbouring AR. The results expressed in this Section will also apply equally to a heterogeneous environment because the MHD-CAR operation is defined at the network layer.

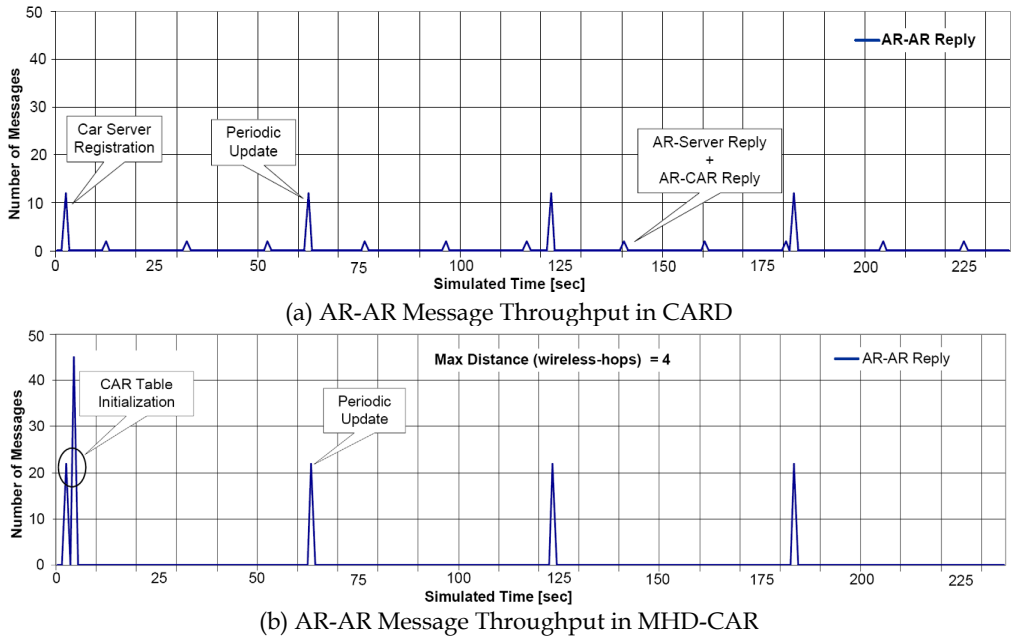


Fig. 8. Inter AR Message Throughput in (a) CARD, and (b) MHD-CAR

A single simulation run consists of a MN moving across 12 ARs, starting from AR_A and undergoing 11 handover instances, discovering and resolving the next CAP(s)/CAR(s) while it is still connected to its current AR. The experiments are repeated 100 times for each

of the 25 seed values generated using the OMNeT++'s *sedtool* thereby constituting a total of 2500 runs each for simulating CARD and MHD-CAR protocol. The results are then expressed as average sum of the measured parameters.

6.1 Impact of MHD-CAR on Signaling Load

The average signalling load over the wireless link (MN-AR Messages) and the wired links (AR-AR Messages) for both the CARD and MHD-CAR is compared in Figure 10 and the results tabulated in Table 3. The ARs in both CARD and MHD-CAR periodically update their CAR Tables after 60 seconds and in our simulation this update takes place three times.

6.1.1 Signaling load over the Inter-AR links

In CARD each AR, during boot-up time, register its identity information with the CARD server and keeps the CARD server updated by sending *unsolicited AR-Server Reply Messages* periodically after every 60 seconds. The number of *unsolicited AR-Server Reply Messages* transmitted during boot-up time and during periodic updates remains equal and corresponds to the number of ARs. In between the periodic updates, the current AR will exchange *AR-AR CARD Request/Reply* message pair with the CARD Server for performing RAT function (upon MN's request) and *AR-AR CARD Request/Reply* message pair with the resolved CAR(s) for performing the DCC function. This is seen in Figure 8(a) where we show the occurrence of the average inter-AR (Server & CAR) reply message throughput. In MHD-CAR however, since the MN is able to locally resolve the identity and capabilities of CAR(s) based on the information in NAN Cache (see Section 4.2), the AR does not need to exchange any request/reply message pair with some server or neighbouring AR(s) thereby resulting in 100% reduction of inter-AR signalling load related to discovering CAR(s). This is depicted in Figure 8(b) where the ARs only exchange *unsolicited AR-AR Reply* messages with their neighbouring ARs as part of the periodic update of the local CAR Table after every 60 seconds.

	Messages			
	MN-AR		AR-AR ²	
	Requests	Replies	Requests	Replies ³
CARD	23	23	22	58
MHD-CAR	12	12	0	66
Message Load (%)	-47.8 %	-47.8 %	-100 %	+13.8%
Total Load Reduction (%)	47.8%		17.5%	

Table 3. MN-AR & AR-AR Message Load Comparison

Since there is no central server, the inter AR *update message* throughput in MHD-CAR is 83.33% higher than CARD but overall recording a 17.5% reduction in the *post boot-up* exchange of inter-AR message (see Table 3).

² The inter-AR messages are inclusive of the AR-Server messages for the CARD operation

³ The AR-AR Reply messages take into account only the replies messages exchanged after the initialization and distribution process.

During boot-up MHD-CAR records a much higher throughput than CARD due to the iterative inter-AR exchange of local CAR Tables (see Section 4.1 & 4.2) till the Table entries have converged to the maximum specified distance, which is four in our scenario. Evidently the value of maximum distance will have a direct impact on the number of inter-AR messages, but this is of no serious consequence as it happens only once and that too during initialisation via bootstrapping, afterwards which the CAR Tables converge very quickly. Figure 8(b) shows a sharp surge of inter-AR messages during CAR Table initialisation and distribution period but it is seen that the ARs converge within the first 3-4 seconds. Table 3 lists and compares the average number of inter-AR messages exchanged for both the protocols and depicts the percentage reduction in the overall inter-AR messages induced by MHD-CAR.

6.1.2 Signalling load over wireless link

As depicted in Figure 9, a MN in a CARD protocol will exchange *MN-AR CARD Request/Reply* message pair with its current AR twice, once in the beginning (soon after connecting to it) to request a list of CAR(s) information maintained in its CAR Table; and the second time when handover is imminent and the MN will need to resolve the L2-ID(s) (received in the beacon messages from in-range AP(s)) that may not have been present in the current AR's CAR Table, and/or to find out the capabilities of the CAP(s)/CAR(s), so that the MN may select a suitable target AR based on some selection criteria (Dario et al., 2006(b)). This approach is suitable for a slow moving MN with a long dwell time in a single cell but not suitable for a fast moving MN.

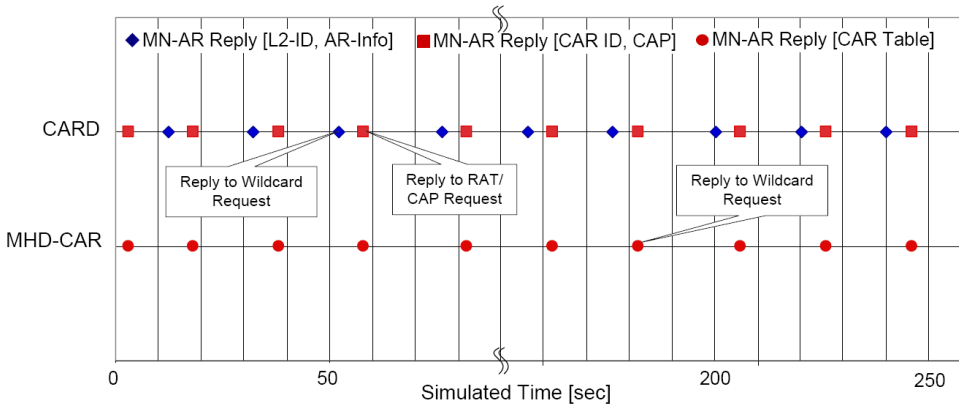


Fig. 9. MN-AR Reply Message Timestamps for CARD and MHD-CAR over the Wireless Link

In contrast, MHD-CAR requires only a single exchange of *MN-AR CARD Request/Reply* Message pair, in which the MN sends a wildcard request and the current AR will send the contents of its CAR Table (see Table 1) in the corresponding reply message and the MN will cache the information in its NAN Cache (see Table 2). This will contribute towards the reduction of the signalling load over the wireless link by almost 48%, as seen from the comparison of the message timestamps of *MN-AR Reply* messages transmitted by the CARD and MHD-CAR depicted in Figure 9 and hence the throughput reduction achieved by MHD-CAR over the wireless link is clearly evident. Table 3 lists and compares the average

number of MN-AR messages exchanged over the wireless link for both the protocols and depicts the percentage reduction of message load induced by MHD-CAR with reference to the scenario depicted in Figure 7.

6.2 Impact of MHD-CAR on Scanning Delay

It is a known fact that the L2 handover delay is a major delay component adding to the overall handover latency for MIPv6 (Yousaf et al., 2008(c)) and thus a major impediment to the performance of higher layer mobility management schemes. The main contributing factor to the L2 handover latency is the delay incurred by the *channel scan* operation as part of the CAP discovery process (see Section 3). Typically a MN *after* losing its connectivity with its current AP will perform the 802.11 *all-channel scan* (active/passive) on, in our case, 13 frequency channels. This scan-delay is certainly not suitable for attaining seamless handover performance for fast moving MNs and various methods have been proposed in this respect. One of the consensus solutions to reduce the L2 handover delay is to perform scanning on selective channels (Park et al., 2004) or perform a pro-active scan, i.e., before a MN loses its connection with its current AP (Haito et al., 2007).

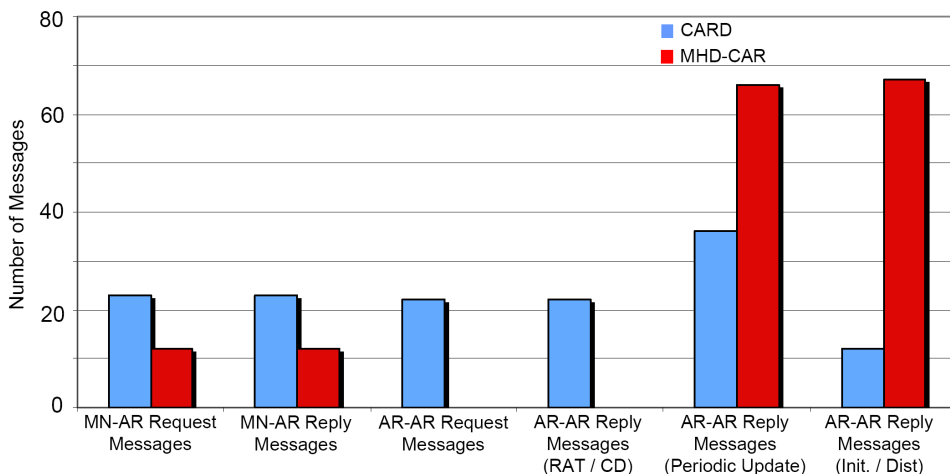


Fig. 10. Signalling Load Comparison of CARD and MHD-CAR

The MHD-CAR based on the information available in the NAN Cache proposes to reduce the scan delay by enabling a MN to perform *target scanning* on selective frequency channels. The MN, instead of waiting to disconnect from the current AP, will start to scan for only those channels that correspond to the nearest CAP(s) as specified by the *Distance* parameter in the NAN Cache, and if the CAP(s) are not located (or are out of range) will proceed to select and scan the next set of frequency channels corresponding to next farther CAP(s). The MN will typically start the scan process when the RSSI from the current AP falls below a certain threshold. Figure 11 compares the performance of MHD-CAR's target scanning with the 802.11 all-channel scan.

From the Figure 11 it is evident that the performance of MHD-CAR's target scanning incurs less delay than the full channel scan, however the delay performance is a function of the size

of the NAN Cache and location of the MN. For example it is observed that when the *Distance* value is 4 and MN is undergoing the 7th handover, target scanning incurs more delay. This increase in delay, however still less than the full-channel scan, is due to the fact that the MN is at AR_G and the CAR Table in the AR_G will have a total of 8 CAR entries and thus the MN has to scan all the 8 channels.

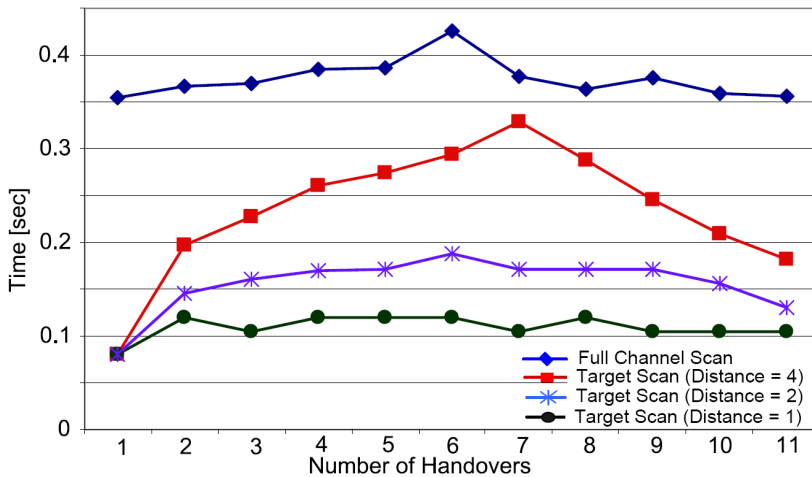


Fig. 11. Performance of MHD-CAR's Target Channel Scanning

7. Comparison with IEEE 802.21 MIHS

The IEEE 802.21 is an emerging standard (IEEE Std 802.21) that aims at imparting seamless mobility to MNs traversing diverse radio access technologies in a heterogeneous network environment. The motivation is to facilitate inter-WAT handover between IEEE 802 and non-IEEE 802 access networks in such a way that the process negotiation is independent from the specific features of the underlying access network technology. In other words, IEEE 802.21 is being designed to provide *Media Independent Handover (MIH)* service that is tasked with the *handover initiation* and *handover preparation* whereas leaving the task of *handover execution* to other protocols like FMIPv6. The prerequisite to preparing a MN for a handover is the discovery of new links in the vicinity and this, according to (George et al., 2008), may involve maintaining a remote information server (similar to a CARD Server) that can be queried to get information about the available networks in the area of a specific MN and/or relying on the MN's scanning operation to select the network of its choice.

The MIH services are being realised by defining *Media Independent Handover Function (MIHF)*, which is implemented as a separate protocol stack located between the *Network Layer* and the *Data Link layer* as shown in Figure 12. The MIH operation is dictated by the MIHF exchanging events, commands and information messages with the adjacent layers. It should be noted that the MIHF is supposed to be located within the protocol stack of both the MN and the Network Node mandating major revision of the present network infrastructure and nodes. This would also mean major enhancements and modifications to the individual media specific technologies (802.11, 802.16, CDMA, UMTS, GSM, GPRS, etc)

in the form of defining and developing new SAPs and primitives that would interface with the generic SAP and primitives defined for the IEEE 802.21 before a MN can take advantage of the MIH services (Eastwood & Migaldi, 2008). This would entail a major re-engineering effort involving the upgrade of the whole network infrastructure and protocol standards. Besides, the IEEE 802.21 model diverges from the standard ISO/OSI reference protocol model by introducing an intermediate layer between L2 and L3.

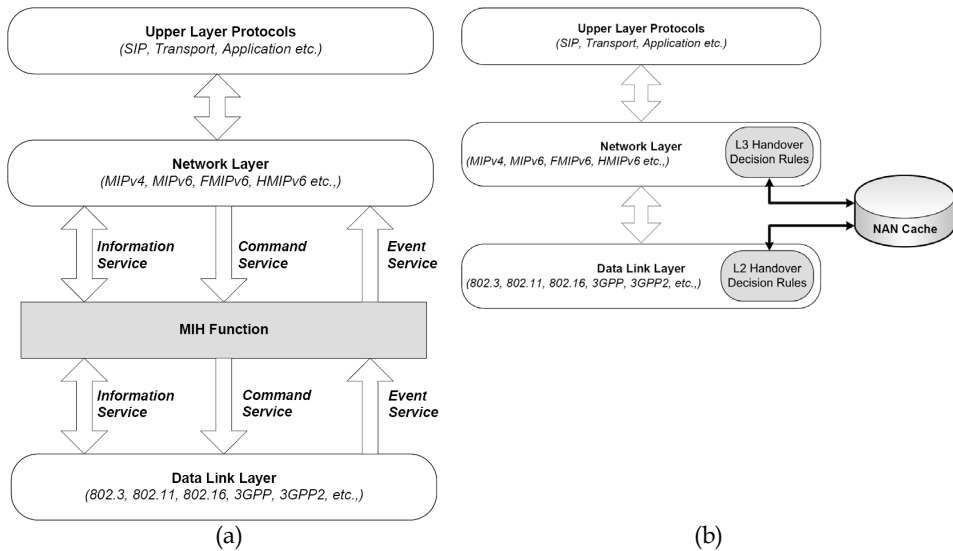


Fig. 12. Conceptual Models of (a) The IEEE 802.21 MIH Service Model, and (b) The MHD-CAR Protocol

Similar to IEEE 802.21, the operational scope of MHD-CAR is to provide a mechanism to enable a multi-homed MN to undergo inter-RAT handover by providing the requisite information content that would enable a MN to choose the best network that would suit its application service requirements. However, in sharp contrast to IEEE 802.21 service model depicted in Figure 12(a) which is implemented as a protocol stack, the MHD-CAR is based on managing and maintaining a NAN Cache inside the MN that can be accessed by the mobility functions defined at the network layer and the data link layer. This translates into defining simpler interfaces at L2 and L3 for interaction with the NAN cache rather than demanding major revisions from the access technologies as in the case of IEEE 802.21 highlighted above. The NAN Cache thus provides cross-layer capabilities.

In contrast to IEEE 802.21, the MHD-CAR protocol is simply an optimised version of the existing CARD protocol without introducing any new messages, interfaces and/or network entities, or deviating from the reference ISO/OSI reference model making it scalable and deployable and without any burden on the network itself.

8. Conclusions

In this chapter we have provided operational and functional details of a proposed protocol called MHD-CAR that has been designed in view of the stringent performance requirements imposed by fast moving MNs in terms of seamless and fast handovers in a heterogeneous wireless network environment. MHD-CAR optimises the standard CARD protocol in enabling a MN to discover on the fly the identity and capabilities of not only the neighbouring CARs but also CARs that may be located multiple hops away, and all this is achieved with minimum reliance on the network. This is expected to augment the performance of seamless handover protocols like FMIPv6 by ensuring accurate selection of NAR with minimum discovery latency.

The MHD-CAR is a distributed mechanism in which the ARs are able to inter-communicate their identities and capabilities information to neighbouring ARs and to ARs that are located multiple wireless-hops away without relying on maintaining and managing a central CARD server. Each AR stores this information in their local CAR Tables which are then communicated to the MN upon request. Due to the distributed mechanism, MHD-CAR is more efficient, reliable, survivable and scalable protocol than the CARD protocol. Since the MHD-CAR does not introduce any new protocol messages it can therefore be easily integrated into the present deployment infrastructure.

It exhibits far better performance over the IETF's CARD protocol in terms of the substantial reduction of signalling load over both the inter-AR links (by 17.5%) and crucially over the error prone wireless link (by 48%), while utilizing the CARD protocol messages. This reduction in signalling load is achieved because the MN is able to perform RAT and DCC functions locally, based on the information content of the NAN cache, and without relying on the network.

Another very important aspect of the MHD-CAR scheme is that it provides cross-layer liaison between L2 and L3 mobility function. This is achieved by having a NAN cache in the MN, which provides the MN with a topological snapshot of the identity and capabilities of the access networks that may be multiple hops away from its present point of attachment. This enables a MN to perform target scanning on selected channels greatly reducing the CAP discovery latency and enhancing the accuracy of the TAR selection process. This alone will have a direct impact on the overall handover latency and fast moving MNs will greatly benefit from it.

In contrast to the IEEE 802.21 standard, it is observed that the MHD-CAR is a light weight and much simpler alternative solution that provides the main functional services of the 802.21 MIHS. Although MHD-CAR has not been designed as an alternative to 802.21 but it does share its motivational, operational and functional scope. The IEEE 802.21 WG was developed to provide a unified global mechanism by defining a common MIH layer sandwiched between the Network Layer and the Data Link Layer and defined common triggers that would be generated independent of the underlying access technology. The motivation was to enable the MN to make accurate selection of the network and to provide triggers that would aid the IP mobility protocols like FMIPv6. However all this is being introduced at the cost of high complexity while deviating from the base ISO/OSI prescribed layered approach by introducing a new layer between the L2 and L3. Also it would mandate changes to the existing access technologies to conform to the MIHS scheme of signalling. For example different SAPs are required to be defined for each of the access technology. It

would also involve the exchange of signalling message between the network and the MN, even over the air interface.

MHD-CAR therefore provides the same conceptual functionalities defined for MIHS but without transgressing the functional boundaries of the standard OSI/ISO protocol reference model and with a much simpler and scalable approach.

9. References

- Dario, D.; Femminella, M.; Piacentini, L. & Reali, G. (2006(a)). Performance Evaluation of the Push-Mode-Multicast based Candidate Access Router Discovery (PMM CARD), *Computer Networks*, Volume 50, No. 3, (February 2006) page numbers (367-397), ISSN 1389-1286.
- Dario, D.; Femminella, M.; Piacentini, L. & Reali, G. (2006(b)). Target Access Router Selection in Advanced Mobility Scenarios. *Computer Communications*, Volume 29, No. 3, (February 2006), page numbers (337-357), ISSN 0140-3664.
- Eastwood, L. & Migaldi, S. (2008). Mobility Using IEEE 802.21 in a Heterogeneous IEEE 802.16/802.11 -Based, IMT-Advanced (4G) Network, *IEEE Wireless Communications*, Vol. 15, No. 2, (April 2008) pp. 26-34, ISSN: 1536-1284.
- Funato, D.; He, X.; Williams, C.; Takeshita, A.; Ali, M. & Nakfour, J. (2002). Geographically Adjacent Access Router Discovery Protocol. *draft-funato-seamoby-gaard-01.txt*, Internet Draft (IETF), June 2002. Work in Progress.
- George, L.; Salkintzis, A. & Passas, N. (2008). Media-Independent Handover for Seamless Service Provision in Heterogeneous Networks. *IEEE Communication Magazine*, Vol. 46, No. 1, (January 2008) pp. 64-71, ISSN 0163-6804.
- Haito, W.; Tan, K.; Zhang, Y. & Zhang, Q. (2007). Proactive Scan: Fast Handoff with Smart Triggers for 802.11 Wireless LAN, *Proceedings of 26th IEEE International Conference on Computer Communications (INFOCOM)*, pp. 749-757, ISBN 1-4244-1047-9, Alaska (USA), May 2007, IEEE, USA.
- IEEE Std 802.21. Draft D9.0, IEEE Standard for Local & Metropolitan Area Networks: Media Independent Handover Services.
- Koodli, R. (2008). Mobile IPv6 Fast Handovers, *Request for Comments (Proposed Standard) 5268*, Internet Engineering Task Force, June 2008.
- Kwon, D.; Kim, Y.; Bae, K. & Suh, Y. (2005). Access Router Information Protocol with FMIPv6 for Efficient Handovers and their Implementations. *Proceedings of Global Telecommunications Conference (GlobeCom)*, pp. 3814-3819, ISBN 0-7803-9414-3, St. Louis (Missouri), IEEE, December 2005, USA.
- Liebsch, M.; Singh A.; Chaskar, H.; Finato, D. & Shim, E. (2005). Candidate Access Router Discovery Protocol (CARD), *Request for Comments (Proposed Standard) 4066*, Internet Engineering Task Force, July 2005.
- Mishra, A.; Shin, M. & Arbaugh, W (2003). An Empirical Analysis of the IEEE 802.11 MAC Layer Handoff Process, *Proceedings of ACM SIGCOMM Computer Communication Review*, Volume 33, No. 2, pp. 93-102, ISSN 0146-4833, April 2003.
- Narten, T.; Nordmark, E., Simpson, W. & H. Soliman, (2007). Neighbor Discovery for IP Version 6 (IPv6)", *Request for Comments (Draft Standard) 4861*, September 2007.
- OMNeT++. Community Site, <http://www.omnetpp.org>, July 2009.

- Ono, N.; Kimura, T. & Fujii, T. (2003). A Study on Autonomous Neighbour Access Router Discovery for Mobile IP. *Proceedings of 57th IEEE Vehicular Technology Conference (VTC)*, pp. 2241-2245, ISBN 0-7803-7757-5, Jeju, April 2003, Korea.
- Perkins, C.; Johnson, D. & J. Arkko, (2004). Mobility Support in IPv6, *Request for Comments (Proposed Standard) 3775*, Internet Engineering Task Force, June 2004.
- Park, S; Kim, H.; Park, C.; Kim, J. & Ko, S. (2004). Selective Channel Scanning for Fast Handoff in Wireless LAN using Neighbour Graph, In: *Personal Wireless Communications*, Lecture Notes in Computer Science, page numbers (194-203), Springer, ISBN 978-3-540-23162-2, Berlin / Heidelberg, September 2004.
- Shim, E. & Gitlin, R. (2000). Fast Handoff Using Neighbor Graph Information. *draft-shim-mobileip-neighbor-00.txt*, November 2000. Work In Progress.
- Thomson, S.; Narten, T. & Jinmei, T. (2007). IPv6 Stateless Address Autoconfiguration, *Request for Comments (Draft Standard) 4862*, Internet Engineering Task Force, September 2007.
- Trossen, D.; Krishnamurthi, G.; Chaskar, H. & Kempf, J. (2002). Issues in Candidate Access Router Discovery for Seamless IP-Level Handoffs. *draft-IETF-seamoby-cardiscovery-issues-04.txt*, Internet Draft (IETF), October 2002. Work In Progress.
- Trossen, D. ; Krishnamurthi, G.; Chaskar, H.; Chalmers, R. & Shim, E. (2003). A Dynamic Protocol for Candidate Access Router Discovery. *draft-trossen-seamoby-dycard-01.txt*, Internet Draft (IETF), March 2003. Work In Progress.
- Yousaf, F. Z. & Wietfeld, C. (2008, a). Multi-Hop Discovery of Candidate Access Routers (MHD-CAR), *draft-yousaf-ietf-network-mhdcar-00.txt*, Internet Draft (IETF), April 2008. Work in Progress.
- Yousaf, F. Z.; Müller, C. & Wietfeld, C. (2008, b). Multi-Hop Discovery of Candidate Access Routers (MHD-CAR) for Fast Moving Mobile Nodes, *Proceedings of 19th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1-5, ISBN 978-1-4244-2643-0, Cannes, September 2008, France.
- Yousaf, F. Z.; Bauer, C. & Wietfeld, C. (2008, c). An Accurate and Extensible Mobile IPv6 (xMIPv6) Simulation Model for OMNeT++", *1st ACM/ICST International OMNeT++ Workshop on the SIMUTools Conference*, Marseille, March 2008.

Mobility in IP Networks: From Link Layer to Application Layer Protocols and Architectures

Thienne Johnson¹, Eleri Cardozo², Rodrigo Prado², Eduardo Zagari²
and Tomas Badan³

¹*University of São Paulo*

²*State University of Campinas*

³*Federal University of Goiás
Brazil*

1. Introduction

With the popularization of the Internet and mobile devices, like notebooks and PDAs (Personal Digital Assistants), the concept of portability started to become popular, in the sense that the user could take their device anywhere and start a new connection to the Internet.

In the Internet, a node is identified by an IP (Internet Protocol) address that uniquely identifies its point of attachment to the Internet, and packets are routed to the node based on this address. Therefore, a node must be located on the network indicated by its IP address in order to receive datagrams (Akyildiz et al, 2004). However, when moving to different networks, the IP protocol does not allow the current IP address to be valid in the visiting network and the device should ask for a new IP address when entering a new network (Figure 1). It was then necessary to provide a scheme to allow nodes to be reachable and maintain ongoing connections while changing their location within the topology, in a seamless¹ way, when leaving its home (initial) to a visiting (new point of attachment) network.

One of the first solutions proposed to solve this problem for IP networks was the Mobile IP Protocol (Perkins, 1997), also known as MIP. The MIP protocol aims to solve the problem of node mobility by redirecting packets to the mobile node (MN) to its current location. MIP was a scheme suited for interdomain mobility, allowing MN's movements between different networks from different domains.

But protocols proposed for interdomain mobility are not suited to intrasubnet mobility due to drawbacks such as the need for new protocols on the MNs that exceed their processing power what make these solutions simply undeployable in most mobile devices. The problems are related to the complexity of the proposed solutions which make their

¹ Seamless mobility: capability to change the mobile node's point of attachment to an IP-based network, without losing ongoing connections and without disruptions in the communication.

implementation on small mobile devices such as cell phones and handhelds unfeasible. In fact, manufactures of such devices never considered supporting the present solutions (Zagari et al, 2008).

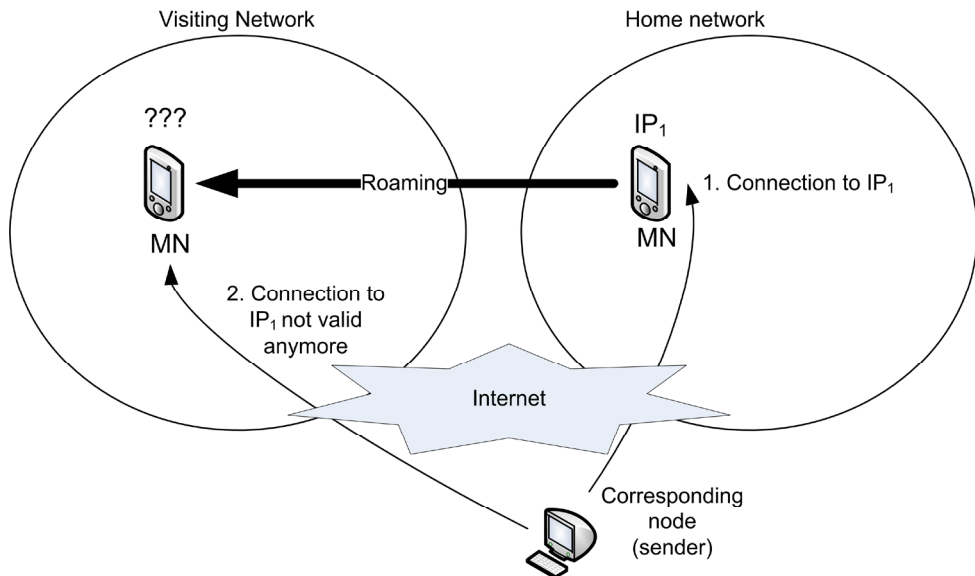


Fig. 1. Current IP address not valid in a visiting network

After MIP, new protocols for mobility between networks in the same domain were proposed. Micromobility protocols aim to improve localized mobility by reducing handover overheads. Other approaches to allow seamless mobility also include mobility provided by transport and session layer schemes.

The objective of this chapter is to provide a major review on mobility protocols and architectures. The protocols and architectures will be classified according to their mobility range (intra and inter domain), layer in which it operates (from the link layer to the application layer), and the support required from the mobile node. We will place emphasis on the solutions called “network-centered”, that is, solutions where mobility is handled entirely by the network without the need of installation of mobility protocols on the mobile nodes. The protocols and architectures discussed in this chapter are being proposed by standardization bodies, e.g., IETF, by industry-driven forums, e.g., 3GPP, by academy and by the industry.

This chapter is divided into 10 more sections. Section 2 presents an overview of mobility issues and a classification schema, which will be used in the protocols sections. Mobile IP is presented in Section 3. Section 4 shows link layer based protocols. Section 5 presents mobility solutions based on L2½ protocols. Section 6 presents network layer protocols. Section 7 presents transport layer protocols. Section 8 presents mobility using application layer protocols. Section 9 presents the Mobility Plane Architecture. Section 10 presents a general classification of the mobility solutions seen in this chapter and future work related to mobility in IP networks. Finally, Section 11 concludes this chapter.

2. Mobility Issues

The mobility process starts with the mobile node's attachment to a local wireless network. The node attachment process happens when a MN enters in the coverage of a wireless access point. In this point a L2 (Layer 2) attachment process is performed. After that, the MN must acquire its IP address from the network. After obtaining its new address, the network is able to route packets to/from the mobile node (Johnson et al, 2008).

When the MN moves away from the current access point, it may detect another wireless access point. Handover is "the process by which an active MN changes its point of attachment to the network, or when such a change is attempted. The access network may provide features to minimize the interruption to sessions in progress" (Manner & Cojo, 2004) by preserving the transport (or higher layers) connections such a way the packets are forwarded to the MN via the new access point.

Mobility management is the key to enable this seamless mobility. It enables wireless or mobile networks to search and locate mobile devices for network communications and to maintain network/applications connections as the MN moves into a new service area. The mobility management is composed of mainly two services: location management and handover management.

Location management consists of two operations: registration or location update and paging, to enable a network to discover the current point of attachment of an MN for information delivery (Saha et al, 2004). Location update is used in support of idle users, and paging is used in support of active communications (Campbell et al, 2002).

Location update procedures need the MN to periodically inform the system to update relevant location databases with its up-to-date location information (Akyildiz et al, 2004). Paging is the ability to track idle mobile hosts. For protocols using this kind of tracking, idle MNs do not have to register if they move within the same paging area, but only if they change paging area (Campbell & Gomez-Castellanos, 2000).

Handover management enables the network to maintain a MN's connection as it continues to move and change its access point to the network (Saha et al, 2004). There are many types of handover, among which are:

- horizontal and vertical: horizontal handover occurs between wireless cells of the same technology; vertical handover occurs between two different networks of different technologies. This chapter is dedicated to study horizontal handover only.
- mobile and network initiated: in MN initiated handovers the MN is responsible for initiating handover requests, while in network initiated handover the network is responsible for indicating that a handover must occur.
- MN and network controlled: in MN controlled handover, the MN must participate in the handover process, while in the network controlled handover the network handles the entire process.
- fast: fast handover tries to reduce the latency during a handover.
- seamless: change the MN's point of attachment to an IP-based network, without losing ongoing connections and without disruptions in the communication.

The basic terminology for mobility is (Perkins, 2002; Manner & Kojo, 2004):

- home network: a network having a network prefix matching that of a mobile node's permanent address;

- home address: the IP address acquired when registering in its home network; a stable address that belongs to the mobile node and is used by correspondent nodes to reach mobile nodes;
- home agent (HA): A router located on the home network that acts on behalf of the mobile node while away from the home network;
- correspondent node (CN): Any node that communicates with the mobile node;
- foreign network: Any network (other than the home network) visited by a mobile node;
- foreign agent (FA): A router located on the foreign network that acts on behalf of the MN in this network;
- care-of-address (CoA): An address that is assigned to the mobile node when located in a foreign link.

2.1 Classification parameters

Existing proposals for mobility can be broadly classified into different types, based on many parameters. We will employ a taxonomy based on 4 axis: Mobility Range, Mobility Routing, Mobility Signaling and Mobility Layer. In this section we will briefly introduce each one of them.

2.1.1 Mobility Range

In this work we adopted the definitions by Manner & Kojo (2004) for the mobility scope.

Micromobility

Also called intradomain mobility or local mobility (Kempf, 2007), is the process of mobility over a small area. Usually this means mobility within an IP domain with an emphasis on support for active mode using handover, although it may include idle mode procedures also. Micromobility protocols exploit the locality of movement by confining movement related changes and signaling to the access network.

Macromobility

Also called interdomain mobility or global mobility (Kempf, 2007), is the process of mobility over a large area. This includes mobility support and associated address registration procedures that are needed when a MN moves between IP domains. Interdomain handovers typically involve macromobility protocols. MIP can be seen as a means to provide macro mobility.

2.1.2 Mobility Routing

Defines how the MNs' location information database is created and maintained.

Routing based

Routing based schemes aim to exploit the robustness of conventional IP forwarding. A distributed mobile host location database is created and maintained within the network domain. The database consists of individual flat mobile-specific address lookup tables and is maintained by all the mobility agents within the domain (Chiussi et al, 2002).

Tunnel based

In tunnel based schemes, the location database is maintained in distributed form by a set of foreign agents in the access network. Each foreign agent reads the incoming packet's

original destination address and searches its visitor list for a corresponding entry. If the entry exists then it contains the address of next lower level foreign agent.

The sequence of visitor list entries corresponding to a particular mobile host constitutes the host's location information and determines the route taken by its downlink packets. Entries are created and maintained by registration messages transmitted by mobile hosts (Campbell & Gomez-Castellanos, 2000). Tunnels may be IP-IP (IP over IP) or MPLS (Multi-protocol Label Switching).

2.1.3 Mobility Signaling

Defines if the mobility signaling is carried out by the network alone or also needs the mobile node participation in the signaling process.

Mobile Node Centric

In this approach, the MN must execute an instance of the mobility protocol, thus participating actively in the mobility management process. But the requirement for modification of MNs software may increase their complexity; considering these nodes have less computational capacity, it may lead to performance degradation on the MN.

Network Centric

In network-based mobility management approach, the serving network handles the mobility management on behalf of the MN; thus, the MN is not required to participate in any mobility-related signaling. Contrary to the latter approach, the MN's performance is not degraded by processing signaling and mobility protocol management.

2.1.4 Mobility Layer

Defines the responsibility of each layer in the mobility management process.

Link Layer

This class includes mobility protocols that use link layer information, when the point of attachment changes, to provide mobility management while the node preserves its network-layer (L3) address. This can fulfill some of the attributes of a micromobility protocol.

Layer 2½

This class uses MPLS to provide mobility management and signaling. MPLS (Rosen et al, 2001) is a technology that substitutes conventional packet forwarding within a network, or part of a network, with a fast operation of label lookup and switching. Each MPLS packet has a label. Label swapping is done by associating labels with routes and using the label value in the packet forwarding process.

In an MPLS cloud, switches are called Label Switching Routers (LSRs) and a connection or tunnel between two endpoints is formed by the union of several LSRs along a route. It is called a Label Switch Path (LSP). When a packet enters into an MPLS cloud, the egress LSR classifies the packet accordingly to the rules defined in its Forwarding Equivalence Class (FEC) and each FEC has an association with a particular LSP.

Through this mapping (FEC - LSP), a label is assigned to the package, which only identifies the LSP to the downstream LSR in the LSP, so that they can continue this procedure until reach the egress (edge) LSR. In core LSRs, the procedure is simpler, since the packet reclassification is no longer required, but just forwarding it to the downstream LSR. Note that the label has only local significance. Before a packet leaves an MPLS domain, its MPLS label is removed (Ren et al, 2001).

Network Layer

All mobility management and signaling is carried out by L3 protocols, based or not in the Mobile IP protocol.

Transport Layer

Mobility on transport layer intends to maintain TCP (Transmission Control Protocol)'s end-to-end reliability and correctness semantics while allowing redirecting the endpoints of an existing transport session (e.g., a TCP connection or a series of UDP - User Datagram Protocol- packets) to arbitrary addresses (Maltz & Bhagwat, 1998).

Application Layer

Mobility provided by application layer protocols intends to allow communication end systems to support mobility, heterogeneity, and multihoming. Terminal mobility also allows a device to move between IP subnets, while continuing to be reachable for incoming requests and maintaining sessions across subnet changes (Schulzrinne & Wedlud, 2000). Session mobility also allows a user to maintain a media session even while changing terminals. For example, a user may want to continue a session initiated on a MN on the desktop PC when entering his/her office. IPv4 or IPv6 mobility does not directly support such session mobility (Nasir & Mah-Rukh, 2006).

3. Mobile IP

The Mobile IP (MIP) (Perkins, 1997; Johnson et al, 2004) uses a stable IP address assigned to mobile nodes. This home address is used to allow the MN to be reachable by having a stable entry in the DNS service, and to hide the IP layer mobility from upper layers. A consequence of keeping a stable address independently of the mobile node's location is that all correspondent nodes try to reach the MN at that address, without knowing the actual location of the mobile node. Therefore, if there are packets forwarded to the home address, and the MN is not at its home network, its home agent is responsible for tunneling packets to the MN's new location.

MIPv4 (Mobile IP for IPv4 networks) solves the mobility problem by allowing the MN to use a second IP address: the CoA. This address changes at each new point of attachment and it indicates the network prefix, identifying the MN's point of attachment with respect to the network topology. The CoA is composed of a valid prefix in a foreign network. Thus, the MN will have a home address and one or more CoAs when moving between networks.

MIPv4 works by the cooperation of three separable mechanisms (Perkins, 1998): discovering the CoA, registering the CoA and tunneling to the CoA. The operation of Mobile IP protocol can be briefly described by the following steps (Figure 2):

1. The mobility agents (HA and FA) announces their presence through messages called Agent Advertisement (optionally, these messages can be requested by mobile agents through messages called Agent Solicitation);
2. A MN receives these messages and determines whether it is on its home network or on a foreign network;
3. When a MN detects it moved to a foreign network, it obtains a CoA in that network. The CoA can be allocated by the foreign agent or some other address configuration mechanism, such as DHCP (Dynamic Host Configuration Protocol);
4. When the MN is operating in the new network, it needs to register its CoA with its HA, through the exchange of Registration Request and Registration Reply messages;

5. Datagrams sent to the MN's home address by a CN are intercepted by the local HA and tunneled to the MN's CoA. The datagram is received at the exit of the tunnel, and finally delivered to the mobile node in the new network;
6. Datagrams sent by the MN are generally delivered to the destination using standard routing mechanisms, not necessarily through the HA.

The cooperation between MN, HA and CN is called triangular routing, as we can see in Figure 2, which summarizes the MIPv4 operation.

The triangular routing generates a processing overhead on HA, in addition to this being a single point of failure in the network. The MIPv6 solves this problem by optimizing the route.

Mobile IPv6 (Johnson, 2004) is intended to provide mobility support in IPv6 networks. In order to know where the MN is found, an association between home address and care-of address should be performed (binding). This combination of CoA is made by the MN and the HA. This association is achieved by a binding registration where the MN sends messages called Binding Updates (BU) to HA, which responds with a message Binding Acknowledgment (BA) (Figure 3).

The correspondent nodes may carry out route optimization, or they can store bindings between MN's home address and CoA. Thus, a MN can supply information about its location to the corresponding nodes, through the Correspondent Binding Procedure, which is a mechanism for authorizing the establishment of binding, called the return routability procedure.

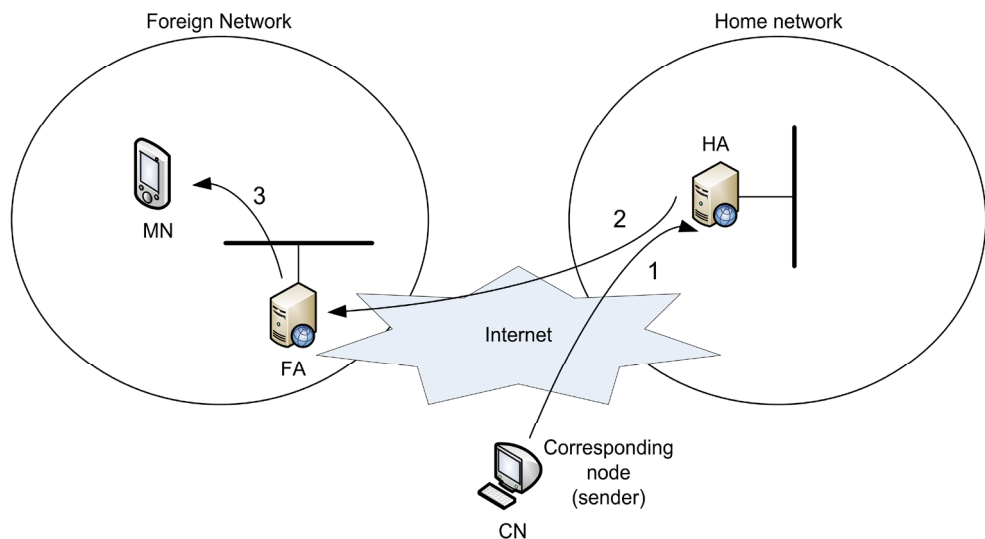


Fig. 2. MIPv4 Operation

Using the Route Optimization process, the CN must support MIPv6 and the MN must register with the CN. In this case, the CN, before sending a package, looks for a cached association between MN's HA and CoA. If there is an association, the package will be

routed to the CoA of mobile node directly. This eliminates congestion at the home link and the HA.

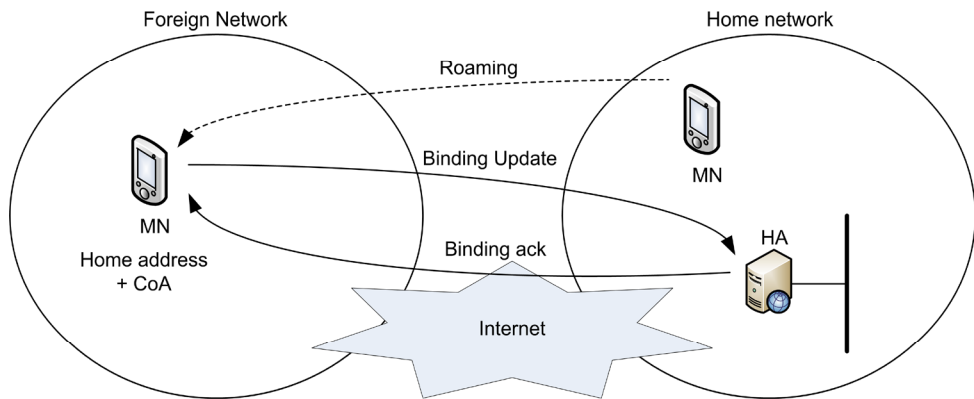


Fig. 3. MIPv6 Operation

MIPv4 and MIPv6 only define means of managing macromobility but do not address micromobility separately. Indeed, it uses the same mechanism in both cases. So, this protocol is not suited for micro mobility management, because of its high signaling load and long handover delay (Habaebi, 2006), namely movement detection, new CoA configuration, and Binding Update, is often unacceptable to real-time traffic such as Voice over IP (Koodli, 2008).

4. Link Layer related Micro-Mobility

FMIP - The Mobile IPv6 Fast Handovers (FMIPv6) protocol (Koodli, 2008) aims to reduce MN movement detection latency and new MN's CoA (Care-of Address) configuration latency by providing information to the MN when it is still connected to its current subnet. After discovering available access points, the MN requests subnet information from these APs: prefix, IP address, and L2 address of their associated routers. If the MN eventually attaches to one of the APs, the movement detection delay is reduced because the MN doesn't need to perform router discovery.

The MN formulates a new CoA (NCoA) based on the prefix of the new subnet and sends a message to its current access router, Previous Access Router (PAR), which communicates with the New Access Router (NAR) to determine whether the NCoA is unique. The PAR also establishes a tunnel to redirect packets arriving for PCoA (Previous CoA) to NCoA.

After performing a handover the MN announces its attachment immediately with an Unsolicited Neighbor Advertisement message (Narten et al, 2007) to circumvent the delay associated to neighbor's address resolution. FMIPv6 also defines a different behavior when the MN doesn't receive acknowledge message prior to its handover. There is also an adaptation of FMIPv6 to IPv4 networks (Koodli & Perkins, 2007).

IP-IAPP - The IP-IAPP proposal (Samprakou et al, 2004) extends 802.11f IAPP (IEEE, 2003) to support inter-network handover via L2 specific methods. IP-IAPP defines the Home

Access Point (HAP) that is the AP to which the MN was last associated inside its home network, similar to the Home Agent in MIP.

When the MN moves to a different network, it sends a modified IEEE 802.11 Reassociation.Request (IEEE, 1997) to an AP, the Foreign Access Point (FAP), informing IP addresses of HAP, MN, and Previous FAP (PAP) and this message triggers the mobility management procedure. The FAP communicates with the HAP to establish a bi-directional HAP-FAP tunnel and the HAP starts mapping the MN IP address to the FAP IP address, the Foreign Agent Care of Address (FACOA).

When the MN reassociates with a new FAP (NAP), the same procedure is performed with the addition of a communication between the PAP e NAP to establish a temporary unidirectional tunnel between them. The proposal has also been improved with the provision of more advanced services: secure inter-AP IP-IAPP communications, zero patching on the clients software, and support of clients which use a dynamic IP address (Samprakou et al, 2007).

The IEEE 802 Executive Committee approved IAPP withdrawal in 2006, because “the trial use period of 802.11F has expired, there has been no significant deployment of 802.11F implementations and, the functionality provided by 802.11F is being addressed in other standards fora”(IEEE P802.11, 2005).

5. Mobility with MPLS

All of the architectures discussed in this section consider that the MPLS cloud is surrounded by the Internet cloud and that micromobility is to be applied in MPLS cloud while MIP is to be applied in Internet cloud.

Mobile MPLS – The Mobile MPLS is a macro mobility protocol that borrows the mechanisms defined in the MIP standard and applies it to MPLS networks, so that the IP-in-IP tunnels are substituted by MPLS tunnels. The main objective in this migration is to improve the delay time in the tunneling packets from the HA to the FA. Another objective is to facilitate the use of QoS services that are native to MPLS networks (Ren et al, 2001).

In order to track the MN location, an entry in the LIB (Label Information Base) table at the HA is created for each MN that is registered with it. When the MN arrives at a foreign network, it registers itself with this FA and obtains a CoA from it. The FA sends this information to the MN's HA and it establishes a new LSP for that FA.

After the completion of the LSP connection, the HA changes the LIB entry for that MN to reflect the out label and port interface gathered from the previously created LSP. In doing that, whatever packet that is sent to the MN's home network will be tunneled to this LSP, arriving at the FA in which the MN is actually connected. A lack of an entry about out label and port interface for the MN at the LIB table at the MN's HA means that the MN returned to its home network.

Three scenarios were discussed. The first one considered was that both FAs and HAs were inside the same administrative MPLS domain. The second one considered was that HAs and FAs were inside different administrative MPLS domains. In order to establish a tunnel between them, Mobile MPLS suggests the use of a border protocol such as BGP (Border Gateway Protocol) (Rekhter & Rosen, 2001). The third scenario considered was that HA and FA were inside a different network tunneling technology, such as MPLS and IP clouds. LER

is responsible for de-tunneling packets from the MPLS cloud and re-tunneling it inside the IP cloud.

H-MPLS - Hierarchical Mobile MPLS (Yang & Makrakis, 2001) extends Mobile MPLS, which is a macro mobility protocol, in order to introduce into it micro mobility features. The main objective is to reduce the signaling overhead in creating a LSP from HA to FA, due to MN frequent handover in a small-size cells wireless environment.

To do that, H-MPLS introduces a new element, called FDA (Foreign Domain Agent) whose function is between that defined for HA and FA, as described in the MIP standards. The role of FDA is to track MN local mobility inside a MPLS domain. There is only one FDA per MPLS domain and many FA per subnetworks inside this domain.

The dynamics of the protocol is as follows: whenever a MN enters a foreign MPLS domain and it is its first registration, it acquires a CoA from its FA LSR and registers with it. This FA sends a Registration Request to its FDA, which has an equivalent function of a FA, but its scope is for a domain. This FDA will send back a Label Request message to FA and put as its FEC, the MN's CoA. At meanwhile, FDA will send a Registration Request message to HA, in the same way that was described in Mobile MPLS, but putting its IP address as a FEC for this LSP. So, at this point, there will be two LSPs, one from HA to FDA and another one from FDA to FA.

Now, if a MN does a handover, but stays in the same domain as the previous FDA, only the LSP from the new FA to FDA needs to be established. To avoid FDA sending packets to MN via the old FA, due to an out-of-date entry cache, the new FA sends a Binding Update message to old FA instructing it to create a LSP to the new FA in order to tunnel packets arriving at the MN's old location.

LEMA - Label Edge Mobility Agent (Chiussi et al, 2002) is a tunnel-based micro mobility architecture that uses MPLS as a network transport technology. This network is composed of an overlay network whose nodes are called LEMA, an LER that has its function augmented with LEMA features. This overlay network tracks MN location by building a set of LEMAs nodes from the highest to the lowest LEMAs that compound a path.

Highest LEMAs are the ingress node which registers its address in the HA database, and acts as FA in the MIP protocol; while the lowest LEMA is the access router, which remove the MPLS tunnel and delivers messages to the AP which the MN is connected to. This kind of scheme makes a hierarchical network, where only the LEMAs that compose a path to track the MN need to be aware of it.

Others features that can be attributed to it are fast handover capability, scalable design, QoS capability and gradual deployment. It is the MN's role to define the set of LEMAs that compose its path inside a LEMA network. This path is chosen based on a set of parameters, such as: available bandwidth, mobility patterns, and so on. The algorithm employed to choose a particular set is an open issue, and could be of high complexity. Finally, all LEMAs are connected to themselves by pre-established LSPs.

MM-MPLS - Micro Mobile MPLS (Langar et al, 2004) extends mobile MPLS with the principles employed by MIP-RR (L3 protocol) to support micro mobility on MPLS networks. It introduces a new component, called Label Edge Router/Gateway (LER/GW) that resides between HA and FA agents, as defined in the MIP protocol. It acts as a foreign domain agent to HA and it is this address that the MN must register at HA database when it first gets into the domain at which LER/GW is administrating.

So, there is a tunnel/LSP from HA to LER/GW, and an LSP from LER/GW to FA, with MN CoA's address as a FEC. FA here is an AR (Access Router) that remove the MPLS tunnel and delivers the packet to MN that is registered on it. Whenever MN moves to another FA, which is under the same LER/GW, only a regional registration is required that will create a new LSP from LER/GW to the new FA, using the new MN CoA's address as FEC. MM-MPLS uses LDP (Label Distribution Protocol) as signaling protocol in MPLS cloud.

I-LIB - Intermediate Label Information Base (Fowler & Zeadally, 2006) maintains the same idea of MM-MPLS architecture in general, where a FDA is placed between HA and FA. FDA has the same role as described early, i.e., its CoA address is registered at HA database and a tunnel connect HA to FDA. On the other side, FDA keeps track of MN by establishing a LSP to FA that is directly connected to it. Whenever a MN does a handover and establishes a new connection to a new FA, this new FA will try to establish a LSP to FDA, sending a Registration Request.

Here is where this proposal differs from the previous ones. Instead of establishing a new LSP from scratch, linking the new FA to the FDA, any segments that are common between the new path and the old path will be preserved. As such, any LSR that already has an entry in its LIB could preserve it and just update it to show the new configuration (the new segment that connects the MN). In order to do that, a new LIB is proposed which augmented the old ones with new fields to contemplate mobility issues. Among the fields that are required, the previous and new MN's CoAs must be accounted for. It is necessary to modify the packet since the FDA and HA know the MN by its old CoA, while the FA knows the MN by its new CoA.

6. Network Layer

Cellular IP - The CIP (Valko, 1999) architecture is composed of different wireless access networks (CIP access networks) connected to the Internet through a gateway. MIP manages mobility between these CIP access networks, while Cellular IP handles mobility within one domain. The IP address of the gateway is used as the MIP CoA. Thus, packets are first routed to the host's HA and then tunneled to the gateway, which "detunnels" packets and forwards them toward base stations, using host-specific routing path.

Base station (BS) components serve as wireless access points and also route IP packets, but IP routing is replaced by Cellular IP routing and location management. Base stations cache the path taken by uplink packets from MN to gateway for a period of time and use the reverse path to route downlink packets. In order to route packets to idle MNs, Cellular IP employs paging.

HAWAII - HAWAII divides the network into a hierarchy of domains. All issues related to mobility management within one domain are handled by a gateway called a domain root router, which uses a specialized path setup scheme which installs host-based forwarding entries in specific routers to support intra-domain micromobility (Ramjee et al, 1999). While moving inside its home domain, the MN maintains its stable IP address.

MIP mechanisms are used when the MN moves into a foreign domain. However, if the foreign domain is also based on HAWAII, then the MN is assigned a co-located CoA from its foreign domain to which packets for the MN are tunneled. The domain root router routes the packets to the MN using the host-based routing entries. When the MN moves between different subnets of the same domain, only the route from the domain root router to the BS

serving the MN is modified, and the remaining path remains the same, and connectivity is maintained using dynamically established paths.

The protocol contains three different messages for establishing, updating and refreshing host specific routes for the MN in the domain root router and any intermediate routers on the path towards the mobile host. The protocol also has four different path setup schemes, aiming to reduce disruption to the user traffic during a handoff, classified into two types based on the way packets are delivered to MNs. In the first type, packets are forwarded from the old base station to the new and, in the second type, they are diverted at the crossover router.

MIP-RR - MIP-Regional Registration (Fogelstroem et al, 2007) is an optional extension to the Mobile IPv4 protocol, and proposes a mean for mobile nodes to register locally within a visited domain. By registering locally, the number of signaling messages to the home network is kept to a minimum, and the signaling delay is reduced. This protocol introduces a new network node called the Gateway Foreign Agent (GFA). Besides the regular MIP Registration messages, a new pair of registration messages, Regional Registration Requests/Replies, is used between MNs/FAs/GFAs.

There are two models of how the MN uses Regional Registration. In the first model, the FAs in a visited domain advertise the address of the GFA, and, when a mobile node first arrives at this visited domain, it performs a home registration. At this registration, the mobile node registers the address of the GFA as its CoA with its HA. When moving between different foreign agents within the same visited domain, the mobile node only needs to make a regional registration to the GFA. In the second model, the FA can indicate that dynamic assignment of GFA is to be used, if being the FA's responsibility to choose the GFA after receiving a Registration Request from the MN.

PMIP - The Proxy MIP (Gundavelli et al, 2008) is a network-centric micromobility approach that relies on tunnels inside a domain to direct traffic to mobile nodes. PMIPv6 reuses many concepts of MIPv6, like the HA functionality, and defines two new entities, the Mobile Access Gateway (MAG) and the Local Mobility Anchor (LMA).

A MAG typically runs on the Access Router. It is responsible for detecting the mobile node attachments, and, if security policies are fulfilled, establishes tunnels to the LMA for directing traffic to the mobile nodes reached via this MAG. A MAG also emulates (via Router Advertisements messages) the mobile node's home network in such a way that the mobile node may change the default router in a handover, but preserves the remaining L3 parameters. As the mobile node moves inside the domain, tunnels between MAG and LMA and routes on the LMA are updated.

LMA maintains a binding cache entry for each currently registered MN, providing reachability to the MN's address. When the LMA receives a packet targeted to the mobile node it forwards the packet via the tunnel ending on the MAG to where the node is attached. PMIPv6 employs local binding update messages between MAG and LMA for signaling purposes and only a single hierarchy of tunnels. Indeed, Proxy MIP considers IPv4 support, but this requires an extension to the original protocol, since it uses some IPv6 features such as auto-configuration and extension headers.

HMIPv6 - To support local mobility, Hierarchical Mobile IPv6 (HMIPv6) extends Mobile IPv6 and IPv6 Neighbor Discovery (Narten et al, 2007) and introduces Mobility Anchor Point (MAP), a new Mobile IPv6 node. A mobile node entering an HMIP domain receives

Router Advertisements containing information about one or more local MAPs and configures two CoAs: an on-link CoA (LCoA) and a Regional Care-of Address (RCoA). The LCoA is configured on a mobile node's interface based on the prefix advertised by its default router. It is a standard Mobile IP CoA and has a different name just to be distinguished from RCoA. The RCoA is configured on the MAP's link and is obtained by the MN from the MAP employing the address mechanisms described by RFC 4877 (Devarapalli & Dupon, 2007).

After configuration, the MN sends two binding update (BU) messages. The first is a local BU to the MAP to bind the MN's RCoA to its LCoA and to establish a bi-directional tunnel between them. The second is a BU to the home agent to bind the MN's home address to its RCoA. The MAP receives all packets on behalf of the mobile node it is serving and encapsulates and forwards them directly to the mobile node's LCoA.

When the mobile node moves within the same MAP domain, it only needs to register its new LCoA with its MAP, limiting the amount of Mobile IPv6 signaling outside the local domain. The RCoA remains unchanged and the home agent (HA) or the correspondent nodes (CNs) are not aware of the change in LCoA.

DMA - The Dynamic Mobility Agent (Misra et al, 2001) architecture uses the Intra-Domain Mobility Management Protocol (IDMP) (Das et al, 2002) to manage intradomain mobility. The architecture defines two entities to achieve mobility support: Mobility Agent (MA) that acts as a domain-wide point for packet redirection, and the Subnet Agent (SA) that provides subnet-specific mobility services. Two CoAs are associated with a MN: Global CoA (GCoA) and Local CoA (LCoA).

The GCoA is the address used by macromobility protocols to redirect packets and remains unchanged as long as the MN stays in the current domain. The LCoA is an address from the subnet the MN is attached to. IDMP is used in the communication between the MN and the SA, and between the MN and the MA.

When the MN first arrives at the domain, it obtains an LCoA and the SA assigns the MN a MA. The MN registers its LCoA with the MA and obtains a GCoA. After the macromobility updates process is performed by the MN, the packets from remote hosts, tunneled or directly transmitted to the GCoA, are intercepted by the MA and tunneled to the MN's LCoA. When the MN moves to new subnet it obtains a new LCoA and informs its MA of the new LCoA, updating the GCoA-LCoA mapping.

7. Transport Layer Mobility

MSOCKS is a split-connection proxy-based architecture that uses TCP Splice technique to achieve the same end-to-end semantics as normal TCP connections (Maltz & Bhagwat, 1998). A special host, called a proxy, is placed in the communication path between a mobile node and a correspondent node. An end-to-end TCP connection between a mobile node and a correspondent node are split into two separated connections: one connection between the mobile node and proxy and another between the proxy and the correspondent node. The MSOCKS protocol extends the SOCKS protocol (Leech et al, 1996) to redirect TCP streams to a mobile node's changing location.

When the mobile node changes the address of its network interface it opens a new connection to the proxy and sends an MSOCK message specifying the connection identifier of the original connection. The proxy unsplices the old mobile-node-to-proxy connection

from the proxy-to-correspondent-node connection, and splices in the new mobile-node-to-proxy connection. Only the proxy is aware of the mobile node migration and the communication between the proxy and the correspondent node remains unchanged. This technique also allows the mobile node to change the network interface used to communicate with the proxy.

TCP Migrate – This mobility architecture (Snoeren & Balakrishnan, 2000) (Snoeren et al, 2002) allows an application running on mobile hosts to support transparent connectivity across network address changes. As MNs change their network attachment point, new addresses can be assigned through DHCP, manually or using an auto-configuration protocol. To locate mobile hosts in the new network, Domain Name System (DNS) is used and its ability to support secure dynamic updates.

Because most Internet applications resolve hostnames to an IP address at the beginning of a transaction or connection, this approach is viable for initiating new sessions with mobile hosts. When a host changes its network attachment point (IP address), it sends a secure DNS update to one of the name servers in its home domain updating its current location. The name-to-address mappings for these hosts are un-cacheable by other domains, so stale bindings are eliminated.

Nevertheless, when a MN moves during a previously established connection, it may suspend the open connection and reactivate it from the new address, sending a special packet (Migrate SYN) to the correspondent node, which carries a token that identifies the previous connection. This SYN packet signals the correspondent node to re-synchronize the connection with the MN at the new point of attachment (new address). Thus, it is possible to provide mobility support as an end-to-end service, according to the application's specific requirements, without changes in the network layer.

8. Application Layer Mobility

Mobility using SIP - The SIP (Session Initiation Protocol) is a signaling protocol, widely used for setting up and tearing down multimedia communication sessions over the Internet. It can be used in any application where session initiation is necessary.

The SIP registration mechanism is considered the application-layer equivalent of the MIP registration mechanism. However, while mobile IP binds a permanent IP address identifying a host to a temporary CoA, SIP binds a user-level identifier to a temporary IP address or host name (Schulzrinne & Wedlund, 2000). An INVITE message is sent by a MN to its CN to set up a communication session. The mechanism to provide MN mobility during an active session foresees that the MN needs to send another INVITE message to the CN to communicate the information about the new parameters of the communication session after the handover, using the same call identifier as the original call setup.

This solution has some drawbacks (Salsano et al, 2008). The second INVITE is sent end-to-end, and this could lead to high delays. Moreover, the handover procedure relies on the capability of the CN to handle this procedure, thus increasing MN processing needs. An auxiliary mechanism is necessary if the MN and CN move at the same time.

9. Mobility Plane Architecture

The Mobility Plane Architecture (MPA) (Zagari et al, 2008) is an instance of a reference architecture for micromobility support in IP networks (Prado et al, 2008). The goal of MPA is to speed up the handover process in order to minimize communication disruptions when the mobile node changes its network point of attachment. One of the requirements of this architecture is to place the burden demanded by micromobility on the network, not on the mobile nodes. Another requirement is to use, ideally, only well established network protocols. The key point of MPA is to employ an overlay network built above a transport network for directing traffic to the mobile nodes. This overlay network is composed of network elements called Mobility Aware Router (MAR), which are routers enhanced with MPA's functionalities. MPA employs point-to-multipoint (P2MP) tunnels in order to encapsulate traffic directed to the mobile nodes and allows a gradual deployment once the architecture elements are installed only at the MARs.

MPA addresses the following issues related to mobility in IP networks: tunnel management (tunnel establishment, shutdown, and topology updating); secure mobile node attachment and handover; tracking of mobile node actual point of attachment (location); routing on the overlay network (decoupled from routing on the transport network); and quality and class of service (QoS/CoS) offered to the mobile nodes.

9.1 Functional Description

The architecture defines the following basic elements:

- Transport network - an IP network from which the network operator wishes to offer mobility services.
- Point-to-multipoint (P2MP) tunnel - a tunnel with a topology forming a tree. Nodes in the tree are MARs and arcs are tunnel segments connecting MARs. The tunnel has a single ingress (root) MAR, branch MARs (nodes with branching level greater than one), and egress MARs (leaves of the tree). A packet being forwarded through the tunnel may or may not be replicated at a branch MARs according to the policies enforced by these MARs.
- Access router - a router (usually an egress MAR) connected to a wireless access point.
- Access network - an IP subnetwork formed by the access routers and access points.
- Overlay mobile network - logical network built with one or more P2MP tunnels established through the transport network.

Figure 4 illustrates these basic elements. In addition to the basic elements, four functional blocks (FB) are defined:

- Tunnel Management (TM) Functional Block: TM is the entity responsible for P2MP tunnel establishment, shutdown, and re-routing. It must provide interfaces to the network management system and to the human operator. Tunnel management is a function carried out by MARs.
- Mobile Routing (MR) Functional Block: MR is the entity responsible for tracking the mobile nodes actual point of attachment and for interacting with the MARs forwarding engine in order to route traffic to the mobile nodes correct location. Mobile routing is a function carried out by MARs.
- Address Configuration (AC) Functional Block: AC is the entity responsible for supplying L3 addresses to the mobile nodes when they connects or reconnects to

the access network. Address configuration is a function carried out cooperatively by MARs and mobile nodes.

- **Handover Helper (HH) Functional Block:** HH is the entity responsible for facilitating the handover process with functions including L2 notification (triggering), L2 re-association, secure node attachment, and handover-related signaling. This function can be spread among MARs, network equipments (e.g., wireless switches), and mobile nodes.

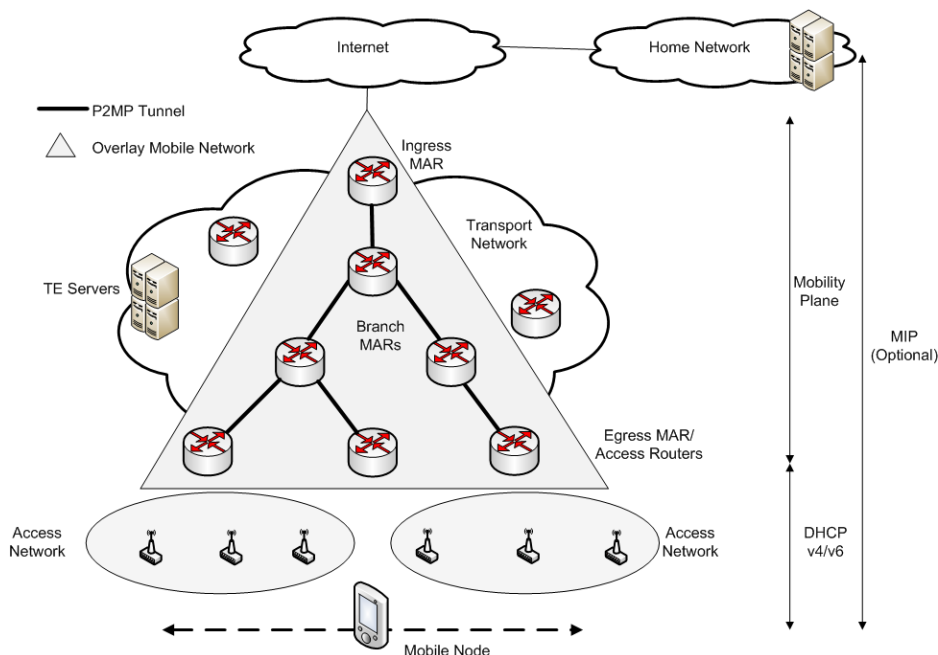


Fig. 4. MPA overview

9.2 Operations Basics

When a packet targeted to a mobile node reaches an ingress MAR, it is tunneled until it reaches an egress MAR able to route the packet to the mobile node. When the mobile node performs a handover, the AC FB presented on the mobile node and on the network interact in order to provide the mobile node with a new L3 address. If the address is identical to the previous one the transport connections are not broken due the handover. The handover also triggers a mobile node location update on the MR FB. The location updating process updates the mobility routing tables on the MARs in such a way that when a packet is targeted to the mobile node the packet is routed to the egress MAR serving the link the mobile node is attached to.

Let us consider the mobile routing and the location update processes. Figure 5 shows a mobile node attached via access router M4. When the mobile node moves to a link served by M5, the entry related to this node on the mobile routing table at M2 must be updated with a different tunnel segment (in this case from segment C to D). If the mobile node roams to a link served by M7, the mobile routing table at M1 and M3 must be updated. Table

updates are performed as soon as the mobile routing protocol messages indicating the new point of attachment are processed by the branch MARs.

The entries on the mobile routing table are soft state, meaning that the entries are dropped if location update messages confirming them cease. Soft state is a clear way to drop routes to mobile nodes when they no longer are reached through these routes. This scheme demands that a mobile node perform network attachments periodically in order to generate location update messages that will refresh the routes to it. A way to force the mobile nodes to perform periodic attachments is to provide them L3 address with a short lease time. When the lease time is close to expire, a mobile node performs address renewal that will trigger address location update messages in its behalf.

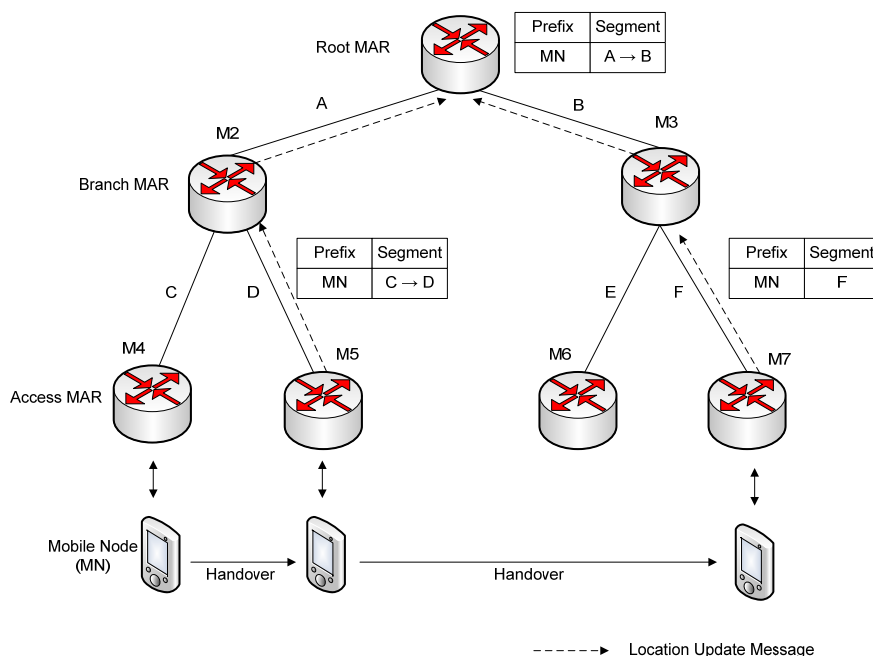


Fig. 5. MN Routing in MPA

9.3 MPA advantages

The MPA architecture presents the following advantages:

1. The solution is not limited to IPv6, being deployable on both IPv4 and IPv6 networks. Since it relies on tunneling, mixed deployments with IPv4 on the access network and IPv6 on the transport network, and vice-versa are possible.
2. The solution is not affected by middle boxes such as firewall and NAT boxes placed anywhere on the access, transport, or backbone networks.
3. The solution demands no complex protocols such as MIPv6 on the mobile nodes. Since it relies only on the standard IP protocol stack, the solution supports all the commercially available mobile nodes based on, for instance, Windows Mobile, Symbian, and PalmOne operating systems. The architecture does not forbid

enhancements deployed on the mobile nodes in order to improve handover speed and security, for instance. Such enhancements can be installed in user space or in the operating system kernel (e.g., as device drivers).

4. The solution preserves the L3 address of the mobile nodes when they roam inside the access network, causing no disruption of transport connections maintained by the mobile nodes.
5. The solution does not restrict the mobile node to employ macromobility protocols such as MIPv6.
6. The solution complies with security mechanisms related to L2 (e.g. WPA), L3 (e.g. IPSec), and L4+ (e.g., SSL, HTTPS).
7. The solution combines P2MP tunneling management, QoS/CoS management, and mobile node location tracking into the same protocol (RSVP-TE), reducing implementation and operating costs.
8. The solution does not interfere on services already deployed on the transport network such as VPN and VoIP.
9. Only well standardized protocols are employed by the architecture. When extensions to protocols are necessary they are introduced as opaque objects already foreseen by these protocols.

For more details on MPA description, implementation (using IPv4 and IPv6) and performance analysis, see Prado (2008), Zagari (2008), Johnson (2008), and Zagari (2009). Badan et al (2009) details the MPA implementation using MPLS.

10. General Analysis and Research Directions

Table 1 summarizes all the protocols and solutions seen in this chapter.

Protocol	Mobility Range	Mobility Routing	Mobility Signaling	Mobility Layer
CIP	Micro	Routing	MN-centric	L3
DMA	Micro	Tunnel	MN-centric	L3
FMIP	Micro	Tunnel	MN-centric	L2
Hawaii	Micro	Routing/Tunnel	MN-centric	L3
HMIP	Micro	Tunnel	MN-centric	L3
H-MPLS	Micro	Tunnel	MN-centric	L2½
I-LIB	Micro	Tunnel	MN-centric	L2½
IP-IAPP	Macro/micro	Tunnel	MN-centric	L2
LEMA	Micro	Tunnel	MN-centric	L2½
MPA	Micro	Tunnel	Network-centric	L3
MIP	Macro	Routing/Tunnel	MN-centric	L3
MIP-RR	Micro	Tunnel	MN-centric	L3
MM-MPLS	Micro	Tunnel	MN-centric	L2½
Mobile MPLS	Macro	Tunnel	MN-centric	L2½
MSOCKS	Micro	Routing	MN-centric	L4
PMIP	Micro	Tunnel	Network-centric	L3
SIP	Macro	Routing	MN-centric	L5
TCP Migrate	Macro	Routing	MN-centric	L4

Table 1. General Classification

As this work is a non-exhaustive selection of mobility solutions, our comments are limited to these protocols revised in this chapter.

There are a majority of MN-centric solutions, since the beginning of mobility management research, with the Mobile IP. Newer solutions, such as MPA and PMIP, work as network-centric solutions, and it is our belief that this class of mobility signaling will have more research focus and implementations, because of its advantages to the final user: they do not need protocol installation of configuration on the MN, no overheads on devices to handle mobility, and no intervention from the final user to adopt a particular mobility solution.

Another interesting issue is the network layer protocol. About 10 years ago nobody believed in IPv4 for mobility, which is why both MIPv6 and its extensions appeared at that time. Since then, however, the IPv6 protocol has not been massively adopted, so there is reawakened interest in IPv4 (PMIP and MIPv4 raised again, for example). Now, 4G researchers and professionals say IPv6 is inevitable because each cell phone and mobile devices must have a fixed (stable) IP address. Solutions compatible only to IPv4 must foresee this IPv6 adoption and implement a compatible version of their protocol, if intended to be used in the next future.

Not mentioned in this chapter, solutions for mobility must adopt security mechanisms for authentication, authorization and general security procedures. For example, for the MPA architecture, the access points can be configured to authenticate mobile nodes based on WPA2 employing Pre-Shared Keys (PSK) or RADIUS. PSK is easy to configure but is not as secure as RADIUS-based authentication. RADIUS authentication can be strengthened by using certificates installed on the mobile nodes. As RADIUS transactions take long time (500ms in our testbed network), RADIUS-based authentication increases considerably the handover overhead.

In order to speed up RADIUS-based authentication, a cache mechanism can be employed such as PMK (Pairwise Master Key) caching (also called proactive key caching). In this mechanism, once a mobile node completes successfully a RADIUS transaction, the access point stores the PMK supplied by the RADIUS server in the cache. When the mobile node connects to a new access point, the access point queries the cache (using the mobile node's MAC address as a search key) in order to recover the PMK assigned to the node. If an entry is found, the access point accepts the mobile node without the need of a RADIUS transaction. In this case, the PMK found on cache is used to secure the communication between the mobile node and the access point.

As suggestions for future research directions in mobility management, we can point out the development of architectures able to:

- integrate macro and micro mobility into a single mobility solution;
- support both vertical and horizontal handover (e.g., between WiFi and 4G networks);
- support clean slate solution, by designing the network from the mobility requirements (and not to incorporate mobility extensions over the existing networks).
- These solutions decouple host identity from network address as suggested by HIP (Host Identity Protocol) and other related solutions;
- restrict handover within the L2, employing, for instance, tunneling over Ethernet (instead of over IP or MPLS), flat routing (based on MAC - Media Access Control - address), etc.

11. Conclusion

This chapter was intended to present a major review on mobility architectures and protocols. Many solutions were shown, ranging from link layer related solutions to application layer solutions, including network, transport and intermediary layer protocols. We also presented MPA, which is our solution for mobility management, using a network centric paradigm.

Some of the reviewed solutions were abandoned, some were investigated till today. But mobility management solution still needs investigation, to allow massive deployment and use. Although the new mobility architectures and protocols are permanently under investigation, all new solutions are constrained by factors such as: be deployable over existing IPv4 (and future IPv6) fixed networks, operate without upgrading with the mobile nodes already in the marked, comply with current network operation practices, be scalable without degrade quality of service, and allow the introduction of new services with stringent communication requirements such as media streaming (e.g., IP TV), location, and entertainment services. The challenge in mobility for IP networks is to comply with these factors that, unfortunately, are not restricted only to the technical issued commonly addressed by the network architects.

12. References

- Akyildiz, I., Xie, J., & Mohanty, S. (2004). A Survey of Mobility Management in Next-Generation All-IP-Based Wireless Systems. *IEEE Wireless Communications*, 11 (4), 16–28.
- Badan, T., Zagari, E., Prado, R., Cardozo, E., Magalhaes, M., Carrilho, J., Pinto, R., Berenguel, A., Moraes, D., Johnson, T., & Westberg, L. (2009). A Network Architecture for Providing Micro-Mobility in MPLS/GMPLS Networks. *Proceedings of IEEE Wireless Communications and Networking Conference* (pp. 1-6).
- Campbell, A. T., & Gomez-Castellanos, J. (2000). IP Micro-Mobility Protocols. *ACM SIGMOBILE Mobile Computing and Communications Review*, 4 (4), 45 – 53.
- Campbell, A. T., Gomez, J., Kim, S., Wan, C.-Y., Turany, Z. R., & Valko, A. G. (2002). Comparison of IP Micromobility Protocols. *IEEE Wireless Communications*, 9 (1), 2–12.
- Chiussi, F. M., Khotimsky, D. A., & Krishnan, S. (2002). A Network Architecture for MPLS-Based Micro-Mobility, *Proceedings of the IEEE Wireless Communications and Networking Conference*, vol. 2, pp. 549-555.
- Chiussi, F., Khotimsky, D., & Krishnan, S. (2002). Mobility Management in Third-Generation All-IP Networks. *IEEE Communications Magazine*, 40 (9), 124-135.
- Das, S., Mcauley, A., Dutta, A., Misra, A., Chakraborty, K., & Das, S.K. (2002). IDMP: an intradomain mobility management protocol for next-generation wireless networks. *IEEE Wireless Communications*, 9 (3), 1536-1284.
- Devarapalli, V., & Dupon, F. (2007). Mobile IPv6 Operation with IKEv2 and the Revised IPsec Architecture. *Request For Comment (RFC) 4877*, Available at: www.faqs.org/rfcs/rfc4877.html.
- Fogelstroem, E., Jonsson, A., & Perkins, C. (2007). Mobile IPv4 Regional Registration. *Request For Comments (RFC) 4857*, Available at: www.ietf.org/rfc/rfc4857.txt.

- Fowler, S., & Zeadally, S. (2006). Fast Handover over Micro-MPLS-based Wireless Networks, *Proceedings of the 11th Symposium on Computers and Communications* (pp. 181-186).
- Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., & Patil, B. (2008). Proxy Mobile IPv6. *Request For Comment (RFC) 5213*, Available at: <http://tools.ietf.org/html/rfc5213>.
- Habaebi, M. H. (2006). Macro/Micro-Mobility Fast Handover in Hierarchical Mobile IPv6. *Computer Communications*, 29 (51), 611– 617.
- IEEE Document P802.11/D6.1.97/5 Wireless LAN, MAC and Physical Specifications, June 1997.
- IEEE Recommended Practice for Multi-Vendor Access Point Interoperability via an Inter-Access Point Protocol Across Distribution Systems Supporting IEEE 802.11TM Operation. Mar, 2003.
- IEEE P802.11 Working Group. (2005). Approved Minutes of the IEEE P802.11 Full Working Group. Available at: <https://mentor.ieee.org/802.11/dcn/05/11-05-1136-00-0000-minutes-working-group-nov-2005.doc>.
- Johnson, D., Perkins, C., & Arkko, J. (2004). Mobility Support in IPv6. *Request For Comments (RFC) 3775*, Available at: www.ietf.org/rfc/rfc3775.txt.
- Johnson, T., Zagari, E., Prado, R., Badan, T., Cardozo, E., & Westberg, L. (2008). Performance Analysis of a New Architecture for Mobility Support in IP Networks, *Proceedings of the International Wireless Communications and Mobile Computing Conference* (pp. 706-711).
- Kempf, E. (2007). Problem Statement for Network-Based Localized Mobility Management. *Request For Comments (RFC) 4830*, Available at <http://www.ietf.org/rfc/rfc4830.txt>.
- Koodli, R., & Perkins, C. (2007). Mobile IPv4 Fast Handovers. *Request For Comment (RFC) 4988*, Available at: www.faqs.org/rfcs/rfc4988.html.
- Koodli, R. (2008). Mobile IPv6 Fast Handovers. *Request For Comment (RFC) 5268*, Available at: www.faqs.org/rfcs/rfc5268.html.
- Langar, R., Le Grand, G., & Tohme, S. (2004). Micro Mobile MPLS Protocol in Next Generation Wireless Access Networks, *Proceedings of the IEEE Symposium on Computer and Communication* (pp. 14-17).
- Leech, M., Ganis, M., Lee, Y., Kuris, R., Koblas, D., & Jones, L. (1996). SOCKS Protocol Version 5. *Request For Comment (RFC) 1928*, Available at: www.faqs.org/rfcs/rfc1928.html.
- Maltz, D.A., & Bhagwat, P. (1998). Msocks: an Architecture for Transport Layer Mobility, *Proceedings of the Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies*, vol.3, pp. 1037-1045.
- Manner, J., & Kojo, M. (2004). Mobility Related Terminology. *Request For Comment (RFC) 3753*, Available at: www.faqs.org/rfcs/rfc3753.html.
- Misra, A., Das, S., McAuley, A. & Das, S.K. (2001). Autoconfiguration, Registration and Mobility Management for Pervasive Computing. *IEEE Personal Communications Magazine*, 8 (4), 24-31.
- Narten, T., Nordmark, E., Simpson, W., & Soliman, H. (2007). Neighbor Discovery for IP version 6 (IPv6). *Request For Comment (RFC) 4861*, Available at: www.faqs.org/rfcs/rfc4861.html.

- Nasir, A., & Mah-Rukh. (2006). Internet Mobility using SIP and MIP. In *Proceedings of the Third International Conference on Information Technology: New Generations* (pp 334 - 339).
- Perkins, C. (1997). Mobile IP. *IEEE Communications Magazine*, 35 (5), 84-99.
- Perkins, C. (1998). Mobile Networking through Mobile IP. *IEEE Internet Computing*, 2 (1), 58-69.
- Perkins, C. (2002). IP Mobility Support for IPv4. *Request for Comments (RFC) 3344*, Available at: www.ietf.org/rfc/rfc3344.txt.
- Prado, R., Zagari, E., Cardozo, E., Magalhaes, M., Badan, T., Carrilho, J., Pinto, R., Berenguel, A., Barboza, D., Moraes, D., Johnson, T., & Westberg, L. (2008). A Reference Architecture for Micro-Mobility Support in IP Networks, *Proceedings of the Thirteenth IEEE Symposium on Computers and Communications* (pp. 624 - 630).
- Rekhter, Y., & Rosen, E. (2001). Carrying Label Information in BGP-4. *Request for Comment (RFC 3107)*, Available at: <http://www.ietf.org/rfc/rfc3107.txt>.
- Ramjee, R., Porta, T. L., Thuel, S., Varadhan, K., & Wang S. (1999). Hawaii: A Domain-Based Approach for Supporting Mobility in Wide-Area Wireless Network, *Proceedings of the IEEE International Conference on Network Protocols* (pp. 283-292).
- Ren, Z., Tham, C.-K., Foo, C.-C., & Ko, C.-C. (2001). Integration of Mobile IP and Multi-Protocol Label Switching. *Proceedings of the IEEE International Conference on Communications* (pp. 2123-2127).
- Rosen, E., Viswanathan, A., & Callon, R. (2001). Multiprotocol Label Switching Architecture. *Request For Comments (RFC) 3031*, Available at: <http://www.ietf.org/rfc/rfc3031.txt>
- Saha, D., Mukherjee, A., Misra, I., Chakraborty, M., & Subhash, N. (2004). Mobility Support in IP: a Survey of Related Protocols. *IEEE Network*, 18 (6), 34-40.
- Salsano, S., Polidoro, A., Mingardi, C., Niccolini, S., & Veltri, L. (2008). Sip-based Mobility Management in Next Generation Networks. *IEEE Wireless Communications*, 15(2), 92-99.
- Samprakou, I., Bouras, C., & Karoubalis, T. (2004). Fast and Efficient IP Handover in IEEE 802.11 Wireless LANs. *Proceedings of the International Conference on Wireless Networks* (pp. 249-255).
- Samprakou, I., Bouras, C., & Karoubalis, T. (2007). Improvements on IP IAPP: A Fast IP Handover Protocol for IEEE 802.11 Wireless and Mobile Clients. *Wireless Network*, 13 (4), 497-510.
- Schulzrinne, H., & Wedlund, E. (2000). Application-layer Mobility Using SIP. *SIGMOBILE Mobile Computing and Communications Review*, 4 (3), 47-57.
- Snoeren, A. C., & Balakrishnan, H. (2000). An End-to-end Approach to Host Mobility, *Proceedings of the 6th ACM/IEEE International Conference on Mobile Computing and Networking* (pp. 155-166).
- Snoeren, A. C., Balakrishnan, H., & Kaashoek, M. F. (2002). The Migrate Approach to Internet Mobility, *Proceedings of the Student Oxygen Workshop* (pp. 14-17).
- Valko, A. G. (1999). Cellular IP: A New Approach to Internet Host Mobility. *SIGCOMM Computer Communication Review*, 29 (1), 50-65.
- Yang, T., & Makrakis, D. (2001). Hierarchical Mobile MPLS: Supporting Delay Sensitive Applications Over Wireless Internet. *Proceedings of the International Conferences on Info-tech and Info-net* (vol. 2, pp. 453-458).

- Zagari, E., Prado, R., Cardozo, E., Magalhaes, M., Badan, T., Carrilho, J., Pinto, R., Berenguel, A., Barboza, D., Moraes, D., Johnson, T., & Westberg, L. (2008). MPA: a Network-Centric Proposal for Micro-Mobility Support in IP Networks, *Proceedings of The 6th Annual Communication Networks and Services Research Conference* (pp. 609-616).
- Zagari, E., Prado, R., Badan, T., Cardozo, E., Magalhaes, M., Carrilho, J., Berenguel, A., Moraes, D., Dolphine, T., Johnson, T., & Westberg, L. (2009). Design and Implementation of a Network-Centric Micro-Mobility Architecture. *IEEE Wireless Communications and Networking Conference* (pp. 1-6).

Positioning in Indoor Mobile Systems

Miloš Borenović and Aleksandar Nešković
*School of Electrical Engineering, University of Belgrade
Serbia*

1. Introduction

At present times people travel far greater distances on daily bases than our not so distanced ancestors had travelled in their lifetimes. Technological revolution had brought human race in an excited state and steered it towards globalization. Nevertheless, the process of globalization is not all about new and faster means of transportation or about people covering superior distances. Immense amount of information, ubiquitous and easily accessible, formulate the essence of this process. Consequently, ways through which the information flows are getting too saturated for free usage so, for example, frequency spectrum had become a vital natural resource with a price tagged on its lease. However, the price of not having the information is usually much higher. By employing various wireless technologies we are trying to make the most efficient use of frequency spectrum. These new technologies have brought along the inherent habit of users to be able to exchange information regardless of their whereabouts. Higher uncertainty of the user's position has produced increase in the amount of information contained in its position. As a result, services built on the location awareness capabilities of the mobile devices and/or networks, usually referred to as Location Based Services (LBS, also referred to as LoCation Services – LCS), have been created. Example of services using the mobile location can be: location of emergency calls, mobile yellow pages, tracking and monitoring, location sensitive billing, commercials, etc. With the development of these services, more efforts are being pushed into producing the maximum of location-dependent information from a wireless technology. Simply, greater the amount of information available – more accurate the location^φ estimate is.

Whereas in outdoor environment the satellite-based positioning techniques, such as the Global Positioning System (GPS), have considerable advantages in terms of accuracy, the problem of position determination in an indoor environment is much farther from having a unique solution. Cellular-based, Computer vision, IrDA (Infrared Data Association), ultrasound, satellite-based (Indoor GPS) and RF (Radio Frequency) systems can be used to

^φ Sometimes, in literature, the words position and location have different meaning. Most often, position translates to the set of numerical values (such as geographical coordinates) which describe the user's placement, whereas the location usually refers to the descriptive information depicting the user's whereabouts (such as Picadilly Circus, London, UK). Nevertheless, this work treats both words interchangeably.

obtain the user's position indoors. Positioning technologies, specific for indoor environment, such as computer vision, IrDA and ultrasound require deployment of additional infrastructural elements. On the other hand, the performances of the satellite- and cellular-based positioning technologies are often unsatisfactory for typical LBSs in an indoor environment. Due to the proliferation of IEEE 802.11 clients and infrastructure networks, and the fact that a broad scope of LBSs can be brought into an existing WLAN network without the need for additional infrastructure, WLAN positioning techniques are relevant and established subjects to intensive research.

2. Performance Parameters and Approaches to Positioning

The determination of user's location can be seen as a simple mechanism consisting in calculating the whereabouts of the user. Those whereabouts could be descriptively expressed or in terms of geographic or some other coordinates. Nonetheless, it is practically impossible to obtain the exact location of a user, in 100% of the cases, regardless of the user and its environment (Collomb, 2002). Therefore, it is only an estimate of the user's location that can be obtained and, it is very important to know how proximate the actual location and location estimate are. To achieve that, it is necessary to characterise this location estimate. On the other hand, it is also significant to describe the positioning technique itself in terms of its practicality and viability. All this is generally done through a set of performance parameters: Accuracy (Distance Error, Uncertainty, Confidence, and Distance error's Cumulative Distribution Function), Coverage and Availability, Latency, Direction and Velocity, Scalability, Complexity and Cost effectiveness.

The first group of performance parameters is used to characterise the quality of a location estimate.

Accuracy – This is undoubtedly the most important performance parameter as it illustrates the essential characteristic of a positioning technique. This parameter enables to determine whether the calculated position is close to the exact position. This parameter is composite and consists of three different values that must be taken into account:

- Distance Error,
- Uncertainty, and
- Confidence.

The Distance Error corresponds to the difference between the exact location of the user (i.e. of his/her terminal) and the calculated position, obtained through a position determination method. It is also referred to as Location Error or Quadratic Error in terms of two-dimensional positioning. Distance Error is generally expressed in units of length, such as meters.

Determining the Distance Error can be very useful in depicting the particular position determination cases. However, in order to express the positioning capabilities of a technique it is usually much more suitable to exploit the Distance Error statistics via Uncertainty and Confidence parameters.

Bearing in mind that the calculated user's location is not the exact location but is biased by the Distance Error, it can be seen that the calculated position does not enable resolving the single point at which the user is located, rather an area. Depending on the positioning techniques used, this area may have different shapes (e.g. a circle, an ellipse, an annuli, etc). For that reason, the Uncertainty value represents the distance from the "centre" of this area

to the edge of the furthest boundary of this area. In other words, the Uncertainty value can be seen as the maximum potential Distance Error. The value of uncertainty is expressed with the same unit as for the Distance Error.

However, the Uncertainty value is not sufficient to describe the Accuracy of a positioning technique. The determination of the Uncertainty value goes through a statistical process and does not enable to guarantee that 100% of the calculated positions have a Distance Error lower than the Uncertainty value. That is the reason why the Uncertainty value is usually associated with a Confidence value, which expresses the degree of confidence that one can have into the position estimate. This degree of confidence is generally expressed in percentage or as a value of probability.

As a consequence, it is the combination of Uncertainty and Confidence that validly describe the accuracy of a positioning technique.

The other way of expressing the Accuracy, i.e. the performance or requirements associated to location determination, is through the Distance error's Cumulative Distribution Function (CDF). This approach is more comprehensive and inclusive due to the fact that a particular Uncertainty, Confidence pair can always be read of the graph for each and every Confidence or Uncertainty value. When assessing the technique's suitability for LBSs, expressing the Accuracy of a positioning technique through an Uncertainty, Confidence pair might be descriptive enough for a certain LBS. On the other hand, stating a positioning technique's CDF is more general and depicts the technique's accuracy for all potential LBSs.

Coverage and Availability – Accuracy is not the only parameter to be considered in order to characterise a location estimate. Coverage and Availability must be considered too. These two parameters are linked together:

- The Coverage area for a positioning method corresponds to the area in which the location service is potentially available, and
- The Availability expresses the percentage of time during which the location service is available in the coverage area and provides the required level of performance.

Latency – Location information makes sense only if it is obtained within a timeframe which remains acceptable for the provision of the LBSs. Latency represents the period of time between the position request and the provision of the location estimate and it is generally expressed in seconds.

Direction and Velocity – Although the herein presented work is restrained to the initial position determination algorithms, there are additional tracking algorithms that rely on multiple sequential position determinations in order to estimate the speed vector of the user. In such cases, two additional parameters have to be calculated: the Direction followed by the user and his/hers Velocity. These parameters are generally expressed in degrees and meters per second, respectively.

Scalability – The scalability is a desired and welcomed characteristic of a positioning system. It represents the positioning system's ability to readily respond to any augmentation. The augmentation can be in terms of Coverage area, Availability, frequency and total number of positioning requests, etc.

Complexity – There are many definitions for complexity depending on the domain of application. Nevertheless, in terms of positioning systems, complexity is most often referred to as the property that describes the difficulty of setting up the positioning system.

Cost effectiveness – This abstract characteristic of a positioning system is not entirely independent of its other performance parameters (e.g. Complexity and Scalability).

For example, the greater the Complexity of the system, the lower the Cost effectiveness. One of the ways of describing it is as a ratio between the benefits it provides (how broad range of LBSs it enables) and the costs it induces for the user.

As can be seen from the aforementioned, the latter three parameters don't have standardized units and are usually of descriptive nature.

The approaches and metrics used in order to obtain the user's position are also worth discussing. There are a few fundamental methods of acquiring the user's location:

1) Based on the identification of "base station" to which the user is associated (Cell-ID or Cell of Origin - COO) - This simple approach assumes that the estimated location of a user is equal to the location of a "base station" to which the user is associated. In other words, the user is estimated to be in a location of the "nearest" node of the network. This method is used both in indoor and outdoor environments (GSM, UMTS). Its popularity, despite inferior performances, is due to the simplicity of implementation. Obviously, the accuracy is proportional to the density of the network nodes.

2) Based on the time of signal arrival (Time of Arrival - TOA) - Being that the waves (electromagnetic, light and sound) are propagating through the free space at constant speed, it is possible to assess the distance between the transmitter and a receiver based on the time that the wave propagates in-between those two points. This approach assumes that the receiver is informed of the exact time of signal's departure. Being that this is not always easily accomplished, the alternative approach takes into account the time needed for signal to propagate in both directions (Round Trip Time - RTT). This way, one station is transmitting the predefined sequence. The other station, upon receiving the sequence, after a strictly defined time interval (used for allowing the stations of different processing power to process the received information), resends the sequence. The station that initially sent the sequence can now, by subtracting the known interval of time that the signal was delayed at second station from the measured time interval, assess the time that signal propagated to the other station and back and, consequently, the distance between the stations. This approach is less difficult to implement than TOA, since it does not require the stations to be synchronised.

3) The distance between the stations can be measured based on the differences in times of signal arrival (Time Difference of Arrival - TDOA) - With this approach, the problem of precisely synchronised time in transmitter and receiver is resolved by using several receivers that are synchronised whereas the transceiver, whose location is being determined, does not have to be synchronised with the receivers. Upon receipt of the transmitted signal, a network node computes the differences in times of the signal's arrival at different receivers. Based on that calculation, the user's location is determined as a cross section of two or more hyperboles. Owing to that, these techniques are often referred to as hyperbolic techniques.

4) Based on the signal's angle of arrival (Angle of Arrival - AOA or Direction of Arrival - DOA) - The idea, with this approach, is to have directional antennas which can detect the angle of arrival of the signal with the maximal strength or coherent phase. This procedure grants the spatial angle to a point where the signal originated (and whose location is determined). Vice versa, the mobile terminal can determine the angle of arrival of the signal from the known reference transmitters. Being that this approach is often implemented through the use of antenna arrays, the latter approach can have significant impact on the mobile terminal and is, therefore, less commonly exercised.

5) Based on the received signal strength (Received Signal Strength Indication – RSSI) – The free space signal propagation is characterised with predictable attenuation dependent on the distance from the source. Moreover, in real conditions, the attenuation also largely depends on the obstacles and the configuration of the propagation path. That is why there are various mathematical models which describe the wave propagation for diverse surroundings and, ultimately, estimate the signal attenuation for the observed environment. This approach grants the distance of the entity whose position is being determined, to one or more transmitters.

6) Based on the fingerprint of the location (Database Correlation or Location Fingerprinting) – With this approach, the certain, location dependant, information is acquired in as many Reference Points (RPs) across the coverage area of the technique. This data is stored into so called Location Fingerprints Database. Afterwards, when the actual position determination process takes place, the information gathered at the unknown location is compared with the pre-stored data and the entity's position is estimated at a location of a pre-stored fingerprint from the database whose data are "closest" to the measured data.

Most often, the estimated position with TOA and RSSI approaches is determined by lateration. The process of lateration consists of determining the position of the entity when the distance between the entity and one or more points with identified positions (i.e. reference points) is known. To uniquely laterate the position in N-dimensional space, the distances to N+1 reference points ought to be known. With TDOA approach, the estimated position is obtained as a cross-section of two or more hyperbolas in two-dimensional space, or three or more hyperbolic surfaces in case of three-dimensional space. The process of angulation is employed with AOA and DOA approaches. This process estimates the location of a user as a cross-section of at least two rays (half-lines) originating at known locations. The lateration and angulation processes are depicted in Fig. 1. As for the Location Fingerprinting approach, the estimated location is obtained by utilizing the correlation algorithm of some sort. This algorithm determines, following a certain metric, the "closeness" of the gathered data to the pre-stored samples from the location fingerprinting database.

Apart from these, basic, approaches, there are a number of other choices and hybrid techniques that combine the aforementioned approaches when determining the estimated position of the user.

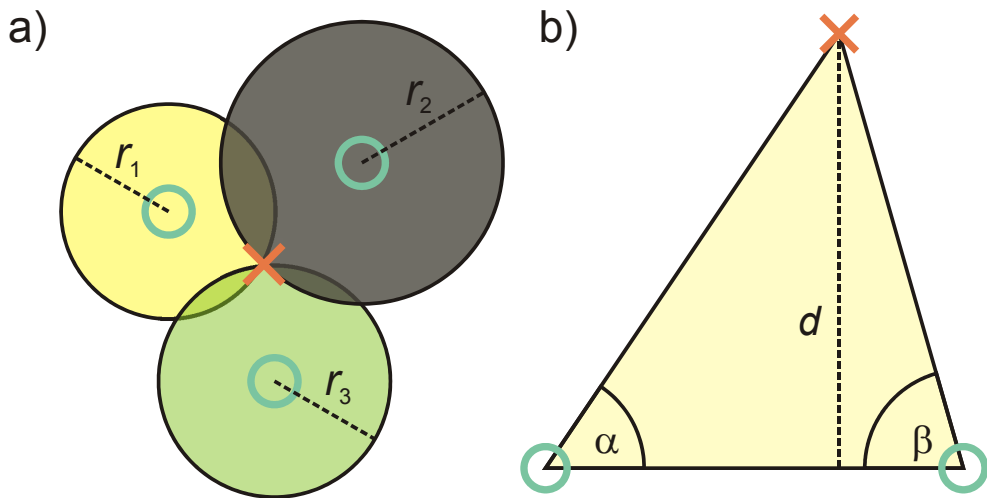


Fig. 1. The processes of estimating a user location: a) Lateration and b) Angulation (Green circles represent the known positions and the red cross stands for the estimated location)

3. Classifications of Positioning Systems

There are more than a few classifications of positioning systems. While some of them are very strict, others can be very arbitrary and overlapping. Without the need to judge or justify any of them, the most common ones are given herein.

Regarding the type of provided information, positioning techniques can be split into two main categories: Absolute and Relative positioning.

Absolute positioning methods consist in determining user location from scratch, generally by using a receiver and a terrestrial or satellite infrastructure. A well-known example of systems based on “absolute positioning” is the American GPS.

Relative positioning methods consist in determining user location by calculating the movements made from an initial position which is known. These methods do not rely on an external infrastructure, but require additional sensors (e.g. accelerometers, gyroscopes, odometers, etc). Inertial Navigation Systems used in commercial and military aircrafts are a good example of systems based on relative positioning.

LBSs currently offered by wireless telecommunication operators or by service providers are all based on absolute positioning methods and not on relative positioning methods, since these services are offered to users whose initial position is generally not known.

Within the “absolute positioning” family, the measurements and processing required for determining user’s location can be performed in many different ways and rely on different means. Thus, many different absolute positioning methods can be used for determining user’s location. These methods can be clustered into different groups, depending on the infrastructure used. Hence, the positioning techniques can be divided into:

- Satellite-based,
- PLMN-based (Public Land Mobile Network based), and
- Other (such as: WLAN, Bluetooth, RFID, UWB, etc).

The first group, which is known by the largest audience, is the “Satellite positioning” group. This group relates to the positioning methods which are based on the use of orbiting satellites, such as the GPS, Glonass or Galileo. Many applications and services based on satellite positioning have been developed during the past years (e.g. in-vehicle navigation, fleet management, tracking and tracing applications, etc). They generally require the use of dedicated receivers. Today, more and more devices such as PDAs or mobile phones include a satellite positioning capability, and this trend should persist in the future.

The second group, the “PLMN positioning” group, corresponds to the location techniques which have been developed for public land mobile networks. Initially deployed in the US under the pressure of the FCC mandate which forces US carriers to locate users placing calls to the 911 emergency number, location technologies are now being implemented in most of European wireless telecommunication networks for commercial purposes. Most of cellular positioning methods are incorporated in mobile telecommunication standards (2G/2.5G/3G/3.5G), but some solutions remain based on proprietary techniques.

The third and last group, the “Other positioning” group, corresponds to those technologies which have not been developed specifically for positioning purposes, but that can be used, in addition to their primary function, for determining user’s location. These technologies encompass WLAN and Bluetooth for instance.

Another distinction can be made, depending on the “place” where the position calculation is made. In some cases, the main processing is performed at the terminal level. In other cases, the main processing is performed in the network. Therefore, the positioning techniques can be classified into:

- Network-based (also referred to as mobile-assisted), and
- Terminal- or Mobile-based (also referred to as network-assisted).

Satellite technologies, as a rule, fit in the Terminal-based positioning techniques. As for the positioning techniques from the PLMN and other groups, they can not be *apriori* associated to either of the Terminal- or Network-based groups.

Finally, the positioning techniques can be classified according to the environment of their coverage. Hence, the positioning techniques can be divided into:

- Outdoor, and
- Indoor.

Although there are intense research efforts to adopt the Satellite-based positioning techniques for Indoor environment, they are still considered to fit into Outdoor group. PLMN-based positioning can be implemented in both Indoor and Outdoor environments, whereas techniques from “Other” group usually fit Indoor environments.

Positioning techniques designed for a particular Indoor environment in most cases fit into Relative positioning group.

Bearing in mind the ongoing convergence process of telecommunication systems and numerous, newly developed, hybrid positioning techniques, the indoor/outdoor categorization as well as other aforementioned classifications ought to be regarded more as guidelines than as strict lines that divide techniques into disjoint sets.

4. Non-Radio Indoor Positioning Systems

This section contains a brief overview of the non-radio positioning systems most commonly used in indoor environment.

4.1 IrDA Positioning Systems

IrDA technologies are based on devices with infrared light transceivers. This light occupies the part of spectrum between the visible light and the radio-waves. Upon encountering an obstacle, such as wall, the major part of the IR light's energy is being absorbed. Therefore, in order to communicate properly, two IR devices must have unobstructed Line of Sight (LoS) path between them. This poses a limitation for employing this technology in positioning purposes.

The most popular application of this technology for positioning use is the "Active Badge" technique (Want et al., 1992). The person or entity, whose position is being determined, possesses a device, badge alike, which periodically emits its ID code via IR transmitter. The IR sensors must be deployed in the coverage area (building). The position of the user is then determined based on the Cell-ID principle. With respect to the attributes of the IR light, the sensors must be deployed in every room in which the positioning feature is needed. Consequently, the accuracy of this technique is on a room level.

Other techniques based on this technology offer various accuracy and applications. The systems with greater number of IR receivers and transmitters on each device are proposed (Krohn et al., 2005). These systems are able to accurately estimate the position of a mobile communication device (e.g. PDA, laptop, digital camera, etc.) in order to allow them to automatically synchronise or perform other location dependent tasks. These activities are supposed to be performed on a flat, table alike surface. The obtained distance error is less than 20cm in more than 90% of the cases. On the other hand, there are systems that augment the "Active Badge" technique by using more IR sensors, micro VGA display and, optionally, video cameras. These systems provide so called Argumented Reality (Maeda et al., 2003). The typical application of an Argumented Reality system would be for the museum environment, where the visitor would be, via micro display (in eyeglasses, for example), fed with the information related to the exhibit he is currently experiencing.

4.2 Ultrasound Positioning Systems

The term ultrasound is related to the high frequency sound waves, above the part of spectrum perceivable to the human ear (20kHz). Although the ultrasound is most frequently used in medicine, there are other areas of application such as: biomedicine, industry (e.g. flow-meters), chemistry, military applications (sonic weapon), etc. As for the positioning purposes, the greatest benefit of using the ultrasound positioning is the product of a fact that ultrasound propagates through the air at limited speed, which is by far smaller than the speed of light. Therefore, the implementation of techniques based on time of flight (i.e. TOA, TDOA) of signal is very much facilitated. Moreover, the mechanic nature of sound waves grants ultrasound positioning techniques immunity to electromagnetic interference which could also be considered as an advantage. It ought to be pointed out that ultrasound waves do not penetrate, but rather reflect of walls. Therefore, the ultrasound receiver, in order to detect the signal, must be in the same room as transmitter but LoS is not necessary.

Ultrasound positioning systems can be classified according to the number of ultrasound "base stations" (transmitters and/or receivers) in each room (Dijk, 2004). The basic ultrasound positioning technique comprises one receiver in each room, and a ultrasound emitting tag which is worn by the entity that needs to be positioned. In this case, the

accuracy is on the level of the room. These systems are commercially available for some time now.

More sophisticated ultrasound positioning systems invoke the use of a greater number of transmitters in each room as well as the use of RF (seldom IR) signals for precise determining the time delay (Fraser, 2006). In this case, the controlling unit, which is connected to all the ultrasound emitters in one room as well as with RF transmitter, determines the exact time when each of the transmitters is about to send its chirps. Commonly, the RF signal is emitted first and then the chirps from all ultrasound transmitters are emitted separated by known time intervals. The receiver, knowing the separating time intervals and the propagation speed of RF and ultrasound waves, can now calculate, based on the time it received each of the chirps, the distance to each of the ultrasound emitters. The position is then determined by lateration. Consequently, for three-dimensional positioning at least four transmitters per room are required. The accuracy is in range of 10cm in 90% of the cases.

Furthermore, the system that eliminates the need for RF transmitter has been developed (McCarthy & Muller, 2003). With this system, the processing power of the receiver can be reduced, and the whole system is less complex. The transmitters are cyclically emitting chirps in constant time intervals whereas the receiver is employing an extended Kalman filter for resolving the chirp transmission and receipt times.

5. Indoor Radio Positioning Systems

The RF positioning techniques employ different parts of the frequency spectrum. Some are implemented on existent short-range radio interfaces and serve as added services, while others are especially developed for positioning. The most common RF technologies which, through the use of these techniques, enable positioning are: RFID, UWB, Bluetooth and WLAN.

5.1 RFID (Radio-Frequency Identification) Positioning Systems

The beginnings of this technology go far back to the time of the Second World War. Over the recent years, due to the cheaper RFID components, the expansion of this technology is occurring.

RFID system consists of tags, reader with antenna and accompanying software. The tags are usually placed on entities whose position needs to be determined. The Line of Sight between the tag and a reader is usually not necessary. The tags can contain additional information apart from its ID code which broadens the usage this technology.

There are three types of RFID tags:

- Passive tags do not have their own power supply. In order to operate, they use the energy, induced on their antenna, from the incoming radio wave from the reader. Using that energy, the passive tag replays by emitting its ID code and, optionally, additional information. Passive tags have very limited range (from a few cm up to a couple of meters). Their advantage is within the scope of cheap construction, compact size and cheap production.
- Active tags are encompassed with power supply which enables them unrestricted signal emission. This kind of tags is more reliable and immune to highly polluted RF environments. Their range can go up to a few hundreds of meters.

- Semi-active tags are equipped with battery power supply. Recent constructions enable a battery life span of more than 10 years.

RFID devices can operate in different frequency bands: 100 – 500 kHz, 10 - 15 MHz, 850 – 900 MHz, and 2.4 – 5.8 GHz (Don Chon et al., 2004).

RFID positioning techniques are based on knowing the position of the reader. When the tagged object enters the range of the reader, its position is assumed to be equal to the position of the reader (similar to Cell-ID). Correspondingly, it is possible to deploy tags across the coverage area. In that case the reader is mounted on the entity whose position is being determined. The accuracy depends on the density of deployed objects (tags/readers) across the coverage area. With active tags, the positioning accuracy can be upgraded with the RSSI information.

Most common application areas of RFID technology are in replacing the barcode readers, product tracking and management, personal documents identification, identification implants for humans and animals, etc. It is interesting to mention that the latter aforementioned application raises numerous ethical issues and there are organized groups worldwide opposing the implementation of this technology.

5.2 Bluetooth Positioning Systems

Bluetooth is a short-range, low-consumption radio interface for data and voice communication (Muller, 2001). Initially conceived in the mid 90s by the Ericsson Mobile Communication as a technology that ought to replace the cable in personal communications, Bluetooth shortly gained significant popularity. Ericsson was joined by IBM, Microsoft, Nokia and Toshiba. They formed Bluetooth Special Interest Group (SIG) with an aim to standardize Bluetooth specifications. Independent group, called Local Positioning Working Group, had a goal of developing the Bluetooth profile which would define the position calculation algorithm as well as the type and format of the messages that would enable Bluetooth devices to exchange position information.

The basic Bluetooth specification does not support positioning services per se (Bluetooth Special Interest Group Specification Volume 1 and 2, 2001). In absence of such support, various research efforts have produced diverse solutions. Bahl and Padmanabhan used the RSSI information for in-building locating and tracking (Bahl & Padmanabhan, 2000). Patil introduced the concept of reference tags and readers (Patil, 2002). He also investigated separately cases when Bluetooth supports and does not give support to RSSI parameter. On the other hand, the research by Hallberg, Nilsson and Synnes goes to saying that RSSI parameter is unreliable for positioning purposes and that its employment ought to be avoided with Bluetooth positioning systems (Thapa & Case, 2003).

In addition, there are ideas of exploiting other parameters than RSSI for positioning purposes. Link Quality and Bit Error Rate (BER) are most commonly referred in this context. However, it should be stated that these solutions are still under development, and that Link Quality is not uniquely defined and is therefore dependent on the equipment manufacturer. Also, BER parameter is not defined in the basic Bluetooth specifications and must be extrapolated from the message received as a response to echo command supported at L2CAP layer. All in all, these parameters undoubtedly contain location dependent information, but the extraction of that information is still subject to research.

The accuracy of Bluetooth positioning systems is decreasing with the increase in the maximal range of the system (Hallberget al., 2003). That is, with the range increase, the

positioning system uncertainty is increased as well, therefore the accuracy is worsened. The improvement of accuracy can be achieved through communicating with more than one Bluetooth nodes and possibly utilizing some of the aforementioned parameters (RSSI, Link Quality, BER). Finally, the major application of Bluetooth technology is expected in ad-hoc networks and the positioning techniques and LBS should be conceived and designed accordingly.

5.3 UWB (Ultra-WideBand) Positioning Systems

Ultra-wideband is a short-range high data throughput technology. The ultra-wideband signal is defined (Harmer, 2004) as a radio-signal that occupies at least either 500MHz of frequency spectrum or 20% of the central frequency of the band. There are many ways in which the UWB signal can be generated. Two, most important from the positioning point of view, are:

- 1) Impulse UWB - By generating very short impulses, with sub nanosecond duration, that are mutually separated several tenths of nanoseconds. Clearly, this signal inherently possesses very wide band.
- 2) Frequency Hopped UWB - By generating the typical DSSS (Direct-Sequence Spread Spectrum) with the signal spectrum ranging from 10 to 20MHz which is then hopped around 1GHz frequency, applying between 10 and 100 thousands of hops per second.

Unlike conventional radio-signals, the impulse UWB signals are practically immune to multipath propagation problems. With conventional signals, the reflected component of the signal is, in its large part, overlapped with the component that is travelling the direct path. Hence, the direct and reflected component interfere at the receiver causing fading. Contrary to that, with impulse UWB technology, due to the very short pulse duration, the reflected component is most often arriving at the receiver after the direct component has been completely received. With respect to this feature, the UWB positioning techniques utilising high resolution TOA approach come as the logical choice. Typically, the position accuracy of 1m in more than 95% of the cases is achievable.

Employing the mobile nodes of the UWB network for accuracy improvement is also under research. Computer simulation (Eltaher & Kaiser, 2005) shows that the positioning error could be further reduced by employing a larger number of antennas with the beamforming capabilities.

Bearing in mind the amount of research in this area, the wider scale commercialisation of indoor UWB positioning systems can be expected in proximate future.

5.4 WLAN Positioning Systems

Positioning techniques in WLAN networks are growing in popularity. The reason for this can be looked in-between the widespread of 802.11 networks and the fact that a broad scope of LBSs can be brought into an existing network without the need for any additional infrastructure. There are a number of approaches to the positioning problem in WLAN networks. Unquestionably, the most popular ones are based on the Received Signal Strength Information (RSSI). Nevertheless, there are other approaches that depend on timing measurements or require additional hardware but offer superior accuracy and/or faster implementation in return (Llombart et al., 2008; King et al., 2006; Sayrafian-Pour & Kaspar, 2005).

Positioning with the use of RSSI parameter can be, in its essence, regarded as the path loss estimation problem. The nature of the path loss prediction in an indoor environment is extremely complex and dependent on a wide variety of assumptions (e.g. type of the building, construction, materials, doors, windows, etc.)(Nešković et al, 2000). Even if these basic parameters are known, precise estimation of the path loss remains a fairly complex task.

Depending on the side on which the position calculation process takes place, positioning in WLAN networks with the use of RSSI parameter can be either network-based or client-based. Whereas the client-based solutions gather the RSSI vector from the radio-visible APs, the network-based solutions have a central positioning engine which collects the client's signal strength vector from the APs and produces the position estimate. The network-based solutions do not require clients to have a specific software installed which is of great essence for security purposes. Moreover, the client does not need to be associated with the network – the positioning can be done solely based on the probe requests the client sends (in case of active scanning). Network-based solutions could also have an important advantage over the client-based ones when used in WLAN networks employing the Automatic Radio Management (ARM). This centralized mechanism is used to obtain the optimal radio coverage by changing the channel assignment and adjusting the output power and/or radiation pattern of the APs. Contrary to the client-based solutions, the network-based positioning engine could take into account the changes made by ARM mechanism while the ARM mechanism would present a setback for the client-based solutions. On the other hand, client's Network Interface Cards do not have to be consistent regarding the radiated power which may, depending on the positioning algorithm used, present an analogue problem for network-based solutions. In this work, for explanatory purposes, usually the client-based solution will be presented. However, the reader should keep an open mind towards the analogue network-based option.

Regarding the approach used to determine the user's position, WLAN positioning techniques can be categorised as: propagation model based, fingerprinting based or hybrid. Propagation model based techniques rely on statistically derived mathematical expressions that relate the distance of an AP with the client's received signal strength. The estimated position of the user is then obtained by lateration. Therefore, if there are less than three radio-visible APs (for two-dimensional positioning) the estimated user's position is ambiguous. Also, the model derived for one specific indoor environment is usually not applicable to other indoor environments.

Fingerprinting techniques are most commonly used for WLAN positioning. They are conducted in two phases: the off-line or training phase, and the on-line or positioning phase. The off-line phase comprises collecting the RSSI vectors from various APs and storing them, along with the position of the measurement, into a fingerprinting database. In the on-line phase, the estimate of the user's position is determined by "comparing the likeliness" of the RSSI vector measured during the on-line phase with the previously stored vectors in the database. The fingerprinting process is shown in Fig. 2. These techniques have yielded better performance than other positioning techniques, but are believed to have a longer set-up time.

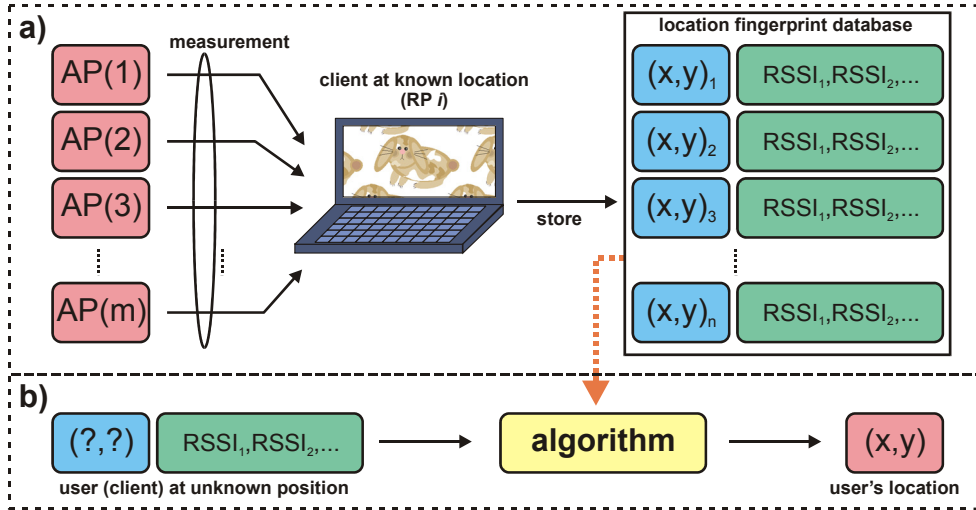


Fig. 2. Two phases of positioning: a) training phase – mobile client is recording RSSI vectors across RPs and stores them in fingerprint database, and b) positioning phase – based on the measured RSSI vector and database access, the algorithm estimates the user's location

Hybrid techniques combine features from both propagation modelling and fingerprinting approaches, opting for better performances than propagation model techniques and shorter set-up time than fingerprinting techniques (Wang & Jia, 2007).

The prospects of using RSSI parameter for indoor positioning were first systematically analysed in "RADAR" (Bahl & Padmanabhan, 2000). According to this research, it is better to use RSSI than SNR (Signal to Noise Ratio) for positioning purposes since the RSSI parameter is much more dependent on the client's position than SNR. Two algorithms to establish the user's location were proposed. The first one is the Nearest Neighbour (NN) algorithm which compares the RSSI vector of a mobile client against the RSSI vectors previously stored in the fingerprinting base. An extension to the proposed algorithm was also considered: the estimated location is not identified as only one RP whose RSSI vector is closest to the observed RSSI vector, but calculated as a "middle" point of k closest RPs (kNN algorithm). This analysis has shown that algorithm performance improved for $k = 2$ and $k = 3$. For larger k , the performance had started to decrease. The second algorithm is based on a simple propagation model with Rician distribution assumed. It ought to be emphasized that both approaches require a minimum of three radio visible access points (APs). The measuring campaign comprised 70 RPs. At each RP measurements were made for four orientations of the receiver, and each measurement was averaged from 20 samples.

To produce the maximum amount of information from the received RSSI vectors, the Bayesian approach was proposed (Li et al., 2006). This concept yields better results than the NN algorithm. The Bayes rule can be written as:

$$p(l_t | o_t) = p(o_t | l_t) p(l_t) / N \quad (1)$$

where l_t is location at time t , o_t is the observed RSSI vector at time t , while N is a normalizing factor that enables the sum of all probabilities to be equal to 1. In other words, at a given

time t , the probability that a client is at location l_i , if the received RSSI vector is o_i , is equal to the product of the probability to observe RSSI vector o_i at location l_i and the probability that the client can be found at location l_i . The process of estimating client's location is based on calculating the conditional probability $p(l_i|o_i)$ for each RP. The estimated client's location is equal to the RP with the greatest conditional probability. To accomplish this task, two terms on the right hand side of Eq. (1) ought to be calculated. The first term, also referred to as the likelihood function, can be calculated based on the RSSI map (for all RP) using any approach that will yield probability density function of observation o_i for all RPs. As for the a priori probability $p(l_i)$, it ought to be calculated according to the client's habits. However, for most cases the assumption of uniform distribution across all RPs is valid. The measurements were made at 70 RPs. As with the previously discussed techniques, the measurements were made for four orientations of a receiver, and each measurement was averaged from 20 samples.

Another project, named Horus (Youssef & Agrawala, 2005; Eckert, 2005), had the goal of providing high positioning accuracy with low computational demands. This is also a probabilistic approach in which time series of the received signal strength are modelled using Gaussian distributions. Due to the time dependence of the signal strength from an observed AP, the authors of this project have shown that the time autocorrelation between the time adjacent samples of signal strength can be as high as 0.9. To describe and benefit from such behaviour, they have suggested the following autoregressive model:

$$s_t = \alpha s_{t-1} + (1 - \alpha) v_t, \quad 0 \leq \alpha \leq 1 \quad (2)$$

where v_t is the noise process and s_t is a stationary array of samples from the observed AP.

Throughout the off-line phase, the value of parameter α is assessed at each RP and stored into the database along with Gaussian distribution parameters μ and σ . In the on-line phase, Gaussian distribution is modified according to the corresponding values of α retrieved from the fingerprinting database. Alike to the kNN algorithm, the Horus system estimates the client's location as a weight centre of k RPs with the highest probabilities. The principal difference to the kNN algorithm is that, in case of Horus system, the k most likely RP are multiplied with their corresponding probabilities. For verification purposes, the authors made measurements at 612 RPs, and each measurement was averaged from 110 samples.

More relevant information about the statistical modelling approach towards location estimation can be found in (Roos et al., 2002) and in the references found therein.

Battiti et al. (2002) were the first to consider using Artificial Neural Networks (ANNs) for positioning in WLAN networks. This approach does not insist upon a detailed knowledge of the indoor structure, propagation characteristics, or the position of APs. A multilayer feedforward network with two layers and one-step secant training function was used. The number of units in the hidden layer was varied. No degradation in performance was observed when the number of units grew above the optimal number. For verification purposes, measurements were made at 56 RPs, and each measurement was averaged from 100 samples.

In most studies, WLAN positioning techniques are compared on the subject of their accuracy while other attributes of a positioning technique such as latency, scalability, and

complexity are neglected. Another aspect that is seldom analyzed is size of the environment in which the technique is implemented.

It also ought to be pointed out that averaging the RSSI vectors in the on-line phase has an immense impact on the technique's latency, so the scope of location based services that could be utilized with such techniques is significantly narrowed. Moreover, bearing in mind that all presented approaches require at least three radio-visible APs in each RP (which is seldom the case in most WLAN installations), feasibility of sound frequency planning is uncertain. Consequently, the degradation of packet data services is inevitable with respect to positioning in larger indoor areas (i.e. large number of APs is required). Enabling the radio-visibility of three APs across the indoor environment is usually constructively irrational and economically unjustified. Hence, the presented techniques cannot be applied to the majority of existing WLAN networks optimized for packet data services.

Finally, there are other studies that accompany the research for sophisticated positioning in WLAN networks. Other relevant research efforts comprise the impact of Network Interface Card on the RSSI parameter, compensation of small-scale variations of RSSI, clustering of locations to reduce the computational cost of positioning, use of spatial and frequency diversity, methods for generating a larger location fingerprinting database by interpolation, and unequal fusing of RSSI from different APs (Kaemarungsi, 2006; Youssef & Agrawala, 2003; Ramachandran & Jagannathan, 2007; Li et al., 2005; Zhang et al., 2008).

6. Cascade-Connected ANN Structures for WLAN Positioning

The ANNs are an optimisation technique known to yield good results with noise polluted processes (Hasoun, 1995). They are generally classified as a fingerprinting technique. In the off-line phase, the set of collected RSSI fingerprints is used to train the network and set its inner coefficients to perform the positioning function. In the on-line phase, the trained network replaces trilateration and position determination processes.

Two basic concepts, a single ANN and a set of cascade-connected ANNs structures with space partitioning, have been presented herein. These models were implemented in Matlab and verified on a 147m x 67m test bed with eight APs. For training purposes, the *traingda* - gradient descent training function with adaptive learning rate was selected. All neural units had the hyperbolic tangent sigmoid transfer function. Being that the input probability distribution function of RSSI values is near Gaussian, the Mean Square Error (MSE) was selected as a criterion function (Hanson, 1988).

Regarding the purpose that ANN is intended for and, moreover, the nature of the problem, it has been concluded that multilayer feedforward neural networks with error backpropagation have substantial advantages in comparison to other structures (Nešković, 2000). The outer interfaces of the ANN must match the number of the APs on the input side (i.e. eight inputs), and the number of coordinates as outputs (i.e. two outputs).

Multilayer feedforward networks can have one or more hidden layers with perceptron units. The hidden layers with corresponding perceptron units form the inner structure of the ANN. There is no exact analytical method for determining the optimal inner structure of the network. However, there are algorithms that, starting with an intentionally oversized network, reduce the number of units and converge to the optimal network structure. Also, there are other algorithms such as the cascade correlation learning architecture (Fahlman & Lebiere, 1990) that build the network towards the optimal structure during the training

process. However, being aware of the fact that these procedures can be complex and that determining the most optimal structure was not the central scope of this research, we intentionally slightly oversized our network's inner structure knowing that an oversized network will not yield degradation in performance. We also adopted that the first hidden layer ought to have more perceptrons than the input layer so that the input information is quantified and fragmented into smaller pieces (Shang & Wah, 1996). The number of perceptrons in the following hidden layers ought to decrease, converging to the number of perceptrons in the output layer. Bearing that in mind, the chosen structure for single ANN (type 1) approach consisted of the input layer, three hidden layers and the output layer. The number of perceptrons per layer was (from input to output) 8-15-9-5-2.

When utilizing space partitioning, the positioning process is split into two stages where each stage could be implemented with the most suitable model. In this case, the two-step space partitioning is implemented utilizing cascade-connected ANNs. The block scheme of this system is shown in Fig. 2.

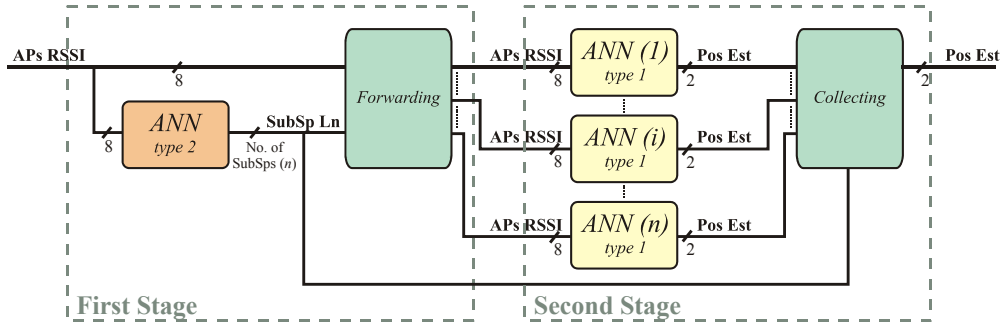


Fig. 2. Cascade-connected ANNs system structure (the input is the observed RSSI signal vector, **APs RSSI**, and the position estimate vector, **Pos Est**, is the output)

In the first stage, an ANN (type 2) is used to determine the likeliness of a measured RSSI vector belonging to one of the subspaces. This ANN (type 2) has 8 inputs and the number of outputs is equal to the number of subspaces the environment is partitioned to. Each output corresponds to the likeliness that a received RSSI vector originates from a particular subspace. The outputs of the type 2 ANN, **SubSp Ln**, are connected to the Forwarding block which, depending on the inputs, employs only one of the second stage networks by forwarding the **APs RSSI** vector.

The inner structure of ANN (type 2) is designed using the same guidelines as with the single ANN model. Therefore, it also has three hidden layers and the number of perceptron units in those layers is varied to fit the different number of subspaces. The second stage ANNs are type 1 networks with structure identical to the previously described ANN used with the single ANN approach.

In the off-line phase, type 2 ANN is trained with the fingerprinting database that originates from the whole environment. The targeted output vector has only one non-zero element (equal to 1). The index of that element corresponds to the number of the subspace from which the RSSI vector originates. Type 1 networks are trained following the training methodology from the single ANN approach with the only difference being that each type 1 ANN is trained with only the part of the fingerprinting database which originates from a particular subspace.

In the on-line phase, the first stage ANN estimates the likeliness that the received RSSI vector originates from a particular subspace. The Forwarding block then determines the most likely subspace by searching for the maximum value in the output vector from the ANN (type 2) and forwards the **APs RSSI** vector only to the second stage ANN that correspond to that subspace. The appropriate second stage ANN then determines the estimated position of the user and, finally, the collecting block forwards that estimate to the structure output. Several space separation patterns were chosen yielding a different number of subspaces ranging from 4 to 44. The space partitioning patterns that have been employed are shown in Fig. 3.

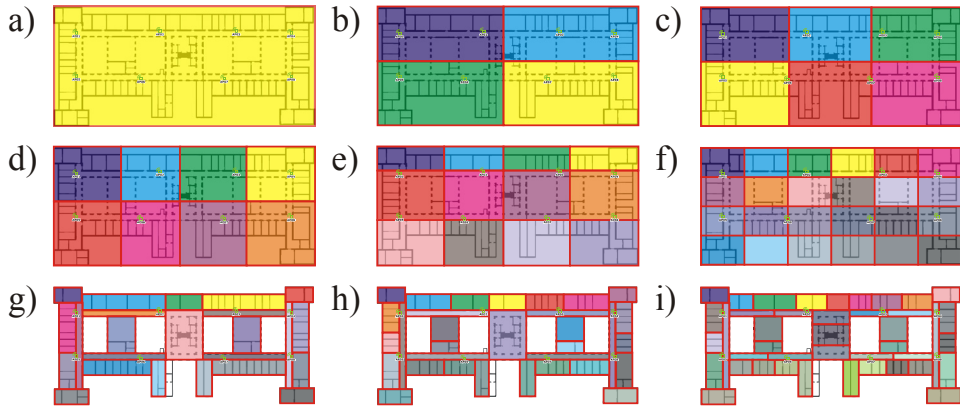


Fig. 3. Space partitioning patterns: a) no space partitioning (1x1), b) 2x2, c) 2x3, d) 2x4, e) 3x4, f) 4x6, g) x24, h) x32, and i) x44

The partitions with a smaller number of subspaces were made on geometrical bases. However, with the increase in the number of subspaces, the subspace size decreased until it came to a room size level. It was then worth to consider partitioning space in an other manner. Starting with 24 subspaces (which was also portioned on geometrical bases), the partitions were made on “logical” bases (i.e. x24, x32 and x44). This logical separation opted for subspaces to be as homogeneous in the propagation manner as possible (e.g. partitioning was made trough walls wherever possible). Note, the single ANN model is herein referred to as 1x1 partitioning.

For the purpose of determining the optimal training parameters, as well as the optimal training duration, the complete set of measurements was split into two subsets. The larger subset was used to train the ANNs, while the smaller, containing measurements from a 100 randomly chosen RPs, was used to validate the obtained models.

The results obtained for different space partition patterns, for optimally trained ANNs, are presented in Table 1.

From Table 1, it can be seen that, with geometrical partitioning, the overall median and average distance errors decrease with the increase in number of subspaces. This behaviour is even more emphasized with the distance errors in the correctly chosen subspace which confirms the influence of environment size on positioning accuracy. When concerning the logical partitioning, slightly better results are obtained for 24 subspaces (4x6 vs. x24) but, with the further increase in the number of subspaces, the average distance error is starting to rise again. Also, with the increase in the number of subspaces the probability of correct

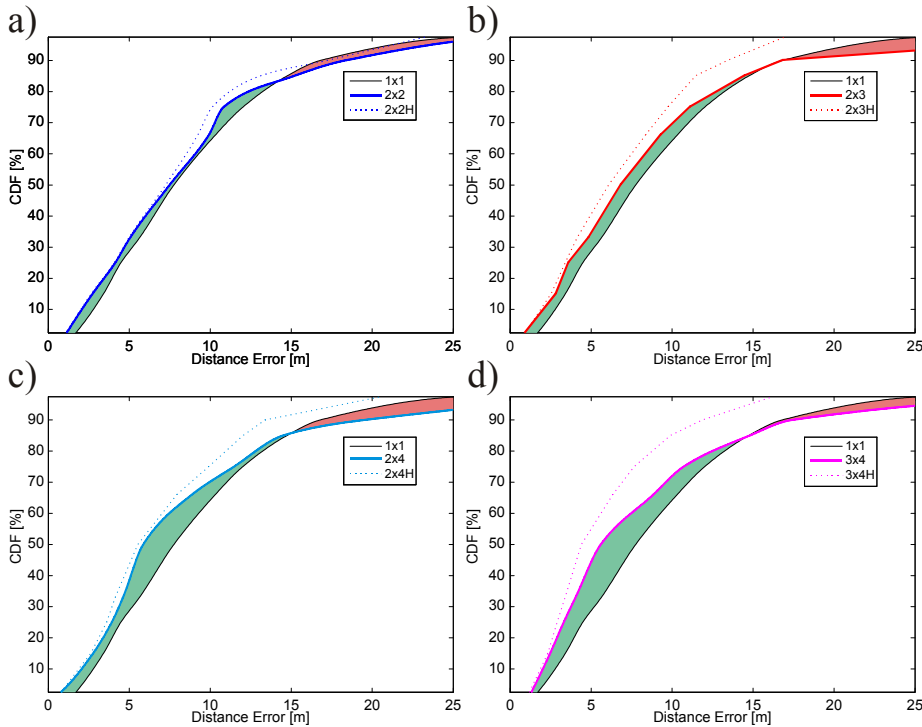
subspace being chosen declines as expected while the probability of correct room estimation rises from 26% for a 1x1 positioning to as much as 66% for a x24 configuration, after which it starts declining a little.

Pattern	1x1	2x2	2x3	2x4	3x4	4x6	x24	x32	x44
Overall Average DE ^a [m]	9.26	9.00	8.97	8.91	8.54	8.28	8.14	8.58	9.11
Overall Median DE ^a [m]	7.75	7.49	6.87	5.86	5.59	5.10	4.57	4.70	4.44
Average DE ^a in IS ^b [m]	-	21.3	22.7	21.2	19.0	18.0	18.4	19.5	19.2
Median DE ^a in IS ^b [m]	-	15.4	17.4	15.3	16.3	14.7	17.5	15.8	16.1
Average DE ^a in CS ^c [m]	9.26	8.35	6.99	6.96	5.76	4.20	4.07	3.78	3.72
Median DE ^a in CS ^c [m]	7.75	7.33	6.13	5.52	4.40	3.87	3.56	3.39	3.32
Probability of CSE ^d	1.00	0.95	0.87	0.86	0.79	0.71	0.72	0.69	0.65
Probability of CRE ^e	0.26	0.42	0.48	0.52	0.58	0.62	0.66	0.62	0.61

^a Distance Error, ^b Incorrect Subspace, ^c Correct Subspace, ^d Correct Subspace Estimation, ^e Correct Room Estimation

Table 1. Performance overview for different partitioning patterns

To better understand and discuss the performances of cascade-connected ANNs with space partitioning, we observed and compared the distance error’s Cumulative Distribution Function (CDF) of a single ANN approach with the cascade-connected ANNs. Fig. 4. shows the obtained CDFs for representative space partitioning patterns.



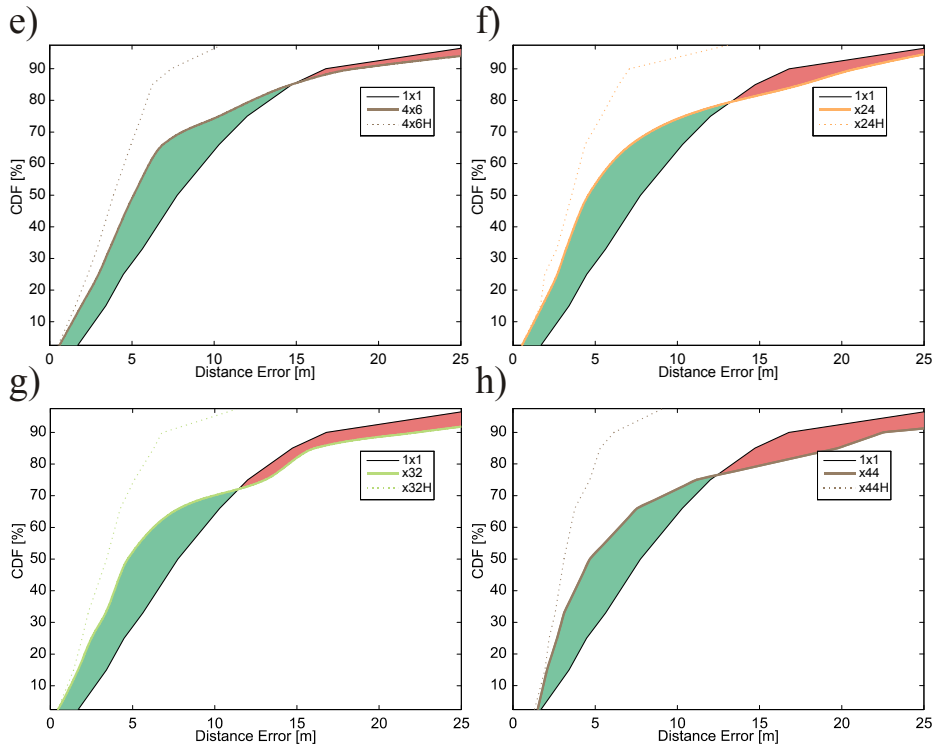


Fig. 4. Cumulative Distribution Function of distance error: a) 1x1 and 2x2 partitioning and correct subspace estimation – 2x2 H; b) 1x1 and 2x3 partitioning and correct subspace estimation – 2x3 H; c) 1x1 and 2x4 partitioning and correct subspace estimation – 2x4 H; d) 1x1 and 3x4 partitioning and correct subspace estimation – 3x4 H; e) 1x1 and 4x6 partitioning and correct subspace estimation – 4x6 H; f) 1x1 and x24 partitioning and correct subspace estimation – x24 H; g) 1x1 and x32 partitioning and correct subspace estimation – x32 H; h) 1x1 and x44 partitioning and correct subspace estimation – x44 H

The green filled areas on Fig. 4. could be considered as a partitioning gain in comparison to 1x1 positioning, while the red filled areas could be considered as partitioning loss. It can be seen that, with geometrical partitioning, Fig. 4. a) – e), the gain areas are increasing with the increase in the number of subspaces. When concerning logical space partitioning Fig. 4 f) – h), it can be noticed that the best performances are obtained with x24 pattern – average distance error 8.14m, median error 4.57m. With the further increase in the number of subspaces, the benefit of decreasing the median error has faded, even though the median error in correct subspace continues to decrease, whereas the average distance error is starting to rise again due to the augmentation in probability of incorrectly chosen subspace. In other words, with the further increase in the number of subspaces, the partitioning gain surfaces are still expanding however, the partitioning loss surfaces are rising as well. Furthermore, it should be noticed that with the increase in the number of subspaces, the CDF is starting to create a knee roughly around 60th percentile. This has two effects: the green surfaces are getting larger as discussed and the crossing angle between the space

partitioning model and 1x1 positioning is increasing while the crossing point between the two is being pushed towards lower percentiles. The latter of the two effects has a negative impact on positioning performances.

Finally, if the Average DE in correct subspace, from Table 1, is compared with the average subspace area (the total area size divided by the number of subspaces), it can be seen that with the increase in size of the subspaces the increase in average error is getting saturated. So, given the constant RPs and APs density, the further increase in size of the test bed should induce only the minor rise of the DE. This also goes to say that the chosen verification environment was large enough to comprehensively explore the influence of the test bed size on positioning accuracy.

7. References

- Bahl, P. & Padmanabhan, V. (2000). „Radar: An in-building RF-based user location and tracking system“, Proceedings of the IEEE Infocom 2000, Tel-Aviv, Israel, vol. 2, Mar. 2000, pp. 775-784.
- Battiti, R., Nhat, T. L., Villani, A. (2002). Location-aware Computing: A Neural Network Model For Determining Location in Wireless LANs. Technical Report # DIT-02-0083 (Feb. 2002)
- Bluetooth Special Interest Group (2001). Specification Volume 1, Specification of the Bluetooth System, Core. Version 1.1, February 22, 2001.
- Bluetooth Special Interest Group (2001). Specification Volume 2, Specification of the Bluetooth System, Profiles. Version 1.1, February 22, 2001.
- Hae Don Chon, Sibum Jun, Heejae Jung, Sang Won An, (2004). „Using RFID for Accurate Positioning“, Journal of Global Positioning Systems, 2004, Vol. 3, No. 1-2: 32-39
- Collomb Frédéric, (2002). Location Service Study Report (Loc_Serv_Study_Rep_PU.doc), Mobile and Vehicles Enhanced Services, 2002.
- Eckert, K. (2005). Overview of Wireless LAN based Indoor Positioning Systems, Mobile Bussiness Seminar, University of Mannheim, Germany, (2005)
- Eltaher A., Kaiser T., (2005). „A Novel Approach based on UWB Beamforming for Indoor Positioning in None-Line-of-Sight Environments“, RadioTeCc, October 26-27, 2005, Berlin, Germany.
- Esko O. Dijk, (2004). „Indoor ultrasonic position estimation using a single base station“, Technische Universiteit Eindhoven, 2004. October 6, 2004, p44-45
- Fahlman, S., Lebiere, C. (1990). The cascade-correlation learning architecture. Advances in Neural Information Processing Systems 2, pp. 524-532 (1990)
- Fraser M., (2006). „Mobile and Ubiquitous Computing: Sensing Location Indoors.“, COMSM0106, 2006.
- Hallberg, Nilsson, Synnes, (2003). „Positioning with Bluetooth“, Telecommunications, 2003. ICT 2003. 10th International Conference on, Volume 2, Issue , 23 Feb.-1 March 2003 Page(s): 954 - 958 vol.2
- Harmer D., (2004). „Ultra Wide-Band (UWB) Indoor Positioning“, Thales Research and Technology UK Ltd. ARTES 4 Project. ESTEC December 2004.
- Hasoun H. M. (1995). Fundamentals of Artificial Neural Networks. Massachusetts Institute of Technology (1995)

- Hanson, S. J., Burr, D. J. (1988). Minkowski-r Backpropagation: Learning in Connectionist Models with non-Euclidean Error Signals, *Neural Information Processing Systems* (Denver, 1987), (editor Anderson D. Z.), pp. 348-357, American Institute of Physics, New York (1988)
- Kaemarungsi, K. (2006). Distribution of WLAN received signal strength indication for indoor location determination. 1st International Symposium on Wireless Pervasive Computing, 2006. pp.6 (Jan. 2006)
- King, T., Kopf, S., Haenselmann, T., Lubberger, C., Effelsberg, W. (2006). COMPASS: A Probabilistic Indoor Positioning System Based on 802.11 and Digital Compasses. University of Mannheim, D-68159 Mannheim, Germany, TR-2006-012
- Krohn, A. Beigl, M. Hazas, M. Gellersen, H.-W. (2005). „Using fine-grained infrared positioning to support the surface-based activities of mobile users“, *Distributed Computing Systems Workshops*, 2005. 25th IEEE International Conference on, Telecooperation Office, Karlsruhe Univ., Germany.
- Li, B., Salter, J., Dempster, A., Rizos, C. (2006). Indoor Positioning Techniques Based on Wireless LAN. *AusWireless '06*, Sydney (March 2006)
- Li, B., Wang, Y., Lee, H.K., Dempster, A., Rizos, C. (2005). Method for yielding a database of location fingerprints in WLAN. *Communications, IEE Proceedings- Volume 152, Issue 5*, pp.580 - 586 (Oct. 2005)
- Llombart, M., Ciurana, M., Barcelo-Arroyo, F. (2008). On the scalability of a novel WLAN positioning system based on time of arrival measurements. 5th Workshop on Positioning, Navigation and Communication, 2008. WPNC 2008, pp. 15 - 21 (March 2008)
- Masaki Maeda, Takefumi Ogawa, Takashi Machida, Kiyoshi Kiyokawa, Haruo Takemura, (2003). „Indoor Localization and Navigation using IR Markers for Augmented Reality“, *HCI International 2003 Interactive demo*
- Michael R. McCarthy, Henk L. Muller, (2003). „RF Free Ultrasonic Positioning“, University of Bristol, 7th International Symposium on Wearable Computers, October 2003.
- Muller N. (2001). „Bluetooth Demystified“, McGraw-Hill, New York, 2001.
- Nešković A., Nešković N., Paunović Dj., (2000). „Indoor Electric Field Level Prediction Model Based on the Artificial Neural Networks“, *IEEE Communications Letters*, vol. 4, No. 6, June 2000
- Patil, A. (2002). „Performance Of Bluetooth Technologies And Their Applications To Location Sensing“, Michigan State University, 2002.
- Ramachandran, A., Jagannathan, S. (2007). Spatial Diversity in Signal Strength based WLAN Location Determination Systems. 32nd IEEE Conference on Local Computer Networks, 2007. pp. 10 - 17 (Oct. 2007)
- Ramachandran, A., Jagannathan, S. (2007). Use of Frequency Diversity in Signal Strength based WLAN Location Determination Systems 32nd IEEE Conference on Local Computer Networks, 2007. pp.117 - 124 (Oct. 2007)
- Roos, T., Myllymaki, P., Tirri, H. (2002). A statistical modeling approach to location estimation. *Mobile Computing, IEEE Transactions on*, Volume 1, Issue 1, pp.59 - 69, (First Quarter 2002)
- Roy Want, Andy Hopper, Veronica Falcao, and Jon Gibbons, (1992). „The Active Badge location system“, *ACM Transactions on Information Systems*, 10(1):91-102, January 1992.

- Sayrafian-Pour, K., Kaspar, D. (2005). Indoor positioning using spatial power spectrum. IEEE PIMRC 2005. Volume: 4, pp. 2722- 2726 (Sept. 2005)
- Shang Y., Wah W. B. (1996). Global Optimization for Neural Network Training, IEEE Computer Society, pp. 45-56 (March 1996)
- Thapa K., Case S., (2003). „An indoor positioning service for Bluetooth Ad Hoc networks“, in: MICS 2003, Duluth, MN, USA.
- Wang, H., Jia, F. (2007). A Hybrid Modeling for WLAN Positioning System. International Conference on Wireless Communications, Networking and Mobile Computing, 2007. pp.2152 - 2155 (Sept. 2007)
- Youssef, M., Agrawala, A. (2005). The Horus WLAN Location Determination System. Int. Conf. on Mobile Systems, Applications And Services, pp.205-218 (2005)
- Youssef, M., Agrawala, A. (2003). Small-scale compensation for WLAN location determination systems. Wireless Communications and Networking, 2003. Volume 3, 20-20 pp.1974 - 1978 (March 2003)
- Youssef, M.A., Agrawala, A., Shankar, A. U. (2003). WLAN location determination via clustering and probability distributions. Proceedings of the First IEEE International Conference on Pervasive Computing and Communications. pp.143 - 150 (March 2003)
- Zhang, M., Zhang, S., Cao, J. (2008). Fusing Received Signal Strength from Multiple Access Points for WLAN User Location Estimation. International Conference on Internet Computing in Science and Engineering, 2008. pp.173 - 180 (Jan. 2008)

Location in Ad Hoc Networks

Israel Martin-Escalona, Marc Ciurana and Francisco Barcelo-Arroyo
Universitat Politècnica de Catalunya (UPC)
Spain

1. Introduction

An *ad hoc*¹ network is defined as a decentralised wireless network that is set up on-the-fly for a specific purpose. These networks were proposed years ago for military use, with the purpose of communicating devices in a highly constrained scenario. Under such a network, devices join and leave the network dynamically; thus, it cannot be expected to have any kind of network infrastructure. This wish for decentralised on-the-fly networks has subsequently expanded to cover several fields besides the military. Today, there are several mobile services requiring the self-organising capabilities that ad hoc networks offer. Examples include packet tracking, online-gaming, and measuring systems, among others. Ad hoc networks have obvious benefits for mobile services, but they also introduce new issues that regular network protocols cannot cope with, including optimum routing, network fragmentation, reduced calculation power, energy-constrained terminals, etc.

In ad hoc networks, positioning takes a significant role, mainly due to the on-the-fly condition. In fact, several services require nodes to know the position of the customers in order to perform their duty properly. Wireless sensor networks concentrate most of the services that need positioning to perform their duty. Such networks constitute a subset of ad hoc networks involving dense topologies operating in an ad hoc fashion, and they are composed of small, energy and computation constrained terminals. In ad hoc networks, and especially in wireless sensor networks, nodes are spread over a certain area without a precise knowledge about the topology. In fact, this topology is variable. Accordingly, there are several unknowns (e.g., node density and coverage, network's energy map, the presence of shadowed zones, nodes' placement in the network coverage area) that are likely to constrain the performance of ad hoc services. Knowledge of the terminals' locations can substantially improve the service performance.

Positioning is not only important for the service provisioning; it is also crucial in the ad hoc protocol stack development. Due to the changes in the topology and the lack of communication infrastructure, ad hoc protocols have to address several issues not present in regular cellular networks. Routing is one of the best examples of the dependence of ad hoc networks on positioning. Studies such as (Stojmenovic, 2002) demonstrate that only position-based routing protocols are scalable, i.e., able to cope with a higher density of

¹ "Ad hoc" is actually a Latin phrase that means "to this (thing, purpose, end, etc.)"

nodes in the network. The same seems to apply to other management and operation tasks in ad hoc networks.

1.1 The Location problem in ad hoc networks

Nodes in an ad hoc network can be grouped into three categories according to the positioning capabilities: beacon nodes, settled nodes, and unknown nodes. *Beacon nodes*, also known as *anchors* or *landmarks*, are those able to compute their position on their own, i.e., without using an ad hoc location algorithm. Accordingly, they implement at least one location technique (e.g., GPS, map matching), which can be used as standalone. Beacon nodes usually constitute the reference frame necessary to set up a location algorithm. *Unknown nodes* are those nodes that do not know their position yet. When an ad hoc network is set up, all nodes except the beacon nodes are unknown. *Settled nodes* are unknown nodes that are able to compute their position from the information that they exchange with beacon nodes and/or other settled nodes. The purpose of the ad hoc location system is thus to position as many nodes as possible, turning them from unknown to settled nodes (Bourkerche et al., 2007).

Location systems in ad hoc networks function in two steps: local positioning and positioning algorithm. The former is responsible for computing the position of an unknown node from the metrics gathered. The second step consists of the positioning algorithm, which indicates how the position information is managed in order to maximise the number of nodes being settled.

1.2 Measuring the performance of location solutions

Performance of location solutions in ad hoc networks can be computed according to several parameters. The main ones are presented below.

1.2.1 Accuracy

Ad hoc positioning requires good accuracy since most of the networks in ad hoc mode are deployed in constrained scenarios, often indoors. In such environments, accuracy is especially relevant, since a few meters of error in the position may cause the node to be identified in another room, floor, or even building. Furthermore, nodes are expected to be very close (e.g., in medical applications), and inaccurate positions could hinder operation and maintenance tasks or even prevent location-based applications from performing their duty. Thus, location algorithms for ad hoc networks must produce positions for settled nodes of the highest possible accuracy.

1.2.2 Latency

The location solutions must be able to converge, i.e., to produce as many settled nodes as possible in the shortest time. Ideally, the location solutions for ad hoc networks should turn all the unknown nodes into settled nodes in a defined time. However, optimality in terms of accuracy (and many other factors) may collide with latency, which means that the convergence (and hence the latency) of the location solution depends on the accuracy requested, among many other factors. Latency is also modulated by the mobility of the nodes in the network. The faster the nodes move, the shorter the convergence period needs to be. It must be noted that, in the case that convergence is not achieved, estimated positions would not involve the actual location of nodes, and this lack of accuracy would be spread by

the network according to the location algorithm used. Hence, it would degrade the accuracy of the location solution.

1.2.3 Form factor of terminals

Ad hoc devices tend to be small, especially in the case of wireless sensor networks. The requirement of small form constrains the capabilities of these devices, which prevents sophisticated (and usually more complex) algorithms from being used. This, in turn, prevents the best QoS from being reached.

1.2.4 Energy-efficient design

Due to restrictions on their size, ad hoc devices tend to include very-limited batteries. Accordingly, location solutions must avoid using complex algorithms or sensing multiple metrics in order to compute the position, since these would limit the lifetime of nodes and hence of the entire network.

1.2.5 Self-organising design

Ad hoc devices are likely to move. Algorithms developed for positioning in such networks should account for the overhead generated by the changes in the topology of the ad hoc network. Changes in the topology are produced by nodes moving around the network area or being added to and removed from the ad hoc network.

1.2.6 Random nature of the ad hoc network

Ad hoc devices can be added and removed from the network during its life cycle. Depending on the scenario, these changes in the topology of the network can be noticeably intense. Location solutions in ad hoc networks should be insensitive to the structure of the network.

1.2.7 Scalability

The algorithm should minimise the impact of adding a new terminal to the network. This means that the amount of resources consumed by the network due to the addition of the new node should be as low as possible. Scalability does not necessarily involve the use of simple algorithms. However, it should allow as many ad hoc devices as possible to be positioned with the same amount of resources.

1.2.8 Node density

In actual deployments of ad hoc networks, devices are not likely to be homogeneously distributed along the layout. The density of terminals is variable in space and time, and hence location algorithms should not assume isotropic conditions.

1.2.9 Beacon percentage

Beacons tend to be fixed nodes often plugged to wired power sources, which make them more durable. Beacons are set up by the network operator, and, consequently, they collide with the ad hoc philosophy. Moreover, it is difficult to ensure that the percentage of beacons visible for an unknown node remains uniform. Accordingly, location algorithms should be as insensitive to beacon percentage and beacon placement as possible.

2. Location metrics

There are several metrics than can be used as input for location techniques. Those metrics are usually known as observables, since they refer to what can be observed and subsequently measured. Timestamps, angles, and signal strength are metrics commonly used by location techniques to compute the position of the nodes. The former usually involves timestamps for the sending and receiving moments associated with one or several signals. Precision of timestamps directly depends on the clock present in the ad hoc devices. These clocks are usually low-profiled, mainly due to the small form-factor and the cost of devices. Consequently, this impacts the accuracy of the time-based observables and ultimately the positions fixed by the location solution. Furthermore, accessing the hardware clock is rather difficult in most of the current devices and technologies. The same does not apply to signal strength, which is usually available in most of the ad hoc technologies. Consequently, there are several location techniques that use this metric for positioning. Furthermore, this metric provides accurate observables, even though it does not mean that positions computed from these observables achieve the same degree of accuracy.

Finally, angular information is proposed for positioning in several solutions. This metric consists of measuring the angle or direction of arrival (AoA / DoA) of signals coming from several nodes to the target one, or vice versa. Fig. 1 illustrates this metric, where the red-coloured angles (i.e., ε_1 and ε_2) stand for the error produced in the angle-of-arrival estimate.

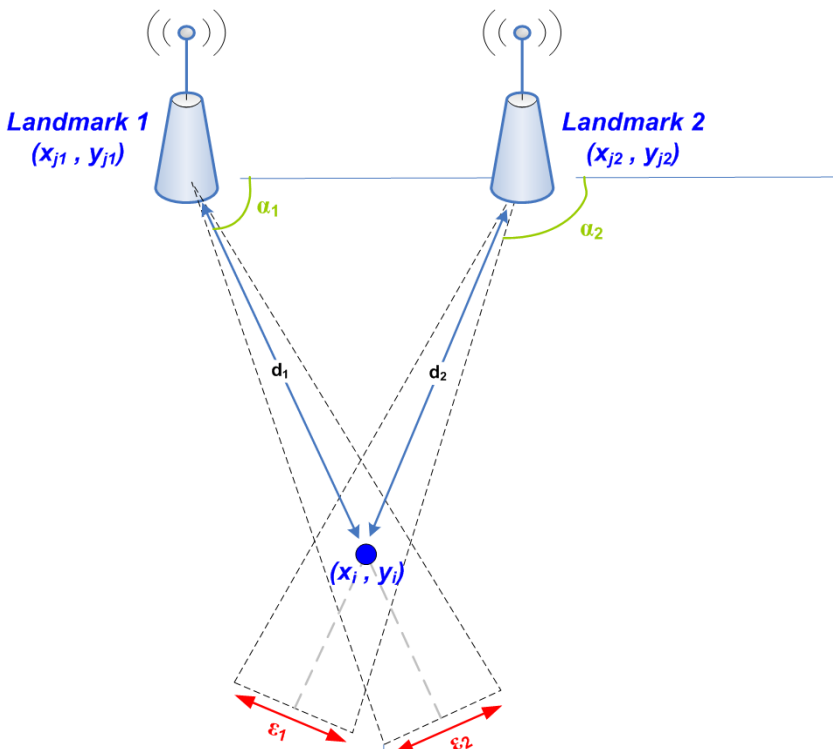


Fig. 1. Positioning according to the angle/direction of arrival

According to Fig. 1, the angles of arrival can be computed as

$$\alpha_k = -\tan^{-1}\left(\frac{y_i - y_{jk}}{x_i - x_{jk}}\right), \quad (1)$$

where (x_i, y_i) is the position (in two dimensions) of the node to be positioned and (x_{jk}, y_{jk}) is the position (in two dimensions) of the landmark k .

The use of this metric involves using arrays of antennas in order to capture the angle in which the signal is being received. Furthermore, the positioning error derived from the angle-estimation error depends on the distance between the pair of entities involved in the angle estimation. Consequently, this metric is rarely used; the hardware is costly, the error is range-dependent, and this metric often involves the customisation of network equipment.

3. Location techniques

Ad hoc networks use a subset of the location techniques proposed for other cellular technologies (e.g., UMTS, IEEE 802.11, etc.). These techniques can be classified into two main groups: ranging-based and angle-based. The following sections describe the techniques in detail.

3.1 Ranging based on signal strength

Ranging-based techniques are based on computing the distance between two nodes (i.e., ranging) and then computing the position of the unknown node by using a multilateration algorithm. Range estimations can be computed from several metrics, but two are preferably used: signal strength and timestamp. Techniques based on the former estimate the range between two nodes according to the received and transmitted power. Radio path models depend on the distance according to a certain power, known as path-loss slope. This means that distance can be computed from transmitted and received signal strength, which is information easily accessible in the network. According to general knowledge on radio propagation, received power can be expressed as

$$P_{rx} = P_{tx} - P_{1m} - 10\alpha \log(d), \quad (2)$$

where P_{rx} and P_{tx} are, respectively, the received and transmitted power in dB(m), P_{1m} stands for the losses at 1 meter from the transmitter location, d is the distance in meters between the transmitter and receiver placements (i.e., the ranging), and α is the path-loss gradient (or slope).

Modelling the radio path losses, such as those in Equation (2), is a difficult task. Obstacles in the propagation path affect the signal in several ways, namely, reflection, diffraction, and absorption. The consequence is that signals reach the receiver following more than a single path (a phenomenon known as multi-path), and, consequently, the received signal strength suffers random variations. Accordingly, different radio path models are proposed depending on the scenario in which the network is going to be deployed. However, the propagation conditions are likely to change (even dramatically) with time as new obstacles appear. Hence, such models would need to be recalibrated periodically (constantly in the

worst case). Indoors is one of the most constrained scenarios, and, consequently, most of the location solutions based on signal strength ranging are proposed for such an environment. An example of a radio propagation model used in location is proposed in (Seidel & Rapport, 1992), where the free space model was adapted to indoor environments by adding several parameters, such as the number of floors in the path or the number of walls. However, it is demonstrated not to be a satisfactory approach since the number of obstacles is not known a priori. Other approaches tried to improve radio signal propagation models for indoors (Wang et al., 2003; Lassabe et al., 2005), but accurate distance estimates are not yet available. Despite all these issues, several proposals are available for signal ranging. One of the first approaches was presented in (Bahl & Padmanabhan, 2000), where several models specifically addressed to ranging-based location solutions were proposed and tested experimentally. Due to the randomness of the received signal, poor results were obtained with all models, if compared with other location techniques based on signal fingerprinting. Better results are reported in (Kotani et al., 2003), where the authors propose processing the signal-strength observables prior to position computation. This previous stage aims to reduce the noise of the measurements so that more accurate positions can be fixed. Furthermore, an extended Kalman filter is used to compute the position, which minimises the variance of the distance estimation. This solution provides accuracy figures of less than three metres, even though worse results are expected under arbitrary propagation conditions.

3.2 Ranging based on time measurements

Considering the number of solutions currently proposed, time-based ranging seems to be a more appealing technology than solutions based on signal strength. This is because, compared to signal strength, time measurements tend to be more stable and less sensitive to environmental conditions. Time-based ranging solutions can be classified into two groups according to the number of signals/paths under consideration: time of arrival and time-difference of arrival. The former is based on estimating the distance between two nodes. It is achieved by marking transmission (t_{tx}) and reception (t_{rx}) times and then applying

$$d = c(t_{rx} - t_{tx}) \quad (3)$$

to compute the distance, where c stands for the propagation time (usually the speed of light). Equation (3) can only be applied if timestamps are taken under the same time line, i.e., when all nodes in the network are time-synchronised. However, this is not the normal case. Thus, the 2-way time-of-arrival approach was proposed to overcome this issue. This approach computes the propagation time under a round-trip-time approach, i.e., measuring the time spent by the signal in travelling the forward (*Node 1* to *Node 2*) and backward (*Node 2* to *Node 1*) paths. Fig. 2 illustrates this procedure, which is explained in detail in (Ciurana et al., 2007). Since all timestamps are taken using the same clock (in *Node 1*), propagation time (T_{prop}) can be computed as half the time measured for both paths (RTT), as long as processing time (T_{proc}) is negligible (or calibrated). The position of the target node (i.e., the one to be located) can be computed by multilateration once enough measurements are achieved (e.g., 3 or more for 2D positioning). The accuracy of time of arrival directly depends on the precision of the time estimations. Accordingly, there are several works addressing improvements in the accuracy of time of arrival observables.

Some examples can be found in (Ibraheem & Schoebel, 2007) and (Reddy & Chandra, 2007), which present approaches improving the traditional correlation-based methods.

Time-difference of arrival consists of observing ranging differences, rather than just observing distances. Therefore, the node that is going to be positioned measures the ranging difference between its position and the position of a pair of landmarks (or even settled nodes). Time difference of arrival is also known as hyperbolic multilateration, since it superposes several hyperbolas in order to fix the node's position. It provides better accuracy than time of arrival and hence is preferred in cellular networks (3GPP, 2002; 3GPP, 2004). However, time-difference solutions present the same issue as time of arrival: nodes have to be time-synchronised. Furthermore, the 2-way approach cannot easily be applied to the time-difference of arrival technique, and hence synchronisation error has to be estimated (by means of specific measurement devices) or removed (taking the location measurements with a common clock).

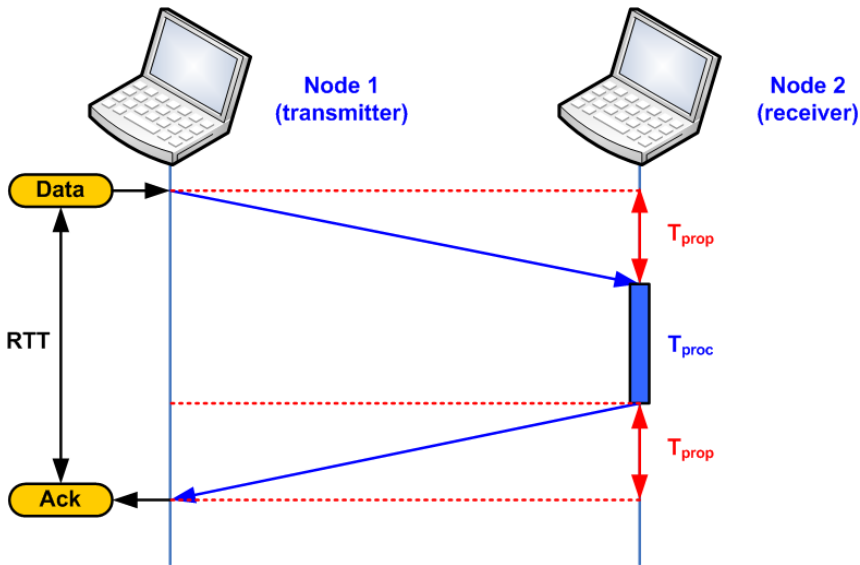


Fig. 2. RTT estimation through 2-way time of arrival approach

3.3 Triangulation based on angular measurements

This location method uses the direction or angles of arrival of several signals as the metric to compute the position. Therefore, the location techniques based on angulation are also known as Angle of Arrival (AoA) or Direction of Arrival (DoA). The position of the user can be computed according to several approaches. One of the simplest consist of intersecting the lines computed as

$$y_i = y_{jk} - (x_i - x_{jk}) \tan \alpha_k, \quad (4)$$

where α_k is computed according to Equation (1). There are several (and more complex) proposals based on numerical approach and closed forms to carry out positioning with angulation, as presented in (Pages-Zamora et al., 2002).

4. Location algorithms

Local positioning has been widely addressed for cellular networks, and the main methods and procedures remain valid for ad hoc networks. Hence, positioning algorithms draw the greatest amount of interest from the research community for location in ad hoc networks.

4.1 Taxonomy of location algorithms in ad hoc networks

4.1.1 Centralised vs. Distributed

Centralised algorithms rely on a network entity (e.g., location server) that gathers the location information from all the unknown nodes and then computes their position. The main advantage is that global optimisation can be performed, as the information of all nodes is available in the location server. Centralised systems are often used in cellular networks such as public land mobile networks (PLMNs), but they collide with the random nature of ad hoc networks. The location server has to be significantly more powerful than regular nodes. Moreover, the data gathered by the server must be synchronised: all measurements must be performed at specific times. If synchronisation is not assured, optimality cannot be reached, and the system may be degraded. Additionally, topology in such systems can be displayed as a tree, with the root at the location server. Therefore, nodes near the server quickly run out their batteries, since they concentrate most of the location traffic coming from unknown nodes; this reduces the ad hoc network lifetime.

In *distributed* location algorithms, some or all nodes are able to compute their position (and the position of other nodes depending on the specific algorithm). Thus, distributed location is more robust to node failures. Distributed algorithms also converge faster than centric solutions after topology changes and are usually insensitive to data synchronisation requirements, since only local or regional data are accounted for. There are several degrees of distributed algorithms. The most common are the localised or pure distributed algorithms, where all unknown nodes are able to compute their position once the necessary local metrics are available.

4.1.2 Incremental vs. concurrent

Incremental positioning algorithms start with only a few beacon nodes. Then, in each step, the position of a reduced amount of unknown nodes is computed using the information provided by settled and beacon nodes. The positions of such nodes are used in subsequent iterations to compute the location of other unknown nodes. The advantage of iterative algorithms is their simplicity. However, they tend to propagate their positioning error, since they use metrics obtained from settled nodes for subsequent local positioning. Furthermore, the convergence of incremental location algorithms is not always guaranteed. In *concurrent* algorithms, all nodes are able to compute their position normally using local information. Accordingly, they are more complex than iterative algorithms, but they can avoid error propagation and hence achieve better accuracy results.

4.1.3 One hop vs. multi-hop

Location involves exchanging information to measure the metrics used to compute the node's position. *One hop* algorithms use only information (i.e., metrics) local to the unknown nodes (e.g., ranging to the node's neighbours). On the other hand, *multi-hop* algorithms use

information of all nodes that can be reached from the node in a certain number of hops, usually two. Multi-hop techniques allow more accurate positions to be computed and fewer beacon nodes to be deployed in the system (Savvides et al., 2001). The main drawbacks are the overhead generated by the multi-hop estimation and the subsequent use of additional resources in the nodes to store such data.

4.1.4 Beacon-Based, Mobile Beacon-Based and Beacon-Less

Ad hoc location algorithms can be classified according to the presence of beacon nodes in three categories: localisation with beacons, localisation with moving beacons, and localisation without beacons (Sun et al., 2005). *Localisation-with-beacons* algorithms are those in which a percentage of nodes are fixed beacons, i.e., beacons that do not change their location. The major challenge of algorithms that rely on beacons is to maximise the accuracy and coverage while at the same time minimising the number of landmarks in the network. Localisation with *moving beacons* algorithms are similar to algorithms based on beacons, but here the beacons are no longer fixed and move through the network. A moving beacon is perceived by unknown nodes as different beacons (i.e., one per message exchanged, from different positions of the mobile beacon). Fewer landmarks are necessary, and more accurate positions can be achieved since beacon density is perceived as higher than it actually is. The main drawback with mobile beacon algorithms is that mobile beacons have to cover the entire ad hoc network and ensure that unknown nodes see the mobile beacons with a suitable frequency, which is often difficult to achieve. The last category, known as *beacon-free* location, involves those algorithms in which no node is aware of its position (i.e., all nodes are unknown). Thus, all nodes work together to compute their position using only their local information. This kind of algorithm usually works with a local coordinate system, which may require translation of the achieved positions into a global coordinate system so that they can be used by a location-based service or protocol.

4.1.5 Range-free vs. range-based

In *range-based* algorithms, the local position is computed according to ranging measurements (i.e., distance or angle estimates). Accordingly, they involve multilateration techniques, which are usually hardware-demanding and therefore energy-consuming. Accordingly, range-based algorithms are suitable for ad hoc networks with powerful terminals (e.g., in technologies such as IEEE 802.11). On the other hand, technologies with more constrained terminals, such as those present in wireless sensor networks, favour the use of range-free algorithms for positioning. *Range-free* algorithms do not rely on ranging to compute the position of unknown nodes; rather, they consist of simpler approaches based on proximity.

4.2 State of the art

The first location solutions proposed for ad hoc networks used centralised algorithms similar to techniques proposed for mobile networks. In (Doherty et al., 2001), the authors propose to manage the location as a convex-optimisation problem: a mobile server is defined to gather all the location data (e.g., distances, angles of arrival, etc.) from unknown nodes in order to compute their positions. The advantage is the simplicity and optimality of the positions computed. However, it involves delivering a significant amount of data to a location server that must be powerful enough to handle complex data structures. Moreover, the cost of the algorithm proposed for this technique is cubic in the number of connections,

which seriously constrains the scalability of the approach.

In order to overcome those drawbacks, distributed algorithms are present in many solutions. The *centroid* algorithm (Bulusu et al., 2000) is a one-hop pure-distributed positioning algorithm in which few beacons are spread in the ad hoc network forming a grid. Unknown nodes compute their position by estimating their range to the three closest beacons and a trilateration algorithm. The main benefit is that it is insensitive to the node density and does not add significant overhead in the network. The main drawback of this method is a larger error in the positioning. Another interesting example of a one-hop algorithm is proposed in the *Lighthouse* project (Römer, 2003), in which a single base station sees all sensors in the network. This full coverage is achieved by means of a beam that rotates at known speed. Stations are able to compute their position knowing the rotation speed, the width of the beam, and the signal time-of-flight. Although the accuracy might not be suitable for many applications in ad hoc networks (it provides a bias up to 14 metres), it represents an improvement in scalability.

Recent proposals emphasise distributed behaviour. The *Ad hoc Localisation System (AhLOS)* presented in (Savvides et al., 2001) is an example of this trend. AhLOS is a one-hop pure-distributed algorithm that uses three trilateration algorithms: atomic, iterative, and collaborative. Atomic trilateration involves a one-hop scenario, where the nodes have three or more beacons in sight so that they can compute their positions directly. The main drawback of this approach is that it relies on a high density of beacons. Iterative trilateration relaxes this assumption, considering all the nodes that compute their position by means of atomic trilateration (i.e., settled nodes) as new beacons. It allows fewer beacons at the cost of less accuracy. Despite covering most of the situations, these two trilateration approaches are not sufficient to position all the nodes in the ad hoc network, since unknown nodes with only one neighbour cannot be positioned. The authors of AhLOS followed a collaborative multilateration approach to overcome this situation, consisting of identifying those unknown nodes that cannot be handled by atomic and iterative algorithms and creating groups that collaborate in order to compute the position of those nodes. This may involve solving large nonlinear systems depending on the size of the groups created.

The *Ad hoc Positioning System (APS)* presented in (Niculescu & Nath, 2003) combines two concepts: beacon-based positioning and ad hoc propagation (i.e., multi-hop). The algorithm proposed in APS consists of four stages. Firstly, some beacon nodes are spread by the network. Secondly, nodes with some landmarks in sight measure their distance to the beacon nodes in terms of some metric, such as propagation time, number of hops, etc. Then, the information gathered by nodes in the neighbourhood of landmarks is propagated (and updated) using a proper algorithm. Finally, once a node has the ranging information of three or more beacons, it computes its position using a multilateration approach. The algorithms for propagating the ranging information in the ad hoc network are the main contribution of (Niculescu & Nath, 2003): *DV-Hop*, *DV-distance*, *Euclidean*, and *DV-Coordinate*. In the first, all nodes build ranging tables containing the position coordinates and the distance in hops to the landmarks. These data are flooded in the network in a controlled way, so that all nodes know how far they are from landmarks. On the other hand, landmarks use such data to compute the average distance of one hop, according to the hops and the distance between them. This is achieved by:

$$c_i = \frac{\sum_{\forall j \neq i} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{\sum_{\forall j \neq i} h_j}, \quad (5)$$

where c_i indicates the hop distance calculated by the landmark i , (x,y,z) are the coordinates of a landmark and h_j stands for the amount of hops from landmark i to j .

This average (i.e., c_i) is then flooded in the network using a singleton approach: once a node receives an announcement packet from a landmark containing the average value computed by such a landmark, it discards any other further announcement packet. The average distance is then stored in the nodes, which then use it to turn distances in hops into real ranges. Finally, nodes compute their position using a multilateration approach. The advantage of this approach is that it is insensitive to the ranging error, since ranging is based on hop counting. However, it introduces more overhead than other algorithms. Moreover, accuracy is degraded in non-dense networks.

The *DV-distance* approach is similar to DV-Hop but exchanges real distances instead of distances in hops. Multi-hop distances are computed as the sum of the distances between nodes involved in the path. Thus, this approach becomes more sensitive to ranging error. Normally, the denser the ad hoc network is, the better the accuracy. On the other hand, it improves the consistency of the DV-Hop approach, working similarly in isotropic and non-isotropic networks.

The *Euclidean* approach involves a multi-hop algorithm, which gathers ranging information up to 2 hops. Thus, the algorithm generates quadrilaterals involving one unknown node, two neighbours, and a landmark, and it infers the distance from the node to the landmark using trigonometric formulation. The advantage of this approach is that the ranging error can be estimated, stored, and subsequently flooded together with the distance; this allows distance-weighted approaches to be used and hence more accurate positions to be achieved. The main drawback is that a 2-hop approach involves additional traffic in the network (even more than in DV-Hop and DV-Distance algorithms) as well as more resources needed in the node to store the additional information, resulting in more quickly depleting node batteries running.

The last approach presented in (Niculescu & Nath, 2003) is the so-called *DV-Coordinate*, which is similar to the solution proposed in (Capkun et al., 2001). DV-Coordinate is based on each node computing the position of its neighbourhood according to a local coordinate-system. Then in a second step, called the registration stage, nodes exchange information to build the transformation matrices, which allow coordinates of local systems to be transformed from one local system to another. A global transformation matrix is necessary to achieve global coherence. DV-Coordinate performs almost the same as Euclidean. However, this approach impacts the scalability of the location system, since it depends on the square of the nodes in the network and involves sending two pieces of data instead of just a distance. In (Niculescu & Nath, 2003/2), the authors extend DV-Coordinate and presented the *Local Positioning System (LPS)*, which uses ranging and angle of arrival information to compute the position of unknown nodes, in a fashion similar to the DV-Coordinate. However, the LPS reduces the overhead of the DV-Coordinate systems, thus improving the scalability and updating only a reduced number of nodes each time, not the whole network.

The *Amorphous Localisation* algorithm (Nagpal et al., 2003) is similar to APS. It consists of computing the distance from nodes to beacons in terms of hops. However, the hop-distance is calculated offline according to the node density expected in the network. Then, a multilateration approach is followed to compute the position. The main drawback of this

algorithm is the offline stage, which seriously constrains the scalability of the algorithm in dynamic ad hoc networks.

A one-hop range-based concurrent pure-distributed algorithm is proposed in (Fu et al., 2006) for networks based on DSSS, such as those based on IEEE 802.11. This algorithm is based on propagating the clock from node to node so that the nodes involved in the positioning work in a synchronised fashion. Then the ranging to neighbour nodes is estimated, and a multilateration algorithm is applied. The synchronisation is achieved by means of the pseudo-noise code used in the DSSS, and, consequently, the time-resolution achieved matches with the code duration. Accordingly, the algorithm only works if times-of-flight are much longer than the code duration, which is expected to be the usual case.

The *Approximation of the Point-In-Triangulation Test (APIT)* is presented in (He et al., 2003) as another example of a one-hop range-based location algorithm. It is based on generating as many triangles as possible involving three beacons. Then, the APIT algorithm evaluates whether the unknown node is inside each triangle. Finally, it overlaps all the triangles, reducing the final positioning error. The authors evaluate the algorithm through simulation and conclude that this approach outperforms the *centroid* algorithm. Furthermore, this approach achieves accuracy figures similar to those obtained in the *APS* and *Amorphous Localisation* algorithms but requiring a lower node density and introducing less overhead. On the other hand, it requires beacons with a radio range longer than that of regular nodes.

All these approaches to the ad hoc location rely on active multilateration; i.e., positioning an unknown node involves a certain amount of location traffic in order to estimate the distances to landmarks. Active location constrains the scalability of location algorithms in ad hoc networks, in which topology and mobility are inherent to the network definition. The next sections introduce a passive algorithm for location in collaborative networks (e.g., ad hoc, wireless sensor networks, etc.), which aims to boost the scalability on positioning systems.

5. Passive positioning

Recent advances in indoor positioning have led to proposals that time-of-arrival (TOA) techniques for locating users are preferable to other techniques, such as fingerprinting. Time-of-arrival solutions achieve accuracy figures that are similar to those obtained by other techniques, but they do not require additional assistance for setup and maintenance. Conversely, time-of-arrival techniques need to calculate a client's range from at least three receivers at known positions in order to obtain a 2D position. In addition, all the signal transmitters involved in the TOA positioning system must be synchronised. Two-way TOA techniques, such as those presented in (Ciurana et al., 2007) and (Yang et al., 2008), cope with this issue by computing the range from the client (i.e., unknown node) to the base station (i.e., landmark) using a round-trip-time (RTT) procedure. Since only the client's clock is used to calculate the range, synchronisation between base stations is no longer necessary. The drawback is that more traffic is generated on the network, thus reducing the available throughput.

A recent proposal on ad hoc location presented in (Martin-Escalona & Barcelo-Arroyo, 2008) extends the capabilities of time-of-arrival location techniques, allowing unknown nodes in a network to position themselves in a passive fashion, i.e., without injecting traffic into the network. The following sections explain this technique, named *passive TDOA*, in detail.

5.1 Description of assisted passive-TDOA algorithm

The passive-TDOA algorithm listens to the access medium for messages that can be used to compute time-difference of arrival (TDOA) figures. The only assumption of the algorithm is that the nodes operate in a collaborative network. Note that this is the case for most of the wireless local area networks, especially those based on ad hoc protocols. In this text, one more assumption is taken only for explanatory purposes: the messages used to compute TDOAs are generated by unknown nodes running a 2-way TOA technique.

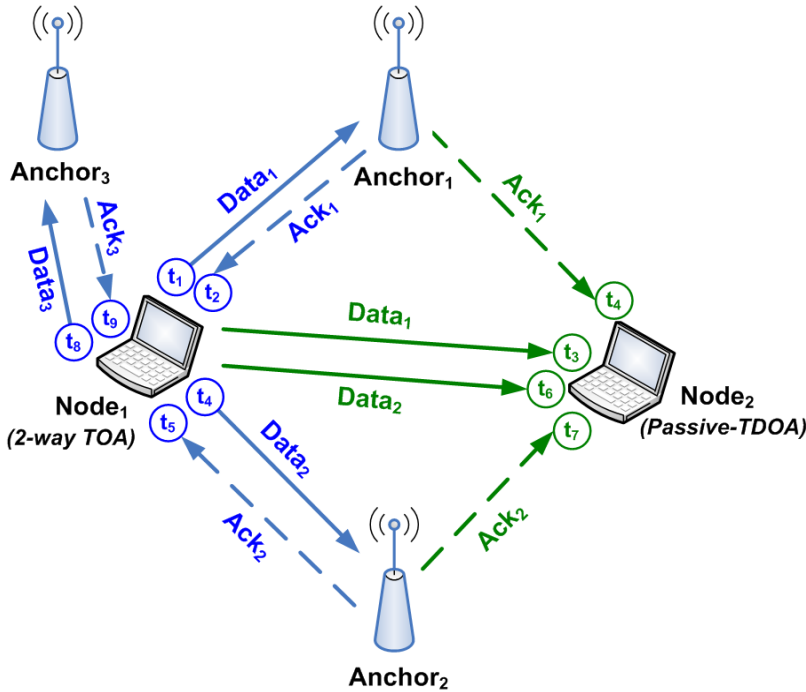


Fig. 3. Operation of the passive-TDOA algorithm

The performance of the positioning algorithm is described in Fig. 3, which shows a network with three anchors or landmarks (i.e., *Anchor₁* to *Anchor₃*) and two regular nodes with positioning capabilities (i.e., *Node₁* and *Node₂*). At a given time, *Node₁* begins a TOA positioning process to locate itself. Thus, *Node₁* sends data message (*Data₁*) to *Anchor₁* at t_1 , which replies with an acknowledgement (*Ack₁*), which reaches *Node₁* at t_2 . The corresponding RTT is hence calculated as $t_2 - t_1$. Note, however, that other nodes in the network also listen to all these messages, since it is a diffusion network. Thus, *Node₂* hears the *Data₁* message at t_3 and the reply to that message, i.e., *Ack₁*, at t_4 . Therefore, a TDOA measurement is generated as $t_4 - t_3$. The same process is followed by *Node₁* to range with *Anchor₂* and *Anchor₃*. Based on the assumption that *Node₂* is only covered at *Anchor₁* and *Anchor₂*, *Node₂* is able to calculate two TDOAs: $t_4 - t_3$ and $t_7 - t_6$. These two measurements are enough to position *Node₂* using a multilateration TDOA algorithm. Note that the TDOA position calculated at *Node₂* involves hearing just two access points, which makes it possible

for positioning to take place where TOA techniques would be ineffective. The only datum needed by MS_2 in addition to the TDOA measurements is the position of $Node_1$. This information can be supplied by the unknown node once it becomes settled (e.g., by broadcasting), or it can be estimated in the passive-TDOA node. Simulation analysis demonstrates that, under line-of-sight (i.e., visibility between nodes), the positioning error achieved by this algorithm is often below 1.4 times the error achieved by the 2-way TOA. Better results are achieved if the technique is deployed under non-line-of-sight conditions, providing figures similar to those achieved by the 2-way TOA technique (Martin-Escalona & Barcelo-Arroyo, 2008). This behaviour is especially relevant for location since non-line-of-sight is the usual condition for location-system operation; hence, the best performance is desired in such scenarios.

5.2 Autonomous passive TDOA: TOA position estimated

One of the most constraining requirements of the passive-TDOA algorithm is the position of the 2-way TOA node. Supplying this information in location procedures where position has been requested by a third party should not involve additional changes in the 2-way TOA algorithm and could be considered the final step for the passive-TDOA algorithm. However, the same does not apply to services in which the user requests his or her own position.

Although the impact of supplying the 2-way TOA positions on the capacity of the network is expected to be small, it does become necessary to define a protocol that guarantees the supply of TOA positions once they have been computed, so that passive-TDOA nodes can figure out their own locations. This protocol could involve some modifications in the 2-way TOA algorithm, which is not a desirable fact. OMA SUPL (OMA, 2008) can be used for such purposes, but security issues need to be addressed before its implementation (e.g., positions should not identify users).

The algorithm initially proposed for passive-TDOA has been modified to cope with TOA position supplying. Accordingly, two operational modes have been defined: *assisted* and *autonomous*. The former consists of the algorithm as defined in the previous sections. The *autonomous* operational mode allows positions of TOA and passive-TDOA nodes to be jointly-computed in the passive-TDOA node, in a passive fashion.

There are several benefits to computing the TOA and passive-TDOA positions jointly. The first one is that the passive-TDOA algorithm does not depend on supplying the TOA position. Accordingly, any 2-way TOA algorithm could be used together with the passive-TDOA algorithm with only slight changes. Furthermore, the passive-TDOA nodes will compute their own position and the position of the 2-way TOA nodes, which gives way to approaches for improving accuracy. The autonomous passive-TDOA algorithm becomes especially interesting in scenarios in which an application needs to locate all the users in the network. Nodes report their positions, as well as the positions of 2-way TOA nodes estimated in the passive-TDOA nodes, and then the application can use the redundancy of positions to improve the accuracy of the 2-way TOA nodes.

The *Autonomous* mode of the passive-TDOA algorithm is based on a usual feature of 2-way TOA algorithms: the redundancy on RTTs. These algorithms tend to measure several RTTs involving the same landmark in order to reduce errors caused by the measurement system and radio channel, hence improving the accuracy. *Autonomous* mode uses two consecutive RTTs on the same landmark to estimate two TDOAs (as defined in the case of the *normal* operational mode), as well as the RTT being measured at the TOA node. As expected, the

RTT estimate in the passive-TDOA node will be noisier than the ones made in the TOA node, but it is expected to be accurate enough to allow the passive-TDOA algorithm to compute its own position.

Fig. 4 shows the procedure that constitutes the *autonomous* operational mode of the passive-TDOA algorithm. The explanation is based on the scenario proposed in Fig. 3, but reduced to one 2-way TOA node (i.e., *Node₁*), a landmark (i.e., *Landmark*), and a passive-TDOA node (i.e., *Node₂*). As explained for the case of the *assisted* operational mode, *Node₁* starts a ping-fashion procedure to compute the range between the landmark and itself. As a result, the RTT_1 (i.e., $t_2 - t_1$) is measured. Consequently, $TDOA_1$ is deducted from the ping procedure as $t_4 - t_3$. Until this point, the procedure is exactly the same as that presented in the case of *assisted* mode.

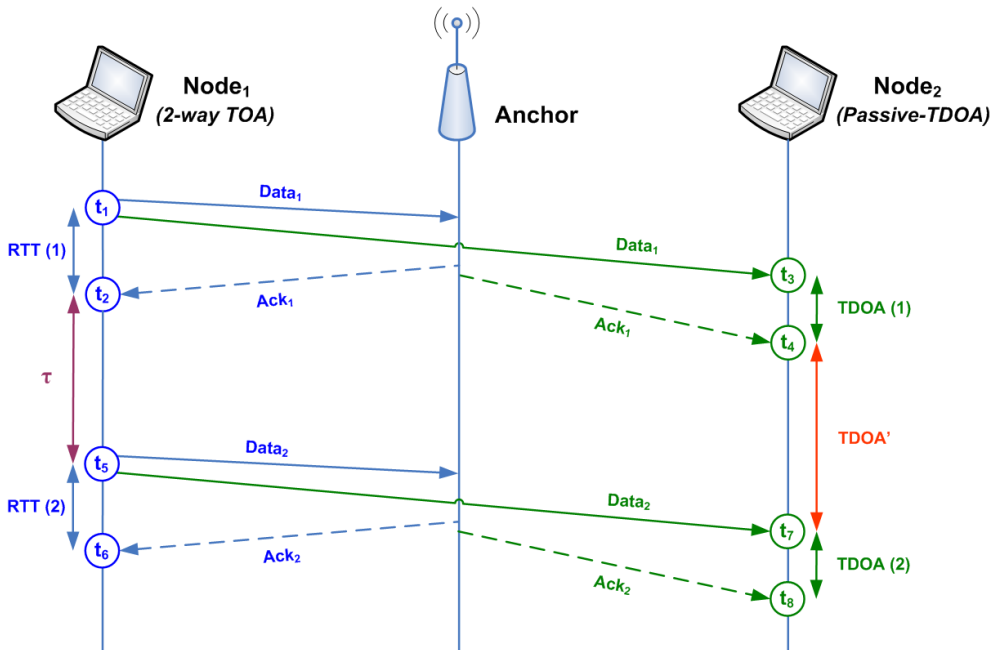


Fig. 4. Flow diagram for the *autonomous* operational mode of passive-TDOA algorithm

Then, it is assumed that *Node₁* starts a new 2-way TOA procedure involving the same entities (i.e., *Landmark* and *Node₂*) after a predefined time (τ), which is known by all nodes in the network. This new procedure provides new estimates for *Node₁* and *Node₂*, i.e., RTT_2 and $TDOA_2$, respectively. Furthermore, *autonomous* mode benefits from this redundancy in the measurements by using it to estimate the ranging between *Landmark* and *Node₁* in *Node₂*. This can be done by simply measuring a new time-difference in *Node₂*: $TDOA'$. This time-difference corresponds to the difference between the arrival time to *Node₂* of the first TOA response and the second TOA request messages, subtracting the time elapsed between the two ping processes (i.e., $t_7 - t_4 - \tau$ in Fig. 4), which is assumed to be known by all nodes in the network. This information, together with the $TDOA_1$ and $TDOA_2$ measurements, allows the

network to deduce the ranging information concerning *Node₂* and the *Landmark*. The formulation starts from

$$T\langle i, j \rangle = R\langle i, j \rangle + R\langle j, k \rangle - R\langle i, k \rangle, \quad (6)$$

which computes the TDOA from the ranging information. R in Equation (6) computes the distance between two nodes, T stands for the distance-difference, and subscripts i, j, k stand for the TOA node, the landmark, and the passive-TDOA node, respectively (as in the rest of the document). According to Equation (6) and the scenario presented in Fig. 4, two TDOAs are computed as

$$T\langle i, j \rangle^1 = R\langle i, j \rangle^1 + R\langle j, k \rangle^1 - R\langle i, k \rangle^1 \quad (7)$$

and

$$T\langle i, j \rangle^2 = R\langle i, j \rangle^2 + R\langle j, k \rangle^2 - R\langle i, k \rangle^2, \quad (8)$$

where the superscript indicates the ping procedure involved in the measurement. These two TDOAs (in distance) are then averaged (under the assumption of providing the same QoS) as

$$T\langle i, j \rangle = \frac{1}{2} (T\langle i, j \rangle^2 + T\langle i, j \rangle^1). \quad (9)$$

According to its definition, $TDOA'$ is computed as

$$TDOA' = T\langle j, i \rangle = R\langle i, j \rangle^1 + R\langle i, k \rangle^2 - R\langle j, k \rangle^1. \quad (10)$$

Under the assumption of noiseless measurements, TOA ranging can be estimated in the passive-TDOA node as

$$R\langle i, j \rangle = \frac{1}{2} (T\langle i, j \rangle + T\langle j, i \rangle). \quad (11)$$

Finally, once $R\langle i, j \rangle$ is estimated, the same algorithm as used in the *assisted* mode is used to compute the position of the passive-TDOA node.

Simulation results indicate that estimating the ranging of the 2-way TOA node results in less accurate positions, as expected. However, under non-line-of-sight conditions, which are the usual case, passive-TDOA provides positions with only 20% more error than the 2-way TOA, with the benefit of no traffic injection. This is especially relevant for group location, i.e., those applications that involve more than a single location process.

5.3 Applications of the passive-TDOA algorithm

The passive-TDOA algorithm has multiple applications in the field of location. The main one has been discussed above and consists of allowing an unknown node to be positioned without injecting traffic into the network. Therefore, the load due to positioning is reduced, and the network throughput remains available for other services. This feature is essential for location algorithms since it improves scalability, which is especially essential for the location platform in the ad hoc environment.

Another application of passive TDOA is the capability of the algorithm to position unknown nodes in environments where TOA techniques cannot. For instance, Fig. 3 shows how passive-TDOA is able to position a customer who only has two access points in sight. Under the same conditions, the TOA technique is not able to provide a location, since this technique requires at least three transmitters (even more depending on the algorithm) to perform a 2D trilateration. The passive-TDOA algorithm would be able to go further in positioning under constrained scenarios. In fact, this technique would be able to compute the position of a station with just one access point in sight, whenever enough settled nodes are in sight and their positions are known. This makes the passive-TDOA algorithm a very interesting solution for positioning under extreme conditions (e.g., scenarios in which there is interference), eventually mitigating the impact of some access points being down (e.g., due to maintenance, fire damage, etc.). Since passive-TDOA can work with fewer landmarks, it helps the system continue to offer location-based services in those circumstances where TOA is not able to provide some positions.

Passive-TDOA can also be used to improve the accuracy of TOA positions. The passive-TDOA node is able to estimate its position and the positions of other TOA nodes involved as long as enough measurements are available. Subsequently, these positions can be coupled to reduce the noise and improve the final accuracy. Furthermore, this operational mode can be used to locate unknown nodes with no location capabilities at all, as explained in more detail previously in this document.

All these applications of the passive-TDOA algorithm give rise to dramatic improvements in the scalability of the system, since more customers can be located while only a few TOA positioning processes are running. Note, however, that all the applications of the passive-TDOA algorithm depend on their expected accuracy, since a large error in the positions computed by passive-TDOA will make these positions useless, and, therefore, system scalability will not increase at all. This work analyses the accuracy expected from passive-TDOA under several conditions and compares it with the positioning error achieved by a regular 2-way TOA algorithm.

It must be noted that, even though the algorithms are addressed to ad hoc networks, they can be implemented in other networks based on infrastructure, such as those operating under the standard IEEE 802.11. In these networks, the anchors would be the access points, and the terminal nodes would be the 802.11 clients. This demonstrates the capabilities of the algorithm presented and the wide range of applications for which it can be used.

6. Conclusion

This chapter presents the positioning problem in the ad hoc context. According to the current literature, ad hoc algorithms are predominantly focused on this topic, since location techniques used for other cellular technologies remain valid in the ad hoc environment. The main algorithms proposed for ad hoc positioning are presented, giving special attention to the passive-TDOA. This algorithm is proposed to improve the scalability of 2-way TOA solutions and while at the same time providing good accuracy figures. Two operational modes are explained in detail, and the main applications for this algorithm are discussed.

7. References

- Stojmenovic, I. (2002). Position-based routing in ad hoc networks, *IEEE Communications Magazine*, vol. 40, July 2002, pp. 128-134, ISSN: 0163-6804.
- Bourkerche, A.; Oliveira, H.; Nakamura, E. & Loureiro, A. (2007). Localization Systems for Wireless Sensor Networks, *IEEE Wireless Communications*, Vol. 14, December 2007, pp. 6-12, ISSN: 1536-1284.
- Seidel, S.Y. & Rapoport, T.S. (1992). 914 MHz path loss prediction Model for Indoor Wireless Communications in Multi-floored buildings. *IEEE Trans. on Antennas & Propagation*, vol. 40, no. 2, February 1992, pp. 207-217, ISSN: 0018-926X.
- Wang, Y.; Jia, X. & Lee, H.K. (2003). An Indoors Wireless Positioning System Based on Wireless Local Area Network Infrastructure, *Proceedings of the 6th International Symposium on Satellite Navigation*, pp. 1-13, University of New South Wales, Melbourne (Australia), July 2003.
- Lassabe, F.; Canalda, P.; Chatonnay, P. & Spies, F. (2005). A Friis-based calibrated model for WiFi terminals positioning, *Proceedings of 6th IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WOWMOM)*, pp. 382-387, ISBN: 0-7695-2342-0, Giardini Naxos, Messina (Italy), June 2005.
- Bahl, P. & Padmanabhan, V. (2000). Radar: An In-Building RF-based User Location and Tracking System, *Proceedings of the 19th IEEE Conference on Computer Communications (INFOCOM)*, vol. 2, pp. 775-784, ISBN: 0-7803-5880-5, Tel Aviv (Israel), March 2000.
- Kotani, A.; Hannikainen, M.; Leppakoski, H. & Hamalainen, T.D. (2003). Positioning with IEEE 802.11b wireless LAN. *Proceedings of the 14th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, vol. 3, pp. 2218-2222, ISBN: 0-7803-7822-9, Beijing (China), September 2003.
- Ciurana, M.; Barcelo-Arroyo F. & Izquierdo, F. (2007). A ranging system with IEEE 802.11 data frames, *Proceedings of the 1st IEEE Radio and Wireless Symposium*, pp. 133-136, ISBN: 1-4244-0445-2, Long Beach, California (USA), January 2007.
- Ibraheem, A. & Schoebel, J. (2007). Time of Arrival Prediction for WLAN Systems Using Prony Algorithm, *Proceedings of the 4th Workshop on Positioning, Navigation and Communication (WPNC)*, pp. 29-32, ISBN: 1-4244-0871-7, Hannover (Germany), March 2007.
- Reddy, H. & Chandra, G. (2007). An improved Time-of-Arrival estimation for WLAN-based local positioning, *Proceedings of the 2nd International Conference on Communication Systems software and middleware (COMSWARE)*, pp. 1-5, ISBN: 1-4244-0613-7, Bangalore (India), January 2007.
- 3GPP. (2004). Functional stage 2 description of LCS; Technical Specification Group Services and System Aspects, 3GPP TS 23.271, version 6.8.0, June 2004.
- 3GPP. (2002). Location Services (LCS); Functional description; Stage 2, 3GPP TS 03.71, version 8.7.0, September 2002.
- Pages-Zamora, A., Vidal, J. & Brooks, D.H. (2002). Closed-form solution for positioning based on angle of arrival measurements. *Proceedings of the 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, vol. 4, pp. 1522-1526, ISBN: 0-7803-7589-0, Pavilhão Atlântico, Lisbon (Portugal), September 2002.
- Savvides, A.; Han, C.C. & Srivastava, M.B. (2001). Dynamic Fine-Grained Localization in Ad-Hoc Wireless Sensor Networks. *Proceedings of the 7th ACM International Conference on Mobile Computing and Networking*, pp. 166-179, ISBN: 1-58113-422-3, Rome (Italy), July 2001.

- Sun, G.; Chen, J.; Guo, W. & Ray-Liu, K.J. (2005). Signal processing techniques in network-aided positioning: a survey of state-of-the-art positioning designs, *IEEE Signal Processing Magazine*, vol. 22, no. 4, July 2005, pp. 12-23, ISSN: 1053-5888.
- Doherty, L.; Ghaoui, L.E. & Pister, K.S.J. (2001). Convex Positioning Estimation in Wireless Sensor Networks. *Proceedings of the 20th IEEE Conference on Computer Communications (INFOCOM)*. vol. 3, pp. 1655-1663, ISBN: 0-7803-7016-3, Anchorage, AK (USA), April 2001.
- Bulusu, N.; Heidemann, J. & Estrin, D. (2000). GPS-less low cost outdoor localization for very small devices, *IEEE Personal Communications*, vol. 7, October 2000, pp. 28-34, ISSN: 1070-9916.
- Römer, K. (2003). The Lighthouse Location System for Smart Dust, *Proceedings of the 1st ACM/USENIX international conference on Mobile systems, applications and services*, pp. 15-30, San Francisco, California (USA), May 2003.
- Niculescu, D. & Nath, B. (2003). DV based positioning in ad hoc networks. *Springer Journal of Telecommunication Systems*, vol. 22, no. 1-4, January 2003, pp. 267-280, ISSN: 1018-4864.
- Capkun, S.; Hamdi, M. & Hubaux, J. (2001). GPS-free positioning in mobile ad-hoc networks. *Proceedings of the Hawaii International Conference on Systems Sciences (HICSS)*, pp. 1-10, ISBN: 0-7695-0981-9, Maui, Hawaii (USA), January 2001.
- Niculescu, D. & Nath, B. (2003). Localized Positioning in Ad Hoc Networks. *Elsevier Ad Hoc Networks*, vol. 1, no. 2-3, September 2003, pp. 247-259.
- Nagpal, R.; Shrobe, H. & Bachrach, J. (2003). Organizing a Global Coordinate System from Local Information on an Ad Hoc Sensor Network. *International Workshop on Information Processing in Sensor Networks (IPSN)*. Published as Lecture Notes in Computer Science, LNCS 2634, pp. 333-348. ISSN 0302-9743.
- Fu, Y.; Liu, H.; Qin, J. & Xing, T. (2006). The localization of wireless sensor networks nodes based on DSSS. *Proceedings of the IEEE International Conference on Electro/Information Technology*. pp. 465-469. ISBN: 0-7803-9592-1, East Lansing, MI (USA), May 2006.
- He, T.; Huang, C.; Blum, B.; Stankovic, J. & Abdelzaher, T. (2003). Range-Free Location Schemes for Large Scale Sensor Networks, *Proceedings of the 9th annual international conference on Mobile computing and networking (MOBICOM)*, pp. 81-95, ISBN: 1-58113-753-2, San Diego, CA (USA), September 2003.
- Yang, S; Kang, D.; Namgoong, Y.; Choi, S. & Shin, Y. (2008). A simple asynchronous UWB position location algorithm based on single round-trip transmission. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*. vol. E91-A, no. 1, January 2008, pp. 430-432, ISSN:0916-8508.
- Martin-Escalona, I. & Barcelo-Arroyo, F. (2008). A New Time-Based Algorithm for Positioning Mobile Terminals in Wireless Networks, *EURASIP Journal on Advances in Signal Processing*, Hindawi Publishing Corporation, vol. 2008, pp. 1-10. ISSN: 1687-6172.
- OMA (2008), Secure User Plane Location Architecture (SUPL), Open Mobile Alliance, [Online], <http://www.openmobilealliance.org>.

Location Tracking Schemes for Broadband Wireless Networks

Po-Hsuan Tseng and Kai-Ten Feng

*Department of Communication Engineering, National Chiao Tung University
Taiwan, R.O.C.*

1. Introduction

In order to enable the delivery of last mile wireless broadband access, the IEEE 802.16-2004 standard (IEEE Std 802.16-2004, 2004) for the wireless metropolitan area networks (WMAN) is designed to fulfill various demands for higher capacity, higher data rate, and advanced multimedia services. Furthermore, the IEEE 802.16e standard (IEEE Std 802.16e-2005, 2006) enhances the original IEEE 802.16-2004 specification by addressing the mobility issues for the mobile stations (MSs). Recently, the IEEE 802.16-2009 standard (IEEE Std 802.16-2009, 2009) has been specified as an integrated version of the IEEE 802.16 specification by the IEEE 802.16 maintenance task group. The IEEE 802.16-2009 standard is as known as the revision of IEEE 802.16-2004 and consolidates material from IEEE 802.16e-2005, IEEE 802.16-2004/Cor1-2005, IEEE 802.16f-2005, and IEEE 802.16g-2007. In order to fulfill the requirement of the E911 phase II requirement advanced by Federal Communications Commission, the location-based services (LBSs) (Perusco & Michael, 2007) are considered one of the key functions of the IEEE 802.16-2009 standard. Moreover, for fulfilling the resource management purpose, location update is also essential to other numerous functions such as the paging processes. Based on the IEEE 802.16 standard, it is required to provide satisfactory location estimation performance under a wide-range of MS's moving speeds. Location tracking is designed as one of the options to provide feasible estimation performance in order to trace the MS's moving behaviors. However, there are several issues required to be considered before enabling the location tracking scheme within IEEE 802.16 system. It is noted that timing information, i.e., the time-difference-of-arrival (TDOA) measurements, from at least four base stations (BSs) is required to perform a two-dimensional location estimation and tracking for an MS. With the stringent synchronization requirement for the IEEE 802.16 OFDMA-based system, the frequent TDOA measurements with other neighbor (a.k.a. non-serving) BSs can be time-consuming and impractical processes for location estimation and tracking. It is a waste of the bandwidth to scan for the neighbor BSs frequently, especially under broadband wireless communication.

In this book chapter, two location tracking schemes are proposed to alleviate the problem that requires frequent connections between the MS and the neighbor BSs. The kinematics-assisted location tracking (KLT) scheme adopts the kinematic relationship to estimate the MS's location at the time instant with unavailable neighbor BSs; while the geometry-assisted location tracking (GLT) algorithm utilizes the geometric constraints for the prediction of MS's position. The two schemes are proposed to interpolate the location of an MS between two direct

location estimations from the MS to the neighbor BSs. It will be shown in the simulations that both proposed schemes can provide feasible performance with significantly reduced communication overhead.

2. TDOA Measurements of IEEE 802.16 Network

To illustrate for the TDOA measurement scheme, the procedures related to the basics of IEEE 802.16 network operations are introduced first. The IEEE 802.16 adopts OFDMA based technique, which implies the MS and the serving BS should synchronize in both time and frequency domain in order to receive the data correctly. Providing that an MS intends to join an IEEE 802.16 network, it conducts initial ranging with the serving BSs to obtain the synchronization parameter. In order to indicate which time slots for the MS to receive and transmit data, the downlink map (DL-MAP) and the uplink map (UL-MAP) are designed respectively. First the MS should listen to the BS's broadcast message to capture the ranging opportunity in the UL-MAP. The MS conducts contention based initial ranging to obtain related parameter and then is granted to join the network. After the MS establishes the link with serving BS, the MS performs periodic ranging in order to maintain the synchronization property while the MS may move to different place or the channel condition may vary at different time. It is noted that the periodic ranging is one of the routine process which is not considered as an overhead in this work. The ranging scheme is conducted by the MS sending assigned CDMA codes to serving BS to measure the distance related parameter, e.g. timing advance value. It is noticed that the timing advance value is the round trip time between the MS and the serving BS. The timing advance information is utilized to reserve the proper timing for the uplink transmission. By performing the ranging scheme, the serving BS measures the timing adjustment according to previously recorded timing advance information to update the distance between MS and serving BS. In order to support for the MS's mobility, the scanning scheme is specified for the MS to perform ranging with neighbor BSs. With the negotiation between serving BS and neighbor BSs, the scanning scheme can directly obtain ranging opportunity without contention. However, the MS is unavailable to serving BS while it performs scanning scheme with the neighbor BSs.

In the IEEE 802.16-2009 standard, two types of TDOA are specified as the downlink TDOA (D-TDOA) and the uplink TDOA (U-TDOA) where the measurements are performed at the MS and the BS respectively. These two schemes are based on ranging and scanning scheme to obtain time difference between the serving BS and the neighbor BSs. Due to the superior timing resolution obtained from the U-TDOA measurement at the BS (i.e., around 25 to 50 nanoseconds measured at the BS compared to microseconds at the MS), the U-TDOA measurement is chosen in this paper for achieving better location estimation accuracy. Moreover, as described in the standard, the general U-TDOA method is adopted while the frequency reuse factor is not equal to one which is considered as a more general case. Several assumptions are specified as follows: (a) both the serving and neighbor BSs are operating with the same frame size; (b) the frames at both the serving and the neighbor BSs are synchronized; and (c) the MS can communicate with both the serving and the neighbor BSs.

Fig. 1 illustrates the timing diagram of the general U-TDOA measurement in IEEE 802.16-2009. The MS ranges sequentially with the serving BS and neighbor BSs. It is noted that the second frame of the MS is operating on the same frequency with the serving BS and the third is the same with the neighbor BS. The serving BS (i.e., the 1st BS) and the neighbor BS (i.e., the 2nd BS) measure the timing adjustment t_{adj_1} and t_{adj_2} respectively, and the neighbor BS reports t_{adj_2} to the serving BS. The timing advance value t_{adv} remains the same since the MS does

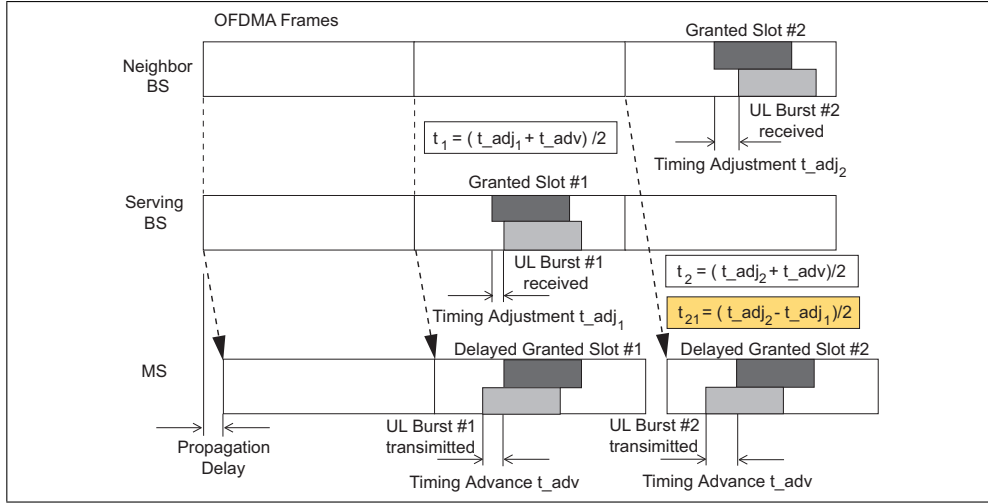


Fig. 1. Timing Diagram of the general U-TDOA measurement proposed in the IEEE 802.16-2009 standard.

not make any timing adjustments while conducting ranging with both the serving and the neighbor BSs. Therefore, serving BS calculates the time difference $t_{21} = (t_{adj_2} - t_{adj_1}) / 2$ and the difference of the MS's distance to the serving BS and neighbor BS is obtained by multiplying this difference by the speed of light.

Fig. 2 illustrates the message exchange sequences for the general U-TDOA measurement. The serving BS requests neighbor BSs to assign a dedicated ranging opportunity for the MS. The MS will first conduct ranging with the serving BS in order to perform the U-TDOA measurement. Through the ranging response (RNG-RSP) message, the serving BS will assign a Rendezvous time, a CDMA code, and a Tx opportunity offset for the MS. It is noted that the Rendezvous time specifies the frame in which the BS transmits an UL-MAP containing the definition of the dedicated ranging region. As the Rendezvous time is expired, the MS will transmit the allocated CDMA code within the regular ranging region. The time-of-arrival (TOA) measurement $t_{1,k}$ performed at the serving BS (BS_1) is obtained as an average value of both the timing adjustment $t_{adj_{1,k}}$ from the measurement and the timing advance t_{adv_k} acquired from the periodic ranging, i.e., $t_{1,k} = (t_{adj_{1,k}} + t_{adv_k}) / 2$ where the subscript denotes the k th time step.

Moreover, the dedicated ranging process is repeated through the neighbor BS (e.g. BS_2) by sending the mobile scanning response (MOB_SCN-RSP) message. The neighbor BS measures $t_{adj_{2,k}}$ and reports the value to the serving BS. The TOA measurement for BS_2 can therefore be acquired as $t_{2,k} = (t_{adj_{2,k}} + t_{adv_k}) / 2$. As a result, the U-TDOA measurement calculated at the serving BS is obtained as $t_{21,k} = (t_{adj_{2,k}} - t_{adj_{1,k}}) / 2$. Similar process can be performed to obtain the timing information from the other neighbor BSs. Nevertheless, with the TDOA measurements, at least four BSs should be involved to perform a 2-D location estimation. There is a significant overhead to perform location tracking scheme as depicted in Fig. 2. In the following section, the proposed location tracking schemes are proposed to alleviate the overhead of frequent ranging with neighbor BSs.

where $\mathbf{x}_k^o = [x_k^o \ y_k^o]^T$ represents the MS's true position at the k th time step, and $\mathbf{x}_{i,k} = [x_{i,k} \ y_{i,k}]^T$ is the location of the i th BS. On the other hand, the relative distance $r_{ij,k}$ from the TDOA measurement $t_{ij,k}$ can be obtained by computing the time difference between the MS w.r.t. the i th and the j th BSs as

$$\begin{aligned} r_{ij,k} &= (r_{i,k} - r_{j,k}) = c \cdot (t_{i,k} - t_{j,k}) = c \cdot t_{ij,k} \\ &= (\zeta_{i,k} - \zeta_{j,k}) + (m_{i,k} - m_{j,k}) + (n_{i,k} - n_{j,k}) \end{aligned} \quad (3)$$

It is noted that in IEEE 802.16-2009 standard each BS equips a GPS in order to achieve time synchronization. Therefore, the relationship $s_{i,k} - s_{j,k} = 0$ is true since the frames at the serving and the neighbor BS are synchronized. As depicted in the previous section for the general U-TDOA method, the TOA measurement (in (1)) is acquired for the purposed of obtaining the TDOA measurement (in (3)).

3.2 Location Estimation and Tracking Algorithms

The main concept for the proposed schemes is to maintain the accuracy for location estimation with a reduced number of dedicated ranging (i.e., the general U-TDOA measurement in Figs. 1 and 2) between the MS and the BSs. The increased number of dedicated ranging with the neighbor BSs can result in unsatisfactory communication performance between the MS and its serving BS, e.g. with degraded scheduling performance for realtime applications. It is noted that the dedicated ranging state indicates the state for the MS to conduct the general U-TDOA method with the serving and the neighbor BSs as shown in Fig. 2. Without communications between the MS and the BSs, on the other hand, the non-dedicated ranging state (i.e., the general U-TDOA measurement is not available) is defined as the state in which the MS's location information is estimated and predicted by the proposed KLT and GLT schemes, which will be explained in the next two subsections.

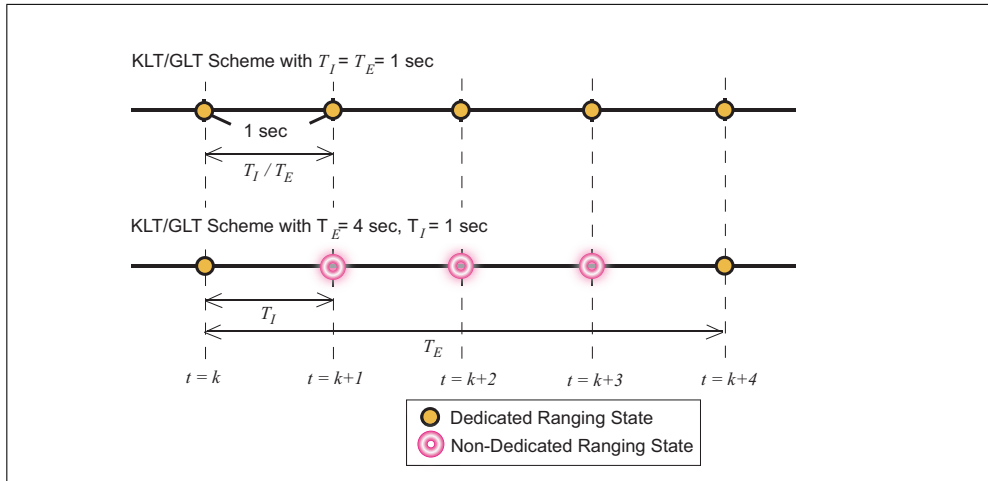


Fig. 3. The timing diagram for the relationship between the dedicated and non-dedicated ranging states.

Fig. 3 illustrates the relationship between the dedicated and the non-dedicated ranging states. The location estimation period (T_E) is defined as the time duration between two ded-

icated ranging states. On the other hand, the location information period (T_I) is designed as $T_I = T_E/m$ with $m \geq 1$, which represents the time interval either between two non-dedicated ranging states or between a dedicated and a non-dedicated states. In other words, T_I is defined as the interleaved period where the MS's location information becomes available, either obtained from the general U-TDOA method of the proposed KLT/GLT scheme. In other word, the U-TDOA method is utilized at the dedicated ranging state; while the KLT/GLT scheme interpolates the position information at the non-dedicated ranging state. Moreover, the sampling time Δt is denoted as the time interval between the k th and the $(k-1)$ th time steps as in (1) to (3). It is selected the same as the location information period, i.e. $\Delta t = T_I$.

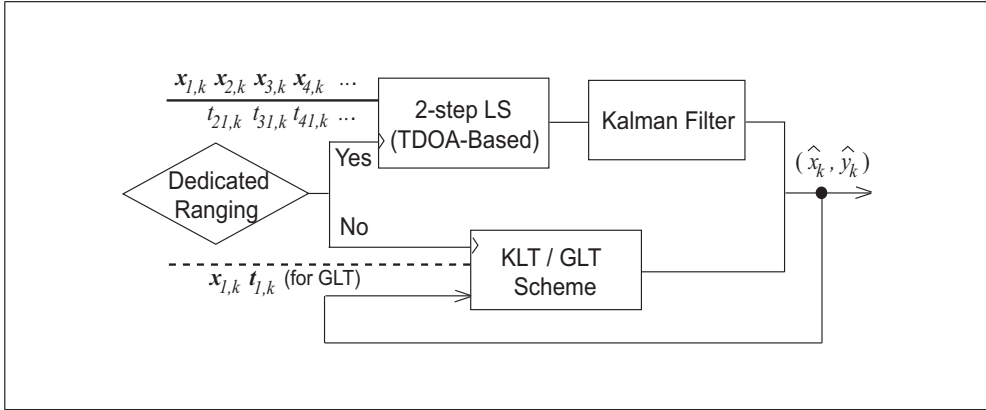


Fig. 4. The schematic diagram of the proposed KLT and GLT algorithms.

Fig. 4 illustrates the schematic diagram of the proposed KLT and GLT algorithms. Either the dedicated or the non-dedicated ranging can happen for obtaining the MS's estimated position $\hat{\mathbf{x}}_k = [\hat{x}_k \hat{y}_k]^T$. For the dedicated ranging case, the cascaded location tracking (CLT) scheme as proposed in (Chen & Feng, 2005) is exploited. The CLT algorithm is cascaded by two functional components, i.e. the two-step least square (LS) method for location estimation and the Kalman filtering technique for location tracking. The two-step LS method obtains the initial location estimation $\hat{\mathbf{x}}_k^{LS} = [\hat{x}_k^{LS} \hat{y}_k^{LS}]^T$ from the TDOA measurement input $t_{ij,k}$ (in (3)) within two computing iterations. Furthermore, the Kalman filter is utilized to smooth out and trace the estimation errors and finally acquires the MS's estimated location $\hat{\mathbf{x}}_k = [\hat{x}_k \hat{y}_k]^T$. On the other hand, since the TDOA measurements are not available during the non-dedicated ranging state, two schemes are proposed to substitute the functionality of the two-step LS method as follows.

3.2.1 Kinematics-Assisted Location Tracking (KLT) Scheme

With the unavailability of the TDOA measurements during the non-dedicated state, the KLT scheme is utilized to adopt the predicted information from the output of the Kalman filter. Considering a three-states linear model, the MS's position, velocity, and acceleration can be estimated via the Kalman filter as $\hat{\mathbf{z}}_k = [\hat{\mathbf{x}}_k \hat{\mathbf{v}}_k \hat{\mathbf{a}}_k]^T$, where $\hat{\mathbf{x}}_k = [\hat{x}_k \hat{y}_k]^T$, $\hat{\mathbf{v}}_k = [\hat{v}_{x,k} \hat{v}_{y,k}]^T$, and $\hat{\mathbf{a}}_k = [\hat{a}_{x,k} \hat{a}_{y,k}]^T$. Assuming that the state vector $\hat{\mathbf{z}}_k$ is available either via the dedicated or non-dedicated ranging, the next non-dedicated states $\hat{\mathbf{z}}_{k+1}$ at the $(k+1)$ th time instant can be acquired by utilizing the feedback information from the output of the Kalman filter at time k .

By adopting the updates from the kinematic relationship, the MS's predicted position $\hat{\mathbf{x}}_{k+1}$ at the $(k+1)$ th time step can be acquired as

$$\hat{\mathbf{x}}_{k+1}^{KLT} = \hat{\mathbf{x}}_k + \hat{\mathbf{v}}_k \cdot \Delta t + \frac{1}{2} \cdot \hat{\mathbf{a}}_k \cdot \Delta t^2 \quad (4)$$

where Δt is the sampling interval as $\Delta t = T_I$. The location estimation and tracking at the non-dedicated ranging state can therefore be performed.

3.2.2 Geometry-Assisted Location Tracking (GLT) Scheme

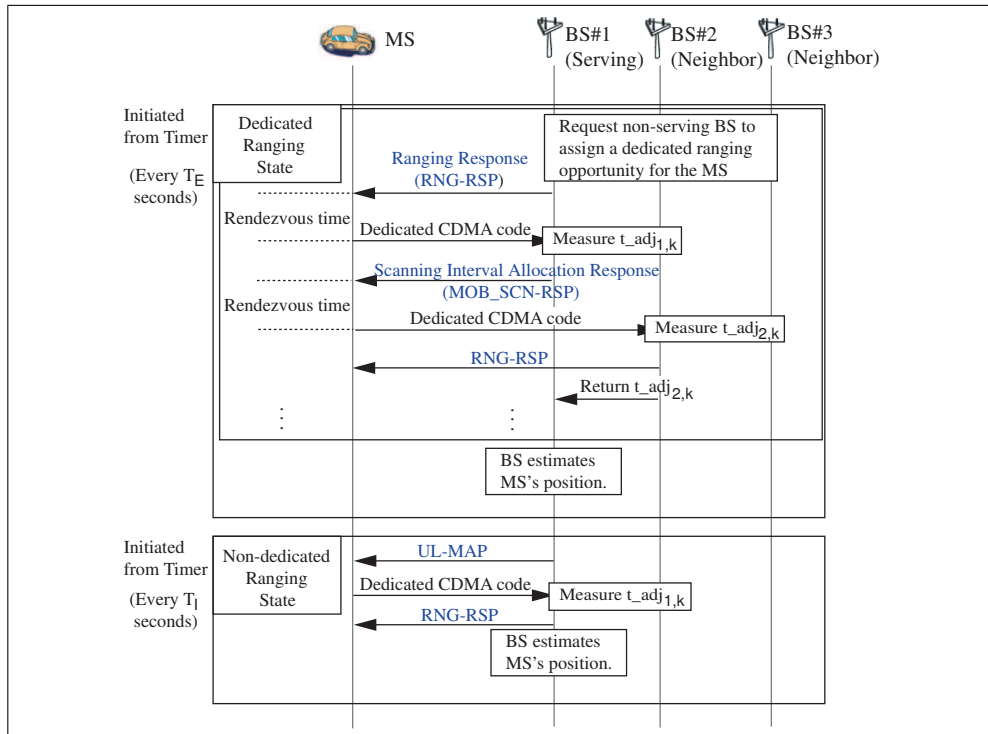


Fig. 5. The message exchange sequences of propose GLT schemes.

Similar to the KLT algorithm as described in the previous subsection, the proposed GLT scheme is utilized to provide location estimation during the non-dedicated ranging state. The concept of the GLT algorithm is to utilize the frequent periodic ranging between the MS and the serving BS. Based on the periodic ranging, the relative distance $r_{1,k}$ between the serving BS and the MS can be obtained from the corresponding TOA measurements $t_{1,k}$. Fig. 5 depicts the message exchange sequences of the proposed GLT schemes within the IEEE 802.16-2009 network. It is noticed that the flowchart is the same as the general U-TDOA scheme as in Figs. 1 and 2 at the dedicated ranging state. However, at the non-dedicated ranging state, the TOA measurement is obtained through periodic ranging scheme as shown in Fig. 5.

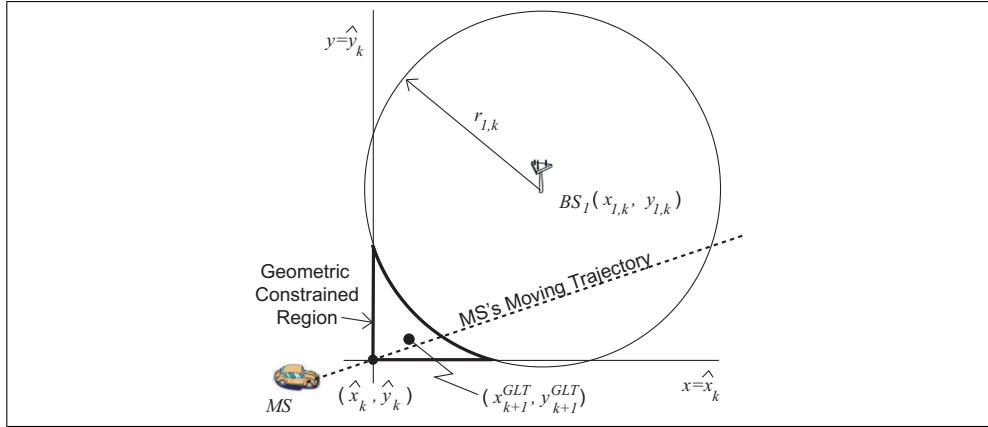


Fig. 6. The schematic diagram of the geometric constraints for the proposed GLT scheme.

As shown in Fig. 6, the circular region can be formed according to the center point $x_{1,k} = [x_{1,k} \ y_{1,k}]^T$ with radius of $r_{1,k}$. Meanwhile, two additional linear equations can be acquired from the feedback of the Kalman filter at the k th time instant, i.e. $x = \hat{x}_k$, $y = \hat{y}_k$. As a result, the two linear and one circular equations can be utilized to provide the geometric constraints for obtaining the MS's position estimation. Based on the constrained region, the LS method is employed to minimize the sum of the square errors for the MS's position. Therefore, the MS's estimated position by adopting the GLT scheme is acquired as

$$\hat{\mathbf{x}}_{k+1}^{GLT} = \mathbf{G} \cdot (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{J} \quad (5)$$

where

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad (6)$$

$$\mathbf{H} = \begin{bmatrix} -2x_{1,k} & -2y_{1,k} & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad (7)$$

$$\mathbf{J} = \begin{bmatrix} r_{1,k}^2 - x_{1,k}^2 - y_{1,k}^2 \\ \hat{x}_k \\ \hat{y}_k \end{bmatrix} \quad (8)$$

4. Performance Evaluation

Simulations are performed to show the effectiveness of the proposed KLT and the GLT schemes. Different noise models (Greenstein et al., 1997) are considered in the simulations in order to represent various environments, including the urban, the suburban, and the rural cases. In the cellular-based network, an exponential distribution is assumed for the NLOS model with the distribution of $p_{n_{i,k}}(v)$ as

$$p_{n_{i,k}}(v) = \begin{cases} \frac{1}{v_{i,k}} e^{\frac{-v}{v_{i,k}}} & v > 0 \\ 0 & v \leq 0 \end{cases} \quad (9)$$

where $v_{i,k} = c \cdot \tau_{i,k} = c \cdot \tau_m \zeta_{i,k}^\varepsilon \omega$ is the RMS delay spread between the i th BS to the MS at the k th time step; τ_m is the median value of $\tau_{i,k}$ whose value depends on various environments, i.e. $\tau_m = 0.4, 0.3$, and 0.1 for urban, suburban, and rural respectively. ε is the path loss exponent which is assumed to be 0.5 . The shadow fading factor ω is a lognormal random variable with zero mean and standard deviation σ_ω chosen as 4 dB in the simulations. Moreover, the measurement noises (i.e. $m_{i,k}$ in (1)) is considered Gaussian-distributed as $\mathcal{N}(0, \sigma_m^2)$ with $\sigma_m = 10$ m. The asynchronous offset (i.e. $s_{i,k}$ in (1)) between the MS and the BS clock time is also assumed to be Gaussian-distributed with $\sigma_s = 7.5$ m.

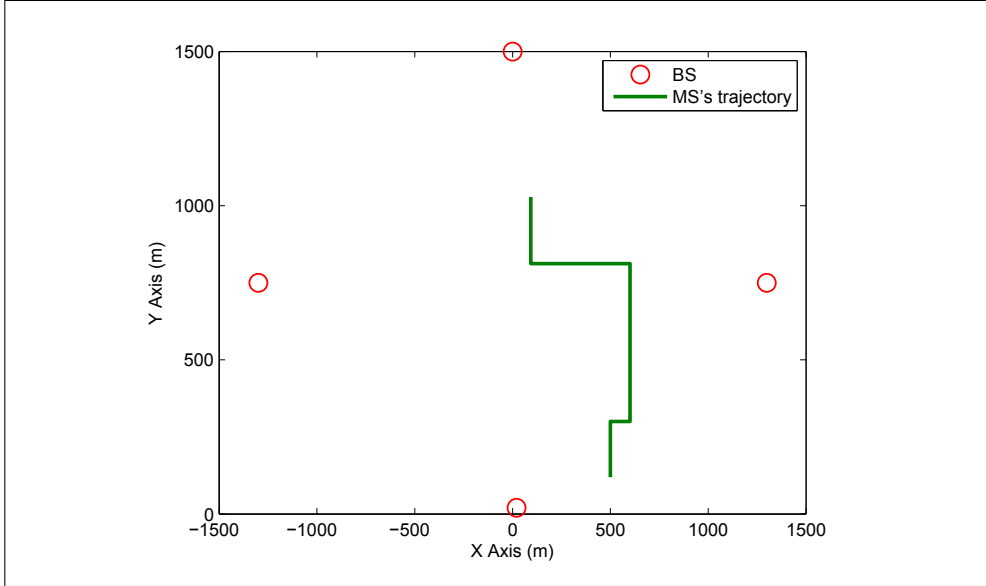


Fig. 7. The geometric layout of the simulation (Green Line: MS's True Trajectory; Red Empty Circles: the Position of the BSs).

Fig. 7 illustrates the MS's trajectory with the Manhattan street scenario in the simulation. The MS's true trajectory is illustrated via the green lines; while the locations of the BSs are represented by the red empty circles as in Fig. 7. The acceleration is designed to vary at time $t = 21, 26, 31, 47, 63, 76$, and 89 sec from $a_k = (a_{x,k}, a_{y,k}) = (0, -1), (4, 0), (-4, 0), (0, 2), (0, -2), (-3, 0)$ to $(3, 0)$ m/sec². The corresponding MS's velocity lies between $(0, 70)$ km/hr.

Fig. 8 shows the performance comparison between the proposed KLT and GLT schemes under the urban environment (with $T_I = 1$ sec and $T_E = 1, 2$, and 4 sec). It is noted that the position error is defined as $P_e = \|\hat{x}_k - x_k^\circ\|$. The case with $T_I = T_E = 1$ sec indicates that the non-dedicated ranging does not exist, which is served as the lower bound of the estimation errors. As shown in the figure, comparably inferior performance is observed with larger values of T_E , which indicates that the dedicated ranging with the TDOA measurement is not frequently available. Moreover, the proposed GLT algorithm outperforms the KLT scheme for each specific case, e.g. around 50 m less of the position error under the case of $T_E = 4$ sec with 67% of position errors.

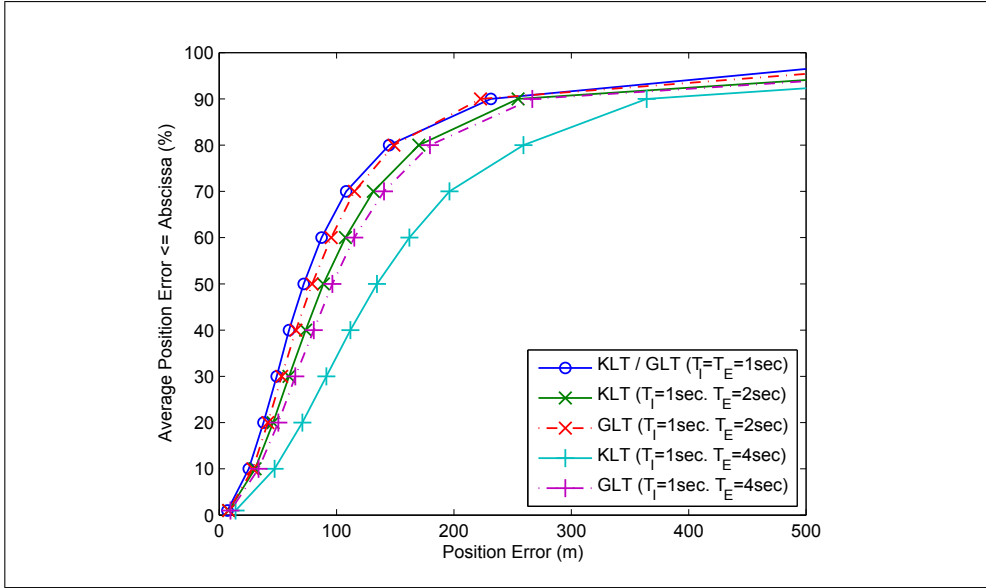


Fig. 8. Performance comparison between the proposed KLT and GLT schemes under the urban environment ($T_I = 1$ sec).

Fig. 9 shows the performance comparison with 67% of position errors between the proposed KLT and GLT schemes under the different noise environment. It is noticed that the KLT scheme performs similar to GLT scheme at the smaller NLOS environment, e.g. $\tau_m = 0.1$ (rural environment). However, the performance of KLT is comparably worse than the performance of GLT scheme under the excessive NLOS errors. The effectiveness of adopting the periodic ranging in GLT scheme can be observed in the case.

In order to evaluate the overhead of the frequent location tracking, serving BS unavailable time (i.e. T_U) is defined for the percentage of the frame communicating between the MS and the neighbor BSs. In the dedicating ranging period, the MS should synchronize with other BSs and therefore the ongoing transmission should be buffered in serving BS. Although the more frequent the dedicating ranging performed brings higher location estimation accuracy, the total usage in the serving BS's point of view would decrease. In the Fig. 2, the MS needs to synchronize to the neighbor BSs first and then performs the ranging process. The following parameter is defined to perform a dedicated ranging with the other BSs:

- T_{syn} : average time to synchronize with the new BS
- T_{rng} : average time to perform range process with a BS

The values for T_{syn} and T_{rng} are chosen as 20 millisecond (msec) and 30 msec in average as reported in (Jiao et al., 2007). As the dedicated ranging specified for the LBS, the time of T_{rng} might be shorter. In terms of the T_{syn} and T_{rng} , the serving BS unavailable time counts in percentage is:

$$T_U = \frac{3 * (T_{syn} + T_{rng})}{T_E} \quad (10)$$

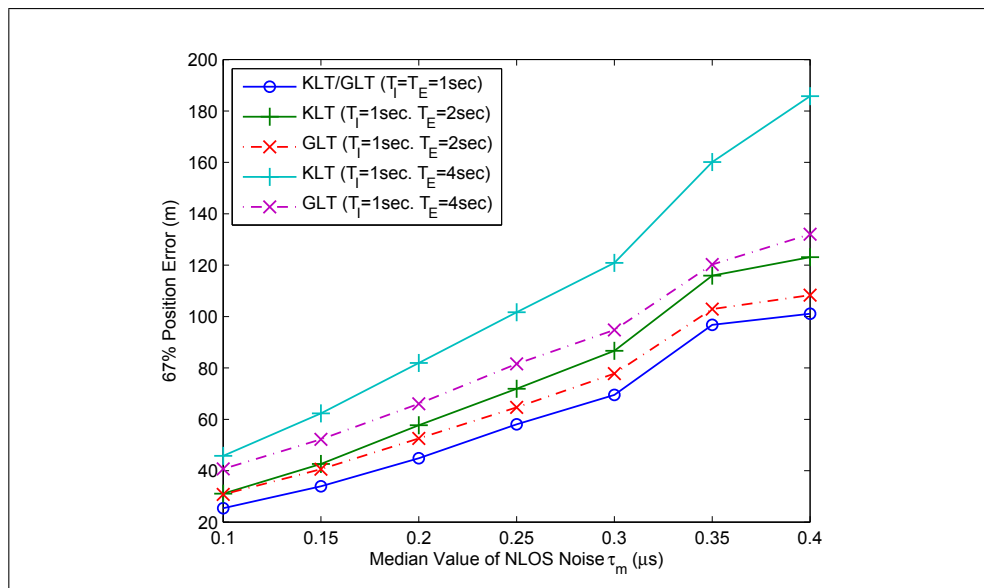


Fig. 9. Performance comparison between the proposed KLT and GLT schemes under different NLOS noise ($T_I = 1$ sec).

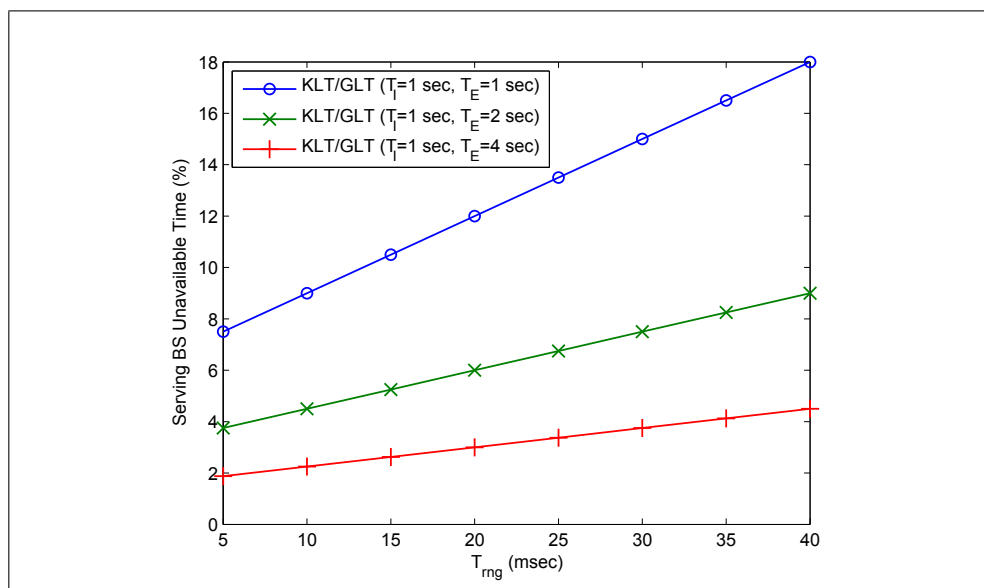


Fig. 10. Serving BS unavailable time via the average time to perform range process.

It is noted that at least three neighbor BSs participate a dedicated ranging. Fig. 10 shows the percentage of the BS unavailable time via the T_{rng} from 5 to 40 msec. As the curves shows that the KLT/GLT scheme with $T_E = 2sec$ has half unavailable time than $T_E = 1sec$. While the overhead of the serving BS is considered as the important factor, comparing to the performance of location estimation in Fig. 9, the KLT/GLT scheme with $T_E = 2sec$ is a better solution with a tradeoff.

5. Conclusion

Two assisted location tracking schemes are proposed in this paper. The schemes are capable of estimating the position, velocity, and acceleration of the MS during the dedicated ranging state. With the non-dedicating ranging state, the assisted methods utilizing the tracking information and the periodic ranging information are proposed. It is shown in the simulation results that the proposed location tracking schemes provide consistent performance and reduces the overhead of the serving BS.

6. References

- Chen, C.-L. & Feng, K.-T. (2005). Hybrid location estimation and tracking system for mobile devices, *Proc. IEEE 61st Vehicular Technology Conference 2005-Spring*, Vol. 4.
- Greenstein, L. J., Erceg, V., Yeh, Y. S. & Clark, M. V. (1997). A new path-gain/ delay-spread propagation model for digital cellular channels, *IEEE Trans. Veh. Technol.* **46**: 477 – 485.
- IEEE Std 802.16-2004 (2004). IEEE standard for local and metropolitan area networks - part 16: Air interference for fixed broadband wireless access systems.
- IEEE Std 802.16-2009 (2009). IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Broadband Wireless Access Systems, pp. 1–2082.
- IEEE Std 802.16e-2005 (2006). IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1.
- Jiao, W., Jiang, P. & Ma, Y. (2007). Fast handover scheme for real-time applications in mobile wimax, *Proc. IEEE International Conference on Communications*, pp. 6038–6042.
- Perusco, L. & Michael, K. (2007). Control, Trust, Privacy, and Security: Evaluating Location-based Services, *IEEE Technol. Soc. Mag.* **26**: 4–16.

Wireless Multi-hop Localization Games for Entertainment Computing

Tomoya Takenaka[†], Hiroshi Mineno[‡] and Tadanori Mizuno[†]

[†]*Graduate School of Science and Technology, Shizuoka University, Japan*

[‡]*Faculty of Informatics, Shizuoka University, Japan*

1. Introduction

Ad-hoc networking capabilities have provided the flexibility needed to construct various types of networks without infrastructure base stations. Emerging products for sensor networks, such as Zigbee (1), use ad-hoc networking capabilities to construct networks. These sensor nodes can construct the network such as in outdoor fields and inside buildings without much effort on establishing the base stations, and monitor neighbor information on area where people usually cannot stay for the monitoring. The technique of ad-hoc networking has been discussed within the Internet Engineering Task Force (IETF) by the Mobile Ad-hoc Network (MANET) Working Group (2). MANET is a promising technique to provide the alternative network infrastructure such as in the disaster case that the existing network infrastructures are destroyed because of fires and earthquakes. The wireless terminals with radio capabilities relay data and deliver to a desired destination. Recently, mobile game consoles with ad-hoc networking capabilities have been produced by companies such as Nintendo (3) and Sony Computer Entertainment (SCE) (4). Networking capabilities play an important role in enabling multiple players to join together to play games. Since the ad-hoc networking technique is independent of the infrastructure network, it is easy for a player to join a game through a wireless network. To utilize this functionality, some games using ad-hoc networking capabilities have been developed. However, the games released thus far only use ad-hoc networking capabilities for joining the game.

We have developed two wireless multi-hop localization games with ad-hoc networking capabilities, and have presented several initial results in (29). The proposed games, a war game and a tag game, are based on classical field games. Players use mobile game consoles with ad-hoc networking capabilities to move around a field. The games use wireless multi-hop localization to estimate node positions. Players on one team jointly establish an ad-hoc network to estimate their positions and compete for positioning accuracy with the other team. We used a previously developed multi-hop localization technique called ROULA (28). We used simulation to evaluate the multi-hop localization games and analyze their characteristics. We found that node velocity and obstruction position controlled the win rate for the games, and maintaining connectivity and local rules led to higher win rates for the games. The results revealed that the proposed games worked well as localization applications using the ad-hoc networking capabilities.

The purpose of this paper is to present new localization applications using ad-hoc networking capabilities. The concept behind multi-hop localization games is presented. Wireless multi-hop localization games are evaluated to find out how well localization-based games with ad-hoc networking capabilities perform in a simulation. These main results obtained from the simulation are summarized as follows.

- The win rate for the games depends on node velocity and obstruction position.
- A higher connectivity constraint leads to a higher win rate for the games.
- Enforcing the local rule of the death penalty enables the win rate to be controlled for the games.

We will next describe the multi-hop localization technique and the current state of mobile games. The localization technique of ROULA is reviewed in Section 3. Section 4 describes our wireless multi-hop localization games, and our evaluation of these is discussed in Section 5. Section 6 concludes the paper with a brief summary and mentions future work.

2. Localization and mobile games

2.1 Multi-hop localization

Multi-hop localization techniques have been discussed for wireless multi-hop networks such as sensor and ad-hoc networks. The motivation behind developing multi-hop localization is wanting to know where the node position is in wireless multi-hop networks by using a small fraction of the anchor nodes. An anchor node is one whose position is known in advance through means such as global positioning system (GPS). Much research has been conducted on how to estimate node positions in wireless multi-hop networks (9–15; 28). Most localization techniques can be categorized into two types. The first is localization by using extra ranging devices, such as ultra sound devices, and the second is localization without using extra ranging devices.

In AHLoS (11), an iterative multilateration by using time-of-arrival (TOA) measurements was proposed to estimate large numbers of node positions with a small number of anchor nodes. The basic idea behind iterative multilateration is that at least three anchor nodes carry out the multilateration to estimate unknown nodes. Once the positions for unknown nodes are estimated by anchor nodes, the nodes are configured as pseudo-anchor nodes. Then, pseudo-anchor nodes join to estimate unknown nodes that remain in the network. Another distance-measurement approach have been extensively discussed in the literature. In (16), robust trilateration using the rigidity of graph theory for flipping avoidance has been proposed. In sweeps (17), algorithms to identify global rigidity were employed to estimate the node positions without flipping for sparse node networks. In (18), an error control algorithm was formulated to mitigate against the error propagation for iterative localization. The distance-measurement approach normally achieves precise positioning accuracy. However it requires extra ranging devices, increasing the cost for all nodes.

The localization scheme without using extra ranging devices has been developed for large-scale sensor networks, and it basically exploits connectivity information of multi-hop networks. In GPS-less (9), anchor nodes first flood beacon packets containing their anchor location information, and unknown nodes estimate their positions by using anchor location information with a Centroid formula. In DV-Hop (10), the positions for unknown nodes in a network are estimated by using average hop-count distances from anchor nodes. First, anchor nodes flood their location information to all other anchor nodes, and calculate the average 1-hop distance. Next, anchor nodes carry out a trilateration to unknown nodes by using

hop-count distances. In AFL (15), the positions of unknown nodes are estimated without using anchor nodes. The basic idea behind AFL is to utilize reference nodes that represents the relative axis in a network. The five reference nodes are automatically selected in the manner described in (15), and they determine relative node positions based on the hop-counts from their reference positions. In REP (19), the hop-count distance in a network with holes is calculated by using boundary detection (20). Boundary detection is a technique that can detect the network boundary with only information on network connectivity. In (21), boundary detection and a Delaunay graph were jointly used to prevent node positions from flipping. The localization scheme without using ranging devices enables nodes to estimate node positions while only using the radio capabilities of a sensor node. Hence, it has great flexibility to enable nodes to be applied to localization in the network.

We previously developed optimized link state routing-based localization (ROULA) (28). ROULA does not require the use of extra ranging devices for any nodes and precisely estimate the node distance by using multipoint relay (MPR) nodes. We thus used ROULA in our proposed games to enable the nodes to estimate their positions. ROULA is described in Section 3.

2.2 Mobile games

Let us now present a brief history of mobile games and discuss different aspects of the proposed multi-hop localization games from these. A number of game consoles have been developed for the entertainment computing market. These game consoles have two basic types: home game consoles and mobile game consoles. The Nintendo Entertainment System™ (“Famicom” in Japan) is the iconic home game console and was introduced in 1983 by Nintendo (3). A user plays the games by using a wired hand-held controller. Famicom supports capabilities for multiplayer games. Two users can play games using two controllers connected by cables to the console.

Mobile game consoles have been developed with ad-hoc networking capabilities, such as Nintendo DS™ in 2004 by Nintendo (3) and Play Station Portable (PSP)™ in 2004 by SCE (4). Many video games on mobile game consoles have been released by game software companies. Hot shots golf™ (6) (“minna no golf” in Japan) is one of popular portable video games for PSP. Hot shots golf supports multiple players by using ad-hoc networking capabilities. However, hot shots golf only uses ad-hoc networking capabilities for joining the game.

Mobile games for ubiquitous computing environments have recently attracted a great deal of attention (23; 26; 27). Many varieties of mobile games have been developed thus far. Geocaching (22) is a GPS-based treasure hunting game for outdoor environments. The basic idea behind geocaching is that players hide and seek out containers called “geocaches”. A player hides a geocache and registers the positions provided by the GPS receiver. Once the geocache is registered and released on the geocaching web site, another player finds the geocache based on the positions. Geocaching is being carried out in the actual field, and everyone can get started by using a GPS receiver and mobile console with Internet capabilities.

Human Pacman (24) is a multiplayer field game using a GPS receiver and wireless networking capabilities. Pacman is a video game for Famicom and was originally developed by Namco (5) in 1980. Human Pacman is real field version of Pacman. Players are assigned to either the Pacman team or the Ghost team. Each team has at least two helpers to assist its own team. Virtual cookies are placed on the map and their positions correspond to a real field. The goals are for the Pacman team to collect all virtual cookies and for the Ghost team to catch all the Pacman players. Can You See Me Now? (CYSMN) (25) is a chase game based on location

Game names	Scoring metrics	Player's behavior for game win	Networking capabilities	Real field use	No. of participants, N_p	Required equipment
Hot shots golf™ (6)	Golf game score	Individual	Used to connect to other players	No	$1 \leq N_p \leq 8$	Mobile console
Classical war game	Hitting with model gun	Individual	Not used	Yes	$2 \leq N_p$	Model guns and goggles
Classical tag game	Avoiding on and elapsed time	Individual	Not used	Yes	$2 \leq N_p$	None
Geocaching (22)	Collecting geocaches	Individual	Used to display locations of geocaches	Yes	$1 \leq N_p$	Mobile console and GPS receiver
Human Pac-man (24)	Collecting virtual cookies	Individual or group	Used to record player's trajectory	Yes	$8 \leq N_p$	HMD, mobile console and GPS receiver
CYSMN (25)	Avoiding runners and elapsed time	Individual	Used to display locations of players	Yes	$4 \leq N_p$	Mobile console and GPS receiver
Proposed games	Increasing positioning accuracy and elapsed time	Group	Used for ad-hoc connections and position estimations	Yes	$20 \lesssim N_p$	Mobile console and GPS receiver

Table 1. Comparisons of conventional mobile games with proposed games.

information with GPS. Three runners are visible to players' locations through an virtual online map and they run through actual city streets. Players avoid the runners and compete for the time elapsed since joining the game. If a runner gets within five virtual meters of a player, the player is seen and is excluded from the game.

Table 1 summarizes the features of current mobile games and the proposed multi-hop localization games. Our proposed game has novel distinct aspects from the other works. First, our proposed games use ad-hoc networking capabilities. Conventional mobile games use networking capabilities such as wireless local area network (WLAN) to access the game server that provide the players' location information or a virtual map through the Internet. Our proposed games have novel uses for ad-hoc networking capabilities to conduct multi-hop localization in mobile games. Second, the scoring metric for the game is based on the positioning accuracy of the multi-hop localization technique. In some of the literature, localization is described as being "cooperative" (8). Nodes in the wireless network help to connect with one another to establish their relative positions. The positioning accuracy depends on the number of nodes in the network. Hence, nodes are required to cooperate to achieve higher positioning accuracy. Players must cooperative with their own team in the games.

Finally, let us discuss the number of participants in the proposed games. The proposed games require large numbers of participants compared with other games. This is because multi-hop localization without ranging devices requires a large number of nodes to estimate the positions of the nodes (28). ROULA is guaranteed to estimate all node positions when connectivity is about 20 (28). Connectivity indicates how many nodes are connected to other nodes in 1-hop on average. The least number of participants can be reduced further by using localization with ranging devices although improving the performance of the localization algorithm is beyond the scope of this paper.

3. Optimized Link State Routing-based Localization

3.1 Overview of ROULA

In our two wireless multi-hop localization games, the players used ROULA (28) to enable them to estimate their positions. Let us briefly describe the ROULA technique. A more detailed description of ROULA and its performance are described in (28).

Figure 1 has a conceptual representation of ROULA in a non-convex network topology. The basic idea behind ROULA is that each node matches regular triangles that form exactly convex curves, and makes them into global coordinates by merging overlapping regular triangles iteratively. ROULA is independent of anchor nodes and can determine the correct node positions in a non-convex network topology. In addition, ROULA is compatible with the optimized link state routing (OLSR) network protocol (30) and uses the inherent distance characteristic of MPR nodes.

A non-convex network topology can occur when nodes cannot be deployed in some areas because of obstructions, e.g., buildings or natural features such as trees or mountains. A non-convex network topology appears to be a non-convex curve if the network is seen from a global point of view. However, if the network is viewed locally, each small set of the network appears to be a convex curve. In other words, a non-convex network topology is composed of partially convex curves. To find these convex curves, nodes in ROULA search for nodes that are arranged into regular triangles.

Nodes in ROULA are assumed to use the OLSR protocol in the network layer. Using the OLSR protocol has two advantages. First, the MPR selection used in OLSR has the inherent characteristic of reducing distance errors in localization without using ranging devices. Second,

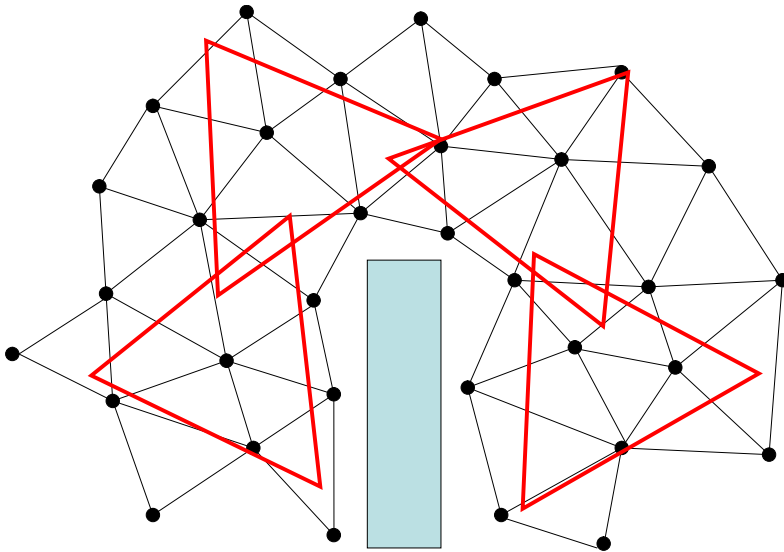


Fig. 1. Conceptual representation of ROULA in non-convex network topology.

nodes in OLSR always hold and update the latest 2-hop node information and MPR nodes in a proactive action that periodically floods hello packets. Node in ROULA localize MPR nodes as their 1-hop nodes without having to make any modifications to the MPR selection. Flooding hello packets and the computational task of MPR selection can be integrated by using the underlying network layer processes.

3.2 Algorithm

The four operations for ROULA are summarized below (28).

1. **Estimating MPR node distances:** Nodes flood hello packets containing their own 1-hop nodes list to their 1-hop nodes. Once a node has a 2-hop nodes list, it selects MPR nodes and estimates the distances between them.
2. **Estimating farthest 2-hop node distances:** Each node selects the farthest 2-hop node for each MPR node and estimates the distances between them.
3. **Estimating relative node positions on regular triangles:** Nodes flood TRI_NOTICE packets to their farthest 2-hop nodes with their farthest 2-hop nodes list. Then, nodes that received TRI_NOTICE packets match regular triangles by using the received farthest 2-hop nodes lists. Nodes then obtain sets of local coordinates by merging their overlapping regular triangles.
4. **Estimating one set of relative coordinates of network:** Sets of relative coordinates are collected and merged into one set of relative coordinates for the network. After that, nodes that have not estimated their positions estimate these by using the Centroid formula (28). If at least three anchor nodes are in the network, the relative coordinates can be converted into absolute coordinates that have the correct network orientation.

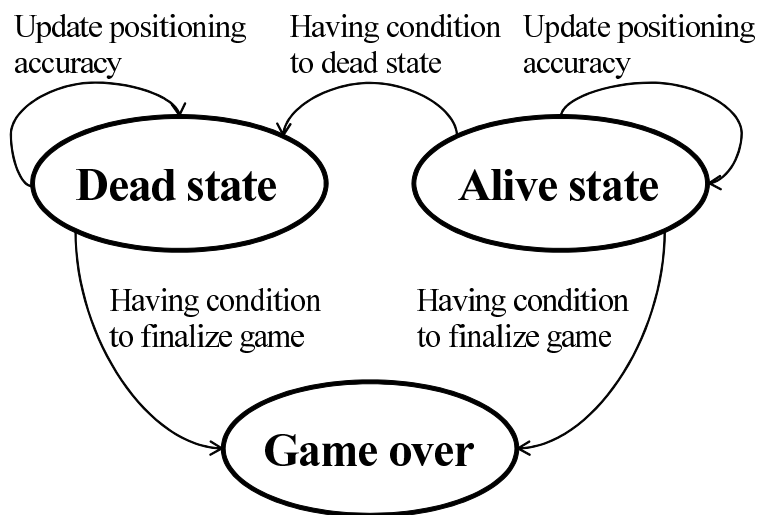


Fig. 2. Principle state transition diagram for multi-hop localization game.

We assume that nodes are deployed in a two-dimensional plane, and a sink node for the network merges all sets of local coordinates in the network. Routing protocol operations are assumed to be done without requiring additional time.

4. Wireless multi-hop localization games

4.1 Overview

The fundamental concept underlying wireless multi-hop localization games is that players on a team establish an ad-hoc network to estimate their positions and then compete for positioning accuracy with the other team by using a multi-hop localization technique. The players use mobile game consoles, called “nodes”, with ad-hoc networking capabilities and play the game on a field.

Figure 2 presents a principle state transition diagram for a multi-hop localization game. Each node is either in a “dead” or an “alive” state. The initial state is alive. Nodes periodically send hello packets and update their positions by multi-hop localization. The condition for transition to a dead state is based on positioning accuracy. The more accurate the positioning obtained by a node, the lower the probability of it transitioning to a dead state. When an alive node satisfies the condition to transition to a dead state, it makes the transition. The condition for finalizing the game is different for each game’s goal.

Two objectives in using ad-hoc networking capabilities are to connect nodes within a limited communications range and to estimate node positions. Once a node is connected to other nodes, it can specify the number of 1-hop nodes. ROULA can estimate node positions by using the number of 1-hop nodes.

Here, we have simplified the mobility characteristics of human motions to win a game as random motions.

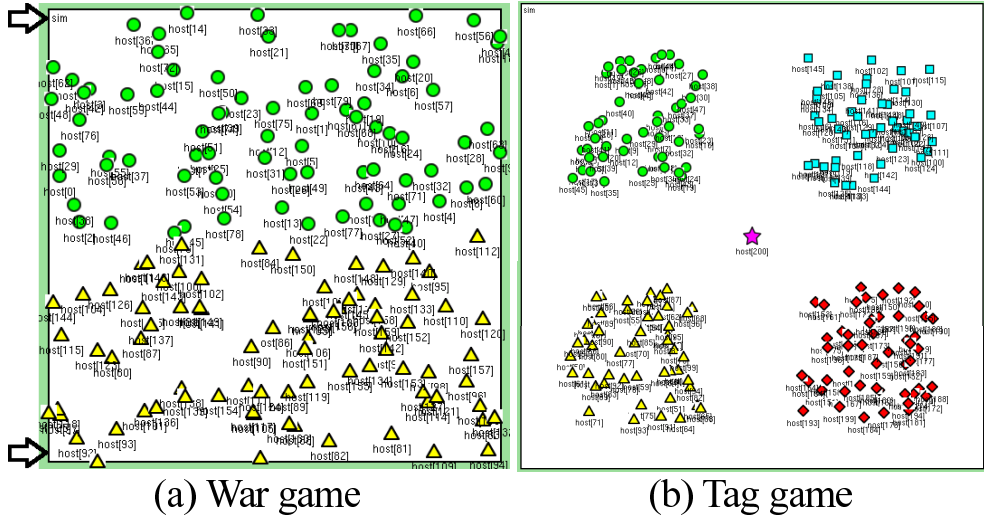


Fig. 3. Initial node placements of (a) war and (b) tag games. Teams 1, 2, 3, and 4 correspond to circles, triangles, squares, and diamonds. Oni is represented by star. Arrows indicate locations of enemy lines.

Algorithm 1 When node i senses hello packet of node j

- 1: **if** node i is on same team as node j **then**
 - 2: receive packet.
 - 3: **else**
 - 4: drop packet.
 - 5: **if** $e_i == e_j$ && $\mathcal{U}(0, 100) \leq 50$ **then**
 - 6: transition to dead state.
 - 7: **else if** $e_i > e_j$ **then**
 - 8: transition to dead state.
-

4.2 War game

Let N_{tm} denote the number of teams and N_g denote the number of nodes required to finalize the war game. Each team has an equal number, N_n , of players. The goal of the war game is for N_g alive nodes on a team to reach the enemy line. $N_{tm}|N_{tm} > 1, N_g|N_g > 0$, and $N_n|N_n \geq N_g$ can be varied. N_{tm} was consistently set to 2 and N_g was set to 1 for the war game discussed here.

Let us consider the case of $N_n = 80$. Figure 3(a) shows the node placement at the beginning of the war game. The field is divided in half, and each team initially occupies one of the two areas. Each team has the same number of nodes. The nodes for team 1 are represented by circles, and those for team 2 are represented by triangles. The arrows in Fig. 3(a) indicate the locations of the enemy lines, which were set 10 [m] from the back end of each team's area. The mobility of each node was modeled as a random waypoint constrained to proceed toward the enemy line. The velocity of each node was determined by using a random variable with an

Algorithm 2 When node i senses hello packet of oni

-
- 1: **if** $\mathcal{U}(0,100) \leq e_i * \kappa$ **then**
 - 2: Node i transitions to dead state.
-

exponential distribution, $\mathcal{E}(v)$, with a mean of v . If a node reaches the enemy line, its team is declared the winner. The nodes on each team periodically run multi-hop localization to estimate their positions. The nodes cannot communicate with the nodes on the other team. We used the positioning error obtained by multi-hop localization as the metric to determine whether a node transitions to a dead state. In the proposed game, all nodes are assumed to be anchor nodes to enable relative coordinates to be converted into absolute coordinates and obtain the positioning error. The positioning error, e_i , is normalized by the communication range, R :

$$e_i = \frac{\sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2}}{R}, \quad (1)$$

$$i = 1 \dots N,$$

where N is the total number of nodes, (x_i, y_i) represents the true position of node i , and (\hat{x}_i, \hat{y}_i) is the estimated position. The true position can be obtained by using GPS. When a node cannot estimate its own position, it is assigned a positioning error of 100%. A node transitions to a dead state depending on its positioning error. State transitions algorithm for node i is described in Algorithm 1. In the proposed games, nodes periodically flood hello packets. When a node senses a hello packet, the node receives the packet only if it has arrived from a node on the same team. Otherwise, the node drops the packet and decides whether to transition to the dead state on the basis of positioning errors (lines 5–8 of Algorithm 1). The $\mathcal{U}(a, b)$ represents a random variable with a uniform distribution in the interval $[a, b]$. If the node has the same positioning error, it transitions to a dead state with a probability of 50% (line 5 of Algorithm 1). Otherwise, the node transitions to a dead state if it has a greater positioning error. Once the node transitions to the dead state, it stops moving and receiving packets for localization. If all the nodes on a team transition to dead states, the other team is declared the winner.

4.3 Tag game

Let N_{oni} denote the number of demons (“oni” in Japanese). Each team has the same number of players. The goal of the tag game is for players on each team to survive N_{oni} oni attacks. Players belong to one of the N_{tm} teams. The N_{oni} oni move around the field and never transition to dead states. $N_{tm} | N_{tm} > 0, N_{oni} > 0$, and $N_n > 0$ can be varied. Here, we have consistently considered the case where $N_{tm} = 4$, $N_{oni} = 1$, and $N_n = 50$ for the tag game.

Figure 3(b) shows the node placement at the beginning of the tag game. Four teams are located in four equal areas. Teams 1, 2, 3, and 4 correspond to the circles, triangles, squares, and diamonds. The oni is located at the center of the field, and is represented by the star. The mobility of the oni and nodes was modeled as a random waypoint. The velocity of each node was chosen by using a random variable with an exponential distribution, $\mathcal{E}(v)$. The nodes periodically run multi-hop localization to estimate their positions.

Algorithm 2 gives the algorithm for node i to transition to a dead state. When node i receives a hello packet from the oni, it transitions to a dead state according to a probability based on its

Field	500 × 500 [m]
Communication range	100 [m]
Node mobility	Random waypoint constrained to advance toward enemy line
Node velocity	$\mathcal{E}(v), v = 1, 2, 4$ [m/step]

Table 2. Simulation parameters for war game

own positioning error. The basic operation of this metric is that if the positioning error is 60%, the probability to transition to the dead state is 60%. In line 1 of Algorithm 2, we introduced design parameter κ to mitigate the impact of positioning error. Our evaluation of the impact of κ is discussed in Section 5.3. We defined max_step , which denotes the maximum length of time for the tag game. The conditions to finalize the tag game are cases where only one team survives or max_step time is up. When max_step is out of time, the winner is the team that has the maximum survival time. The survival time is defined as

$$T_{survive} = \sum_{t=0}^{max_step} \text{number of alive nodes.} \quad (2)$$

In our evaluation, we set max_step to 1000.

5. Evaluation

5.1 Simulation setting

The simulation environment we used was a discrete event simulation environment, OM-NeT++ (31) with Mobility Framework (32). Existing network simulators do not have localization functionality. We implemented localization functionality into OMNeT++. Our localization simulation platform enables to test the localization performance with discrete event simulation. The simulation trials were conducted 30 times with random seeds, and the results were averaged.

5.2 War game

5.2.1 Impact of number of nodes

Table 2 lists the simulation parameters for the war game. The communication range was fixed and we have ignored packet loss in this paper. Hello packets were periodically sent by 5 time step.

Figure 4 shows snapshots of the war game at time steps of 500 and 1000 for 160 nodes. We set the mean node velocity (v) of all nodes to 1 [m/step]. A cross on a node represents a dead state. As we can see from Fig. 4(a), the nodes proceeded toward the enemy lines. As time went by, the nodes got closer to the enemy lines, and more and more nodes died, as seen in Fig. 4(b).

Figure 5 plots the positioning error, ratio of estimated nodes, and number of alive nodes against the time step when each team (T) had 120 nodes. The variance in the positioning error increased as the number of alive nodes decreased, which is consistent with the finding that the number of nodes contributes to positioning accuracy in multi-hop localization (28). As time went by, the number of nodes that could participate in localization decreased. Hence, the variance in positioning accuracy increased. We defined the ratio of estimated nodes as the

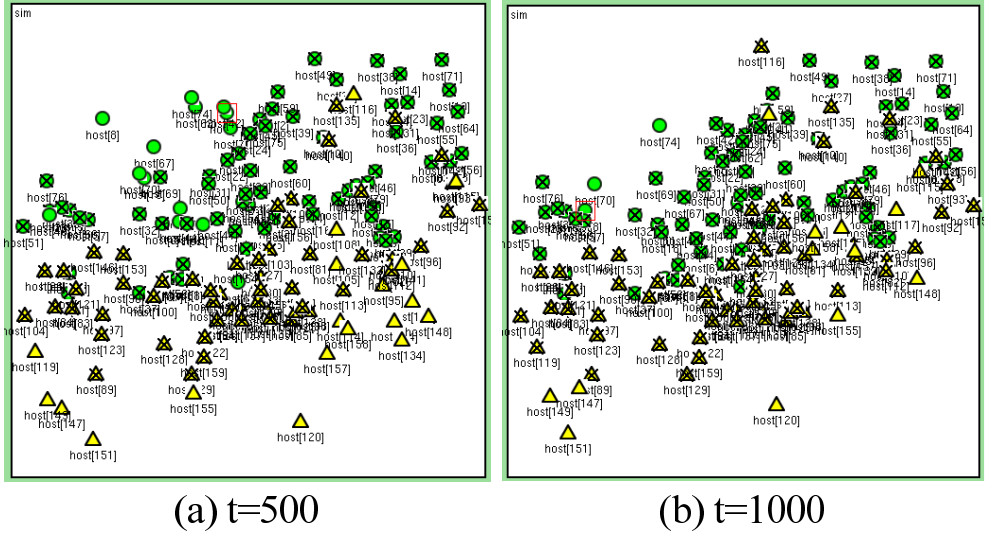


Fig. 4. Snapshots of war games for time steps (a) 500 and (b) 1000 ($N=160$, $v_{T1,T2}=\{1,1\}$). Nodes for team 1 are represented by circles, and those for team 2 are represented by triangles. Cross on node represents dead state.

percentage of nodes that could estimate their positions out of all alive nodes. The number of alive nodes decreased over time. A small number of nodes makes it difficult to estimate node positions using multi-hop localization. Therefore, the ratio of estimated nodes decreased as the number of nodes decreased. The number of alive nodes on both teams remained approximately the same over time. This is because all the nodes had the same velocity and used the same strategy to proceed to the enemy lines.

Figure 6 plots the results for 160 nodes. Compared with the results for 120 nodes, the ratio of estimated nodes was better, confirming that the number of nodes contributed to the ratio of estimated nodes.

Table 3 lists the number of wins by team for 30 trials with the different parameter settings. The row for scenario A in Table 3 shows that the number of wins for teams 1 and 2 for 120 and 160 nodes, corresponded to 16 and 14, and 15 and 15. The number of nodes did not significantly affect the number of wins.

Although the results presented here were for basic scenarios, we observed that the multi-hop localization game using ROULA worked well as a game with ad-hoc networking capabilities.

5.2.2 Impact of velocity

We evaluated the impact of velocity by varying the velocities of nodes on each team. Figure 7 shows snapshots of war games at time steps of 250 and 500 for 160 nodes and the mean node velocities of 1 [m/step] for team 1 and 4 [m/step] for team 2. As seen in Fig. 7(a), the nodes on team 2 were closer to the enemy line than those on team 1. Figure 7(b) shows that many nodes on team 2 died on their enemy's side, and that many nodes on team 1 died on their own side.

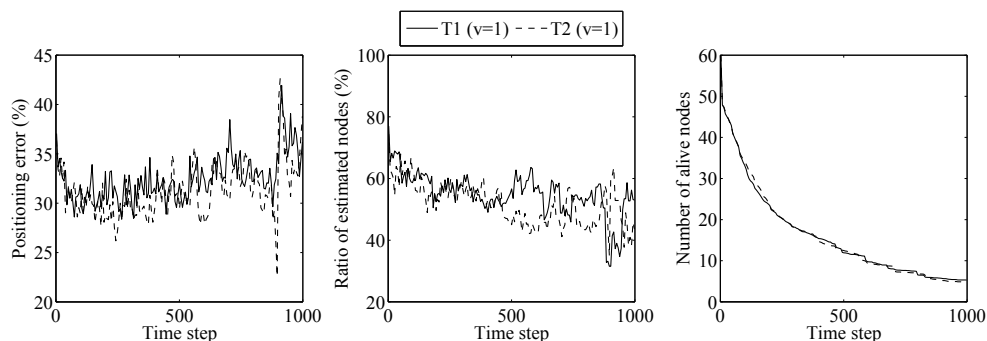


Fig. 5. Positioning error, ratio of estimated nodes, and number of alive nodes ($N=120$, $v_{T1,T2}=\{1,1\}$).

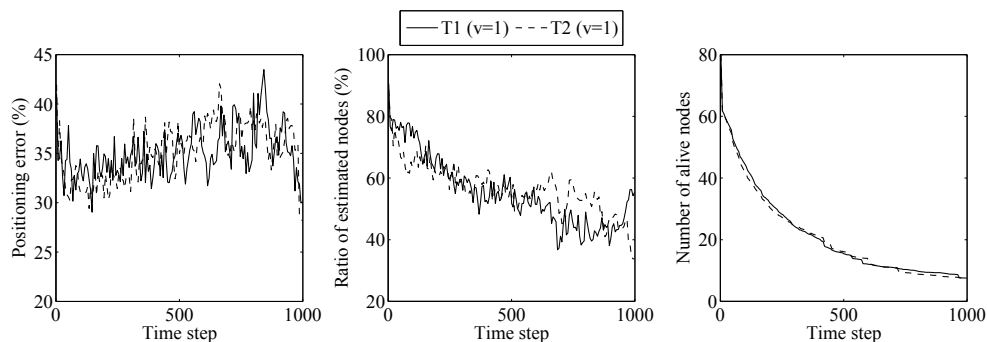


Fig. 6. Positioning error, ratio of estimated nodes, and number of alive nodes ($N=160$, $v_{T1,T2}=\{1,1\}$).

Figure 8 plots the positioning error, ratio of estimated nodes, and number of alive nodes when the velocities of nodes on team 1 were 1 and those on team 2 were 2. The ratio of estimated nodes of team 2 was slightly lower than that on team 1. This is because the nodes on team 2 were more spread out because they moved more quickly, making it more difficult to estimate their node positions using multi-hop localization.

However, the goal of the war game was to reach the enemy line. The row for scenario B in Table 3 indicates the number of wins for teams 1 and 2 for velocities of 1 and 2 [m/step]. Although the ratio of estimated nodes on team 2 was slightly lower than that for team 1, team 2 had more wins. This is because the condition for finalizing the war game was reaching the enemy line; the team with the higher average node velocity had the greater number of wins.

Figure 9 plots the positioning error, ratio of estimated nodes, and the number of alive nodes when the velocities of nodes on team 1 were 1 [m/step] and those on team 2 were 4 [m/step]. The ratio of estimated nodes on team 2 was lower than that on team 1. However, team 2 had

		Team 1	Team 2
Scenario A: N	120	16	14
	160	15	15
Scenario B: $v_{T1,T2}$	{1,2}	12	18
	{1,4}	9	21
Scenario C: y coord. of obst.	250	15	15
	100	29	1

Table 3. Number of team wins in war game. In scenario A, number of nodes N was varied and for $v_{T1,T2} = \{1, 1\}$. In scenario B, the velocity $v_{T1,T2}$ was varied for $N=160$. In scenario C, y coordinate of obstruction was varied for $N=160$ and $v_{T1,T2} = \{1, 1\}$.

more wins as can be seen from the scenario B results in Table 3. This result suggests that the win rate for the war game depends on the node velocity of nodes.

5.2.3 Impact of obstruction position

We evaluated the impact of obstruction position by adding an obstruction to the field and varying its position. Figure 10 shows snapshots of war games assuming that there is an obstruction at position (250, 100). The height and width of the obstruction were 200 and 200 [m], respectively. No node could enter the portion with the obstruction. Figure 10 shows that nodes had trouble moving forward even though the velocities of the nodes on both teams were the same.

To evaluate the impact of the obstruction's position, we fixed its x -axis position at 250 and varied its y -axis position. As seen in Fig. 11, the positioning accuracy and ratio of estimated nodes were almost the same when the obstruction's position was (250, 250). As we can see from in Fig. 12, when the obstruction's position was (250, 100), the ratio of estimated nodes on team 2 was lower than that on team 1. This is because the obstruction made the network topology non-convex, making it difficult to estimate the node positions using multi-hop localization (28). Although the positioning accuracy for team 2 was better than that for team 1, the ratio of estimated nodes was lower. Hence, many nodes were assigned a 100% positioning error. Consequently, the number of alive nodes on team 2 was less than that on team 1.

Not surprisingly, the row for scenario C in Table 3 reveals that the number of team wins was closely related to the obstruction's position. This is because the scoring metric is based on positioning accuracy. The result proved that the win rate for the war game depends on obstruction positions.

Since multi-hop localization increases the positioning error in a non-convex network, the characteristics of an obstruction's position can be considered in team strategies. For example, players can collaborate to move to avoid making a non-convex network in a team's topology. The characteristics of multi-hop localization open the door to creating various game strategies.

5.3 Tag game

5.3.1 Impact of local rule on connectivity constraint

We evaluated the impact of two local rules, i.e., connectivity constraint and death penalty, in the tag game. The connectivity constraint is a rule where the nodes have to avoid situations where the current connectivity becomes less than or equal to a specified connectivity. The connectivity is defined by the number of nodes connected by 1-hop. Thus, nodes have to keep

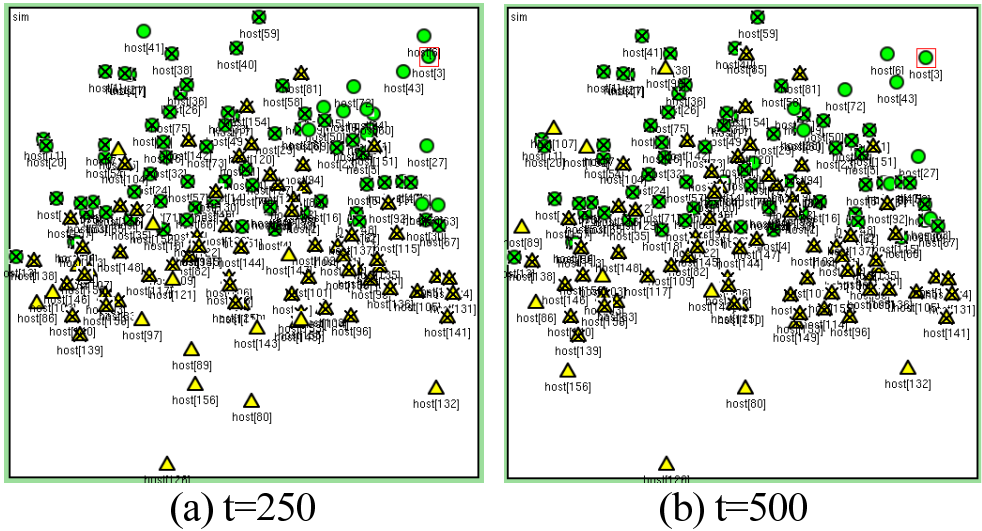


Fig. 7. Snapshots of war games for time steps of (a) 250 and (b) 500 ($N=160$, $v_{T1,T2}=\{1,4\}$). Nodes for team 1 are represented by circles, and those for team 2 are represented by triangles. Cross on node represents dead state.

Field	700×700 [m]
Communication range	100 [m]
Number of nodes	200
Node mobility	Random waypoint
Node velocity	$\mathcal{E}(v)$, $v = 1$ [m/step]
Connectivity constraint	$C_{T1,T2,T3,T4}=\{0 \text{ (not applied)}, 2, 5, 8\}$

Table 4. Simulation parameters for tag game.

moving to prevent the current connectivity from being violated. The rule of the connectivity constraint can be easy to accomplish in an actual game, because each node can know the current connectivity due to the use of mobile game consoles with ad-hoc networking capabilities. First, we evaluated the impact of local rule of the connectivity constraint. Table 4 presents the simulation parameters for the tag game. Figure 13 shows snapshots of a tag game with the local rule of the connectivity constraint (C). Teams 1, 2, 3, and 4 are located at the top left, bottom left, top right, and bottom right, respectively. The connectivity constraints correspond to 0 (not applied), 2, 5, and 8. As shown in Fig. 13(a), the nodes on team 1 spread over the field while those on team 4 are bunched together. Those on teams 3 and 4 spread gradually, as shown in Fig. 13(b).

Figure 14 plots the results for positioning error, ratio of estimated nodes, and number of alive nodes. The higher the connectivity constraint, the lower the positioning error. This is because the positioning accuracy using multi-hop localization depends on the connectivity (28). The greater the connectivity constraints, the higher the ratio of estimated nodes.

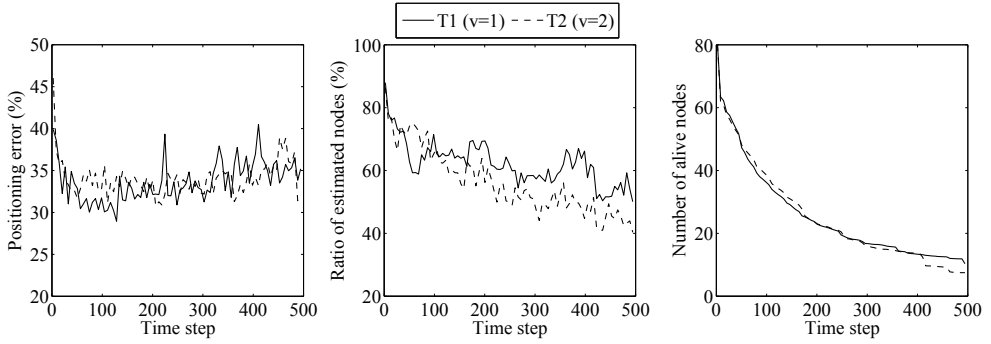


Fig. 8. Positioning error, ratio of estimated nodes, and number of alive nodes ($N=160$, $v_{T1,T2}=\{1,2\}$).

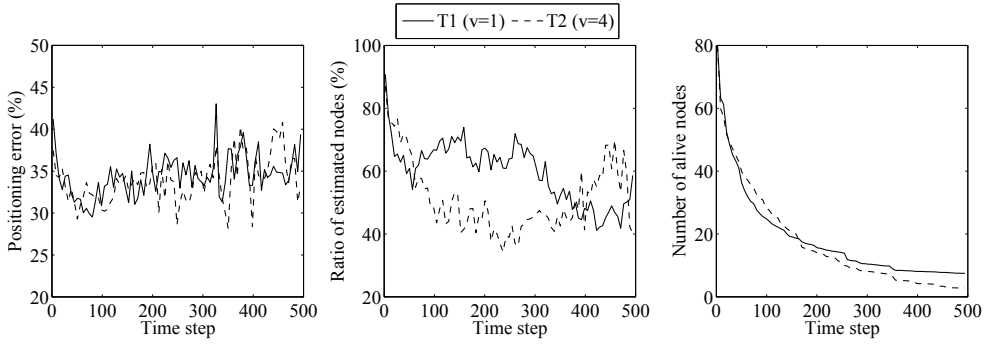


Fig. 9. Positioning error, ratio of estimated nodes, and number of alive nodes ($N=160$, $v_{T1,T2}=\{1,4\}$).

Table 5 summarizes the survival time for the tag game. As seen from the row for scenario D in Table 5, team 1 with a connectivity constraint of 0, had the longest survival time even though it had the lowest localization rate. This is because nodes with a lower connectivity constraint could more readily move around the field. Since the probability of their encountering an oni was lower, nodes with a lower connectivity constraint could survive longer. This result is not suitable for playing the game, because there is no advantage to cooperate for localization, and it does not support the fairness of the game. We thus examined the introduction of a design parameter and a local rule to control the win rate for the game.

5.3.2 Impact of local rule on death penalty

We next evaluated the local rule of a death penalty to impose a penalty for moving alone. The local rule of the death penalty was to impose transition to a dead state when the node could not estimate its own position ω times, consecutively. In addition, we introduced a design parameter κ to mitigate transition to a dead state by using multi-hop localization. The parameter κ encourages the longer survival time in the tag game. κ was introduced in Algorithm 2.

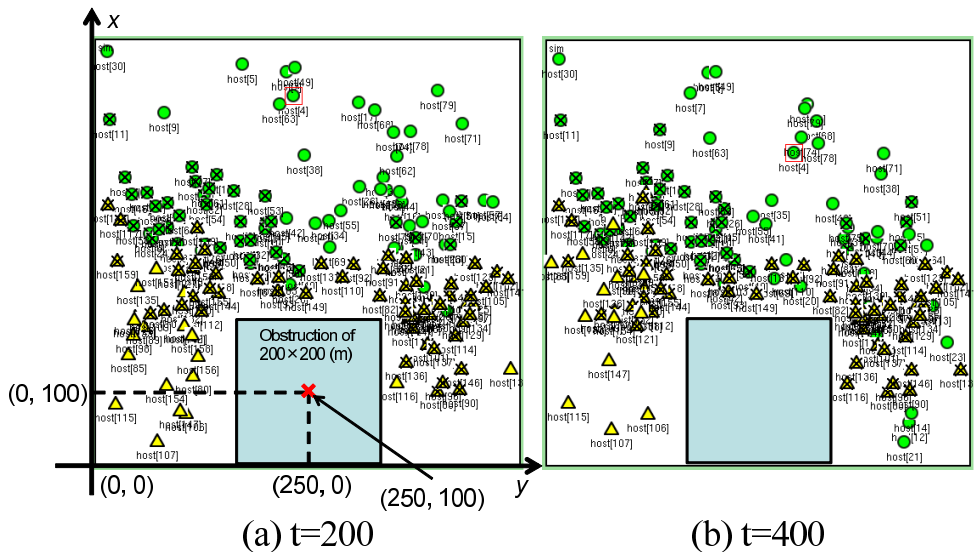


Fig. 10. Snapshots of war game for time steps (a) 200 and (b) 400 with obstruction at (250, 100) ($N=160, v_{T1,T2}=\{1,1\}$). Nodes for team 1 are represented by circles, and those for team 2 are represented by triangles. Cross on node represents dead state. Obstructions are drawn as large gray squares.

	Team 1	Team 2	Team 3	Team 4
Scenario D	6429.4	5460.9	5855.8	6001.9
Scenario E	2237.2	2739.4	2937.4	3454.3

Table 5. Survival time in tag game for $C_{T1,T2,T3,T4}=\{0,2,5,8\}$. Scenario D enabled the rule of connectivity constraint. Scenario E enabled the rule of death penalty.

Figure 15 shows snapshots of the tag game with the local rule of the death penalty enabled. Nodes on team 1 with a connectivity constraint of 0 are still widely spread out, however, their death rate is higher due to the local rule of the death penalty. Figure 16 plots the results for positioning error, ratio of estimated nodes, and number of alive nodes with the local rule of the death penalty. The κ was set to 0.5, and ω was set to 2. The number of alive nodes on team 1 decreased over time, because nodes with a lower connectivity constraint had wider dispersion. As we can see from the row for scenario E in Table 5, the higher the connectivity constraint, the longer the survival time. This result demonstrates that the local rule of the death penalty and the design parameters κ, ω effectively maintained the fairness in the tag game.

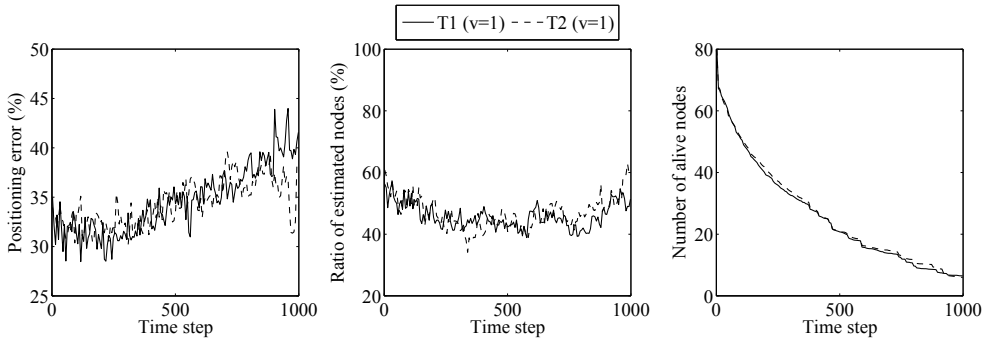


Fig. 11. Positioning error, ratio of estimated nodes, and number of alive nodes with obstruction at (250, 250) ($N=160$, $v_{T1,T2}=\{1,1\}$).

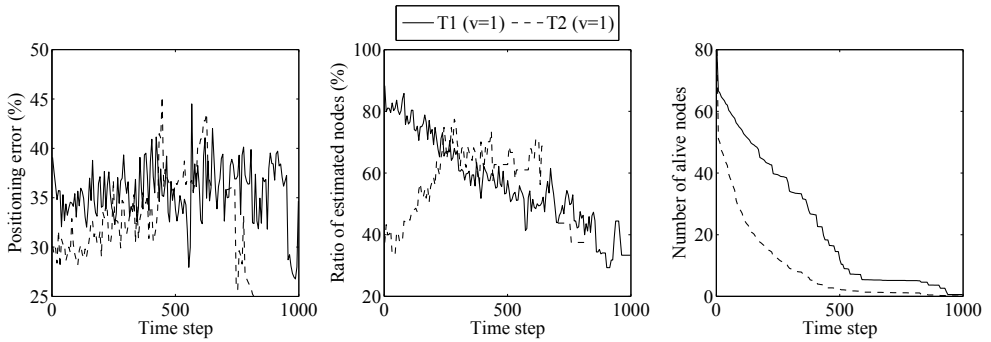


Fig. 12. Positioning error, ratio of estimated nodes, and number of alive nodes with obstruction at (250, 100) ($N=160$, $v_{T1,T2}=\{1,1\}$).

6. Conclusion

We developed two wireless multi-hop localization games, i.e., a war game and a tag game, based on classical field games. The proposed games are played using mobile game consoles with ad-hoc networking capabilities. The fundamental concept underlying a wireless multi-hop localization game is that players on a team establish an ad-hoc network to estimate their positions and then compete for positioning accuracy with other teams obtained using a multi-hop localization technique. Using simulation, we found that node velocity and obstruction positions were parameters to control the win rate for the war game. In the tag game, the higher connectivity constraint led to be surviving longer. The simulations demonstrated that the win rate for the proposed games depends on obstruction positions and connectivity constraint. We also demonstrated that introducing a design parameter and enforcing local rules were needed to control the win rate for the game. The results demonstrated that the proposed games worked well as games with ad-hoc networking capabilities.

In this work, we simply assumed that nodes had random motion to investigate the primitive operations of proposed wireless multi-hop localization games. In the real world, players

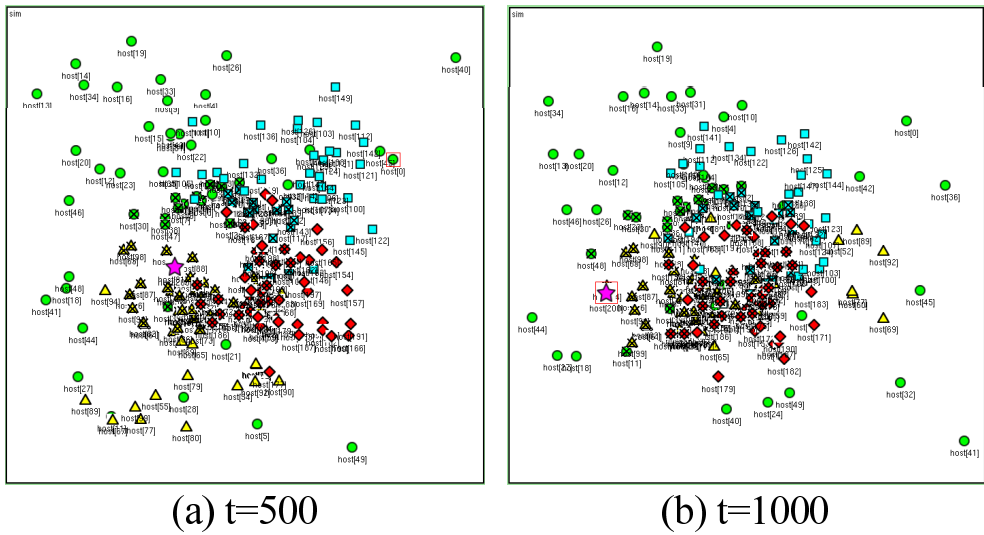


Fig. 13. Snapshots of tag game for time steps (a) 500 and (b) 1000 ($N=200$, $C_{T1,T2,T3,T4}=\{0,2,5,8\}$). Teams 1, 2, 3, and 4 correspond to circles, triangles, squares, and diamonds. Oni is represented by star. Cross on node represents dead state.

would cooperate to minimize their positioning errors, or oni would employ a strategy to track alive nodes. These motions can be embedded into node mobility in simulations to obtain more realistic game results. Since the presented study only covered a range of application proposals on combining ad-hoc networking and multi-hop localization, we suggest that there is a need for further research in terms of appropriateness and effectiveness for the games. Our future work includes detailed evaluations of games with various location-based strate-

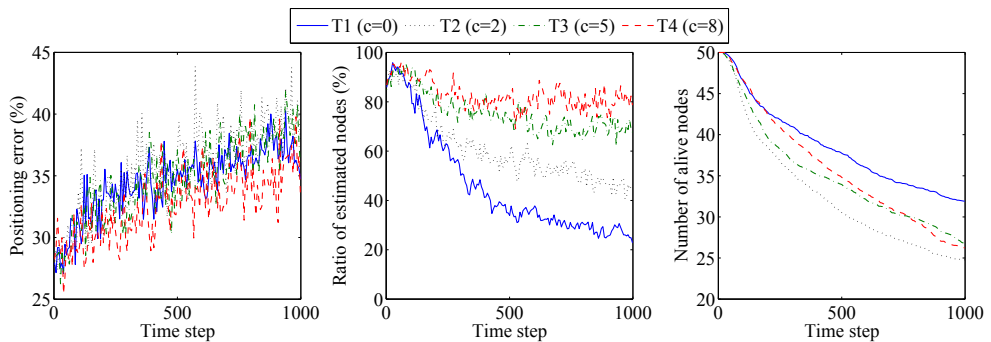


Fig. 14. Positioning error, ratio of estimated nodes, and number of alive nodes ($N=200$)

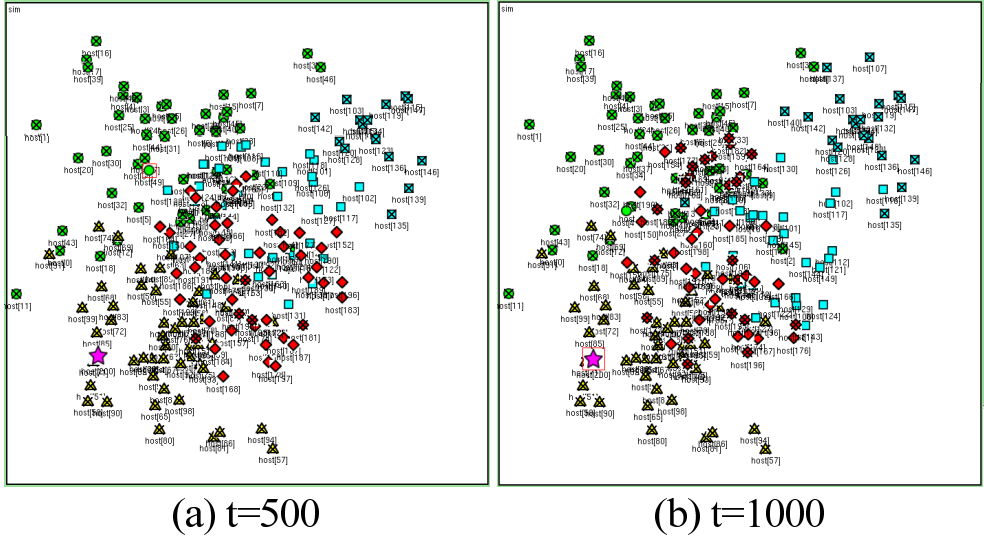


Fig. 15. Snapshots of tag game for time steps (a) 500 and (b) 1000 with local rule of death penalty enabled ($N=200$, $C_{T1,T2,T3,T4}=\{0,2,5,8\}$). Teams 1, 2, 3, and 4 correspond to circles, triangles, squares, and diamonds. The oni is represented by a star. A cross on a node represents a dead state.

gies with positioning errors, actual game testing in the real field, and verifying the degree of user satisfactions when they actually play the proposed games.

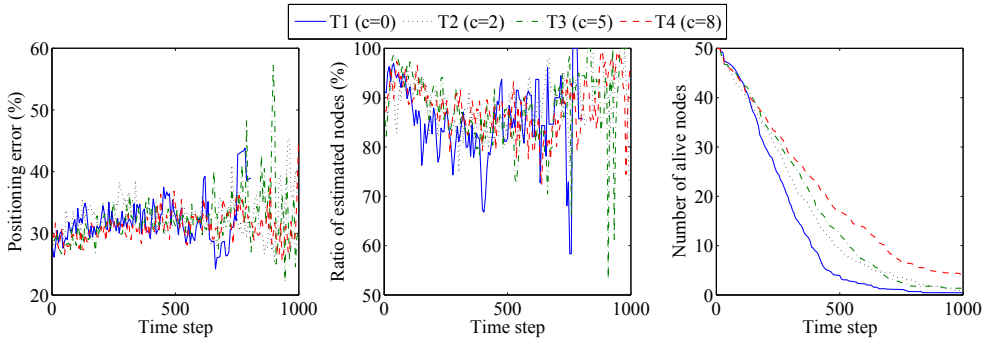


Fig. 16. Positioning error, ratio of estimated nodes, and number of alive nodes with local rule of death penalty ($N=200$, $\kappa = 0.5$ and $\omega = 2$).

7. References

- [1] Zigbee Alliance, <http://www.zigbee.org/>.
- [2] MANET Working Group, <http://www.ietf.org/html.charters/manet-charter.html>.
- [3] Nintendo Co.,Ltd, <http://www.nintendo.com/>.
- [4] Sony Computer Entertainment Inc., http://www.scei.co.jp/index_e.html.
- [5] Namco Limited, <http://www.namco.co.jp/>.
- [6] Hot Shots Golf: Open Tee, http://www.us.playstation.com/PSP/Games/Hot_Shots_Golf_Open_Tee.
- [7] J. Navas and T. Imielinski, "GeoCast - Geographic Addressing and Routing," *Proc. ACM/IEEE Mobicom*, pp. 66–76, 1997.
- [8] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero III, R. L. Moses, and N. S. Correal, "Locating the Nodes—Cooperative Localization in Wireless Sensor Networks," *IEEE Signal Processing Magazine*, vol. 22, pp. 54–69, 2005.
- [9] N. Bulusu, J. Heidemann, and D. Estrin, "GPS-less Low Cost Outdoor Localization For Very Small Devices," *IEEE Personal Communications Magazine*, vol. 7, no. 5, pp. 28–34, 2000.
- [10] D. Niculescu and B. Nath, Ad Hoc Positioning System (APS), *Proc. IEEE Globecom*, vol. 5, pp. 2926–2931, 2001.
- [11] A. Savvides, C. Han, and M. B. Strivastava, "Dynamic Fine-grained Localization in Ad-hoc Networks of Sensors," *Proc. ACM/IEEE Mobicom*, pp. 166–179, 2001.
- [12] C. Savarese, J. Rabaey, and K. Langendoen, "Robust Positioning Algorithms for Distributed Ad-hoc Wireless Sensor Networks," *Proc. USENIX Technical Annual Conference*, pp. 317–327, 2002.
- [13] T. He, C. Huang, B. M. Blum, J. A. Stankovic, and T. Abdelzaher, "Range-free Localization Schemes for Large Scale Sensor Networks," *Proc. ACM/IEEE Mobicom*, pp. 81–95, 2003.
- [14] D. Niculescu and B. Nath, Ad Hoc Positioning System (APS) Using AoA, *Proc. IEEE Infocom*, vol. 3, pp. 1734–1743, 2003.
- [15] N. B. Priyantha, H. Balakrishnan, E. Demaine, and S. Teller, "Anchor-free Distributed Localization in Sensor Networks," *Technical Report TR-892*, MIT LCS, 2003.
- [16] D. Moore, J. Leonard, D. Rus, and S. Teller, "Robust Distributed Network Localization with Noisy Range Measurements," *Proc. ACM Sensys*, pp. 50–61, 2004.
- [17] D. K. Goldenberg, P. Bihler, M. Cao, J. Fang, B. D. O. Anderson, A. S. Morse, and Y. R. Yang, "Localization in Sparse Networks using Sweeps," *Proc. ACM Mobicom*, pp. 110–121, 2006.
- [18] J. Liu, Y. Zhang, and F. Zhao, "Robust Distributed Node Localization with Error Management," *Proc. ACM Mobihoc*, 2006.
- [19] M. Li and Y. Liu, "Rendered Path: Range-free Localization in Anisotropic Sensor Networks with Holes," *Proc. ACM Mobicom*, pp. 51–62, 2007.
- [20] Y. Wang, J. Gao, and J. S.B. Mitchell, "Boundary Recognition in Sensor Networks by Topological Methods," *Proc. ACM Mobicom*, pp. 122–133, 2006.
- [21] S. Lederer, Y. Wang, and J. Gao, "Connectivity-based Localization of Large Scale Sensor Networks with Complex Shape," *Proc. IEEE Infocom*, pp. 13–18, 2008.
- [22] Geocaching, <http://www.geocaching.com/>.
- [23] S. Bjork, J. Holopainen, P. Ljungstrand, and K. Akesson, "Designing Ubiquitous Computing Games – A Report from a Workshop Exploring Ubiquitous Computing Entertainment," *Personal and Ubiquitous Computing*, vol. 6, pp. 443–458, 2002.
- [24] A. D. Cheok, K. H. Goh, W. Liu, F. Farbiz, S. W. Fong, S. L. Teo, Y. Li, and X. Yang, "Human Pacman: A Mobile, Wide-area Entertainment System Based on Physical, Social, and Ubiquitous Computing," *Personal and Ubiquitous Computing*, vol. 8, pp. 71–81, 2004.

- [25] S. Benford, R. Anastasi, M. Flinham, A. Drozd, A. Crabtree, C. Greenhalgh, N. Tandanij, M. Adams, and J. Row-Farr, "Coping with Uncertainty in a Location-based Game," *IEEE Pervasive Computing*, vol. 2, no. 3, pp. 34–41, 2003.
- [26] C. Magerkurth, A. D. Cheok, R. L. Mandryk, and T. Nilsen, "Pervasive Games: Bringing Computer Entertainment Back to the Real World," *ACM Computers in Entertainment*, vol. 3, 2005.
- [27] C. Schlieder, P. Kiefer, S. Matyas, "Geogames: Designing Location-Based Games from Classic Board Games," *IEEE Intelligent Systems*, vol. 21, no. 5, pp. 40–46, 2006.
- [28] T. Takenaka, H. Mineno, Y. Tokunaga, N. Miyauchi, and T. Mizuno, "Performance Analysis of Optimized Link State Routing-based Localization," *Information Processing Society of Japan (IPSJ) Journal*, vol. 48, no. 9, pp. 3286–3299, 2007.
- [29] T. Takenaka, H. Mineno, and T. Mizuno, "Evaluation of Wireless Multi-hop Localization Game for Entertainment Computing," *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2008.
- [30] T. Clausen and P. Jacquet, "Optimized Link State Routing Protocol (OLSR)," *IETF RFC* 3626, 2003.
- [31] OMNeT++ Discrete Event Simulation System, <http://www.omnetpp.org/>.
- [32] Mobility Framework for OMNeT++, <http://mobility-fw.sourceforge.net/>.

Measuring Network Security

Emmanouil Serrelis and Nikolaos Alexandris

University of Piraeus

Greece

1. Introduction

The motive for this research has been the famous quote of Lord Kelvin “You can not improve what you can not measure”. Today’s information era has given new interpretations to this, expressing the need to measure abstract concepts such as Information Security. There are multiple sources, ranging from academic research to industrial reports, such as (Danahy, 2004), (Fisher, 2009) and (Sonnenreich et al., 2006) that share the same view and highlight the importance of measuring security within the context of Information Technologies.

This chapter provides the necessary information as well as the proper tools to measure the security of both IT Systems and business services are based on IT Systems. The main target of the methodologies that described is to provide a better way for managing the security of IT Systems and Infrastructures.

The first section of this chapter covers the basic requirements of any security measurement methodology. The second section of this chapter introduces a taxonomy of the existing security measurement methodologies. The third section highlights the limitations of the existing security measurement methodologies and supports that security should be calculated instead. The fourth and final section of the chapter elevates the need of new measurement methodologies for network security. New methodologies should be taking into account business needs apart from the traditional information technology requirements.

2. The need for security measurement

According to the American dictionary of Princeton the general definition of measurement is “the act or process of assigning numbers to phenomena according to a rule”. Measurement has a very close relationship with metric which, according the same source, is a “a system of related measures that facilitates the quantification of some particular characteristic”.

Focusing on the research area of IT (Maizlitsch, 2005) distinguishes between metrics that are used to quantify values that act as a control of proper functionality and those that act as a performance indicator. Security measurement is a topic that falls under the first category of controlling the proper functionality of security processes.

Although IT Security measurement is very interesting and useful topic in both academic and industrial environments, it deals with the quantification of an abstract concept. Similarly to any other abstract concept, security measurements tend to have rather vague implementations. This is justified by the fact that is hard to provide a figure that could express the current level of security. Thus, the difficulty of measuring security leads to the research of proper methodologies that could define the appropriate metrics as well as describe the necessary measurement process.

The expected benefits from the measurement of security are:

- Enable business strategy: IT Security is essential for the development and the support of trust between organizations, partners, customers and employees. This implies that there measuring the current level of security can held the alignment of business and technology strategies with security aspects and requirements.
- Support the daily business activities: Security measurement can accommodate the increases of the value of information and thus the increased risk levels related to each organizational asset.
- Facilitate risk management: The provision of a better toolset for managing risks can improve decision making as well as taking advantage of business opportunities.
- Reduce costs: A limited understanding of security status could lead to a high-cost operation of IT Systems and business processes, as well as to increased marketing and promotion expenses in order to “protect” the reputation of the products and the services of an organisation.
- Comply to regulatory and legal requirements: Being able to present and report the current status of security and risks is the basic condition for compliance.

A standardized approach of security measurement should aim to enable the operation of an organization without uncertainties or doubts, within a framework that could quantify the probability of a threat occurring, estimate the cost a potential damage, depict the performance overhead of the security processes and evaluate the effectiveness of security measures.

3. Requirements of security measurement methodologies

Having described the expected benefits of security measurement, this paragraph presents the necessary attributes of the security measurement methodologies, by interpreting business requirements in terms of IT Security. The adoption of those requirements is essential in order to make the measurement results utilizable.

A basic requirement is to enable envision to management in the section of security. Each measurement methodology should primarily intend to provide information that would depict the current status as well as the future trends from the security point of view. Additionally, the analysis of the security measurement should:

- Aid an analyst to diagnose the issues that are related to security and evaluate the performance of the existing mechanisms and processes.
- Quantify specific security characteristics and parameters.
- Easy the investigation of hypothetical and “before and after” scenarios.
- Focus the measurement interest to the causes, the media and the meaning of the results instead of the methodologies that were used.

According to (Jaquith, 2007) each measurement methodology should have as many of the following characteristics as possible:

- Consistently measured
- Cheap to gather
- Expressed as a number
- Uses at least one unit of measure
- Contextually specific
- Partial weight
- Repetitiveness
- Comparability

3.1 Consistently measured

The measurement methodologies provide reliability when they can be calculated with a reliable way. Different persons should be in position to apply the method and result the same answers using same set of data. The condition which is required to verify this, is expressed as follows: “Will two different individuals in which is submitted the same question give the same answer with regard to the measurement of some size?”. These measurements should be differentiated from the “measurements” that they depend on the subjective crises of researchers and analysts that are reported as classifications, gradations or estimations.

A measurement methodology can ensure its consistence by recording the individual steps of measurement using a way that will be transparent and explicit to the person that will be asked to measure. Each measurement methodology should explain “how” each step should be applied and “why” it is applied with this particular way.

A particularly efficient way of maintaining consistence is the usage of questions of partial ignorance, that is to say questions that can be answered with “yes” or “no”. Another way is the use of automated processes that would follow each time the same process of measurement without procedural divergences.

3.2 Cheap to gather

Each measurement methodology needs time in order to calculate the results. All measurement methodologies begin row data and afterwards, following the precise steps of each model, generate some useful information. Hence, initially, somebody or something should collect the data from a suitable source, convert them to the desirable form, and finally calculate and format the results.

An efficient measurement methodology should collect those steps of transformation and format using a unified and fast process. If the process of measurement is insufficient, the method of collection of data can cost time and money to the organisation, which could have been spent in the analysis of results.

The high cost of measurement methodology can be caused by a series of factors such as the frequency of measurements, the complexity of process and its non automated nature.

It is therefore reasonable for a model of measurement to also make proposals on the most optimal candidate sources of data, in the light of saving time and money

3.3 Expressed as a number

All measurements should be expressed as an absolute number or percentage, which represents something that measures a quantity of size. Gradations such as “high, intermediate, low” or “1, 2, 3” (from a third degree scale) represent relative grades but do not also measure any size, therefore they cannot be used in a proper model of measurement. Thus, “expressed as an absolute number” implies the number of total elements and not the number that expresses the order of total elements.

Thus, measurement methodologies that are not expressed as numbers are not suitable for the measurement of security. Indicators such as traffic lights with the three possible values “red, yellow, green” they do not constitute some type measurement since they do not include some kind of numerical scale.

It should be noted that the colors of traffic lights can be used as depiction or presentation of the current state but in a more abstract level accompanying the necessary numerical data that should remain the main objective of security measurement.

3.4 Uses at least one unit of measure

Another basic requirement of measurement models of security is that all the related measurements should also include a relative unit of measurement, which will characterize the sizes that are been measured. For instance, the measurement “number of natural invasions in the IT building” uses as a unit of measurement the invasions. With the use of units of measurement, the researcher knows how to express similar measurements using the same way.

In certain cases it is better to use more than one measurement units aiming to facilitate the comparison of different applications. In the previous example the more general measurement unit can be also mentioned as the “number of individuals that tried to invade in the IT building”, which is also another unit of measurement. The use of this unit can be

more suitable for the comparison with another measurable size, that of "total number of individuals that enters in the IT building".

Another requirement for the good measurement methodologies is that they mean something to the persons that examine them. They could reveal issues of infrastructure or service under review improving or demonstrating the value of persons and processes for the organisation. Even if the close relation to the general context does not constitute a main requirement for a good measurement, it helps to maintain measurement results inside the framework of the organisation under discussion while making the results more useful. This should benefit the end recipients (which are usually the management executives of organisation) to comprehend the current security status and decide with rational way based on results of the measurements.

As an example it can be mentioned the use of measurement as "the mean number of attacks" for the entire organisation. This measurement can have the all above characteristics (consistence, numerical price etc.) but it does not help anyone to improve his work. If this measurement is differentiated and connected with the enterprising services that it offers, as the servers of an electronic trade service, it will be a much more important tool for the decision-making process of more specific sectors, such as the protection of specific servers but also the physical protection of personnel.

3.5 Partial weight

The quantification of an individual factor that influences security is without a doubt very useful. Nevertheless, another important issue is the effect of these individual factors to the security of an entire organisation.

This characteristic is related with the previous paragraph ("Contextually Specific") due of the relativity that is implied between measurable sizes and the overall security of the organisation. It differs however in the fact that the requirement of overall estimation includes the way and the size with which a specific measurement influences the security and the operation of organisation.

As an example, the measurement of "numbers of power failures" is precise and relative to security. However, the way with which this measurement influences the operation of organisation can become also the weight of this particular size. Thus, in the case where there is no way to tackle a power failure (eg. A power generator) the weight of this measurement concerning the total estimate of security should be a large figure.

It should also be clarified that the requirement of an overall estimate could also be considered an extension of measurements, which could even require some form of calculation.

3.6 Repetitiveness

The requirement of repetitiveness includes the measurement of same factors while applying the same measurement methods in different time periods. This repetition aims in the verification of previous measurements as well as in the observation and recording of the evolution of a particular size.

Thus, this repetition should not constitute measurement from one only person but be also verified from the measurements of different persons.

Regarding the evolution of factors that are measured, the measurements should be performed periodically, in order to detect unexpected changes, or immediately after a particular known change which will probably influence security.

So, the measurements should be calculated with a frequency proportional with the rate of change of process. The methodologies that use samples at regular time intervals can help the organisations to analyze the effectiveness of security in precise time intervals and prepare them to be in the position to react in time in case of a new security incident. As expected, in a decision for whether a measurement should be calculated often, the cost of measurement should be taken into account in terms of time and money. Alternatively measurements could be performed only before and after each change.

3.7 Comparability

Also, it is very important that one should measure and observe the improvement or the deterioration of security as time advances. For this reason, the results of measurements should be comparable to corresponding results of other organisations or different situations for the same organisation so that they can be contrasted to the current security status.

As reported above, a way to do that is to use common sizes and units of measurement. Additionally, it is possible to measure in equivalent or relevant time periods points that share common characteristics, such as measurement of number of robberies during the last and first day of each month.

4. Taxonomy of existing security methodologies

Currently, there are several approaches for Security measurement. Most of them tend to emphasize on different aspects of security than objective measurements. There are very few approaches that focus on providing the means for quantifying security. The most noticeable solutions are:

- Solutions based on Vulnerability analysis
- Solutions based on Penetration testing
- Solutions based on Baseline comparison
- Solutions based on Best-practice and standards
- Solutions based on Risk management

4.1 Solutions based on vulnerability analysis

Solutions based on Vulnerability Analysis such as Microsoft Security Analyzer are connecting the security status of a networked system to the number its network-related vulnerabilities. Unfortunately vulnerability analyzers can not be used to measure security of

an entire organization because they do not take into account many factors such as operational flaws and personnel security. Moreover, the results are not related at all to the number of actual security incidents.

4.2 Solutions based on penetration testing

Solutions based on penetration testing such as Corsaire Testing, are following the exact same patterns of the attackers without causing real damage to the systems. However the presented outcomes of these approaches are more like subjective ratings and gratings than objective measurements. Additionally, they always focus on the technology related aspects of the organization and neglect other important factors such as operational and physical security.

4.3 Solutions based on baseline comparison

The baseline comparison solutions contain standard security controls, which are applicable to the great majority of IT systems providing basic security. The basis for the decision of whether the organization or specific service fulfils the security requirements is based on the Auditor's personal judgment.

The main issues regarding this kind of approach are that it is very subjective and tends to change every time the Auditor changes. Again the outcome is more like a rating than a proper measurement.

4.4. Solutions based on best-practice and standards

These solutions (e.g. ISO/IEC 17799, BS 7799 and NIST SP 800-33) refer to several suggestions for security countermeasures and controls to improve an organization's information security. Although these approaches are quite thorough and explanatory, they are more useful when developing new infrastructures and services. So far the aspects of quantification and measurement have not been dealt with the same zeal.

4.5 Solutions based on risk management

These solutions assess security by describing, analyzing and evaluating single scenarios. Again, since the estimation of the risk is based on the Auditor's personal judgment, such solutions tend to be very subjective.

4.6 Combining the security's basic elements

Apart from the above categories two more categories of security measurement methodologies can be proposed. The first is concerned with the combination of security's basic elements. These basic elements are Integrity, Availability and Confidentiality. Other additional elements of IT Security are Non-Repudiation as well as Authentication.

An effort that could be included in this category is that of (Knorr, 2000). Within its framework, is proposed a structured approach for the analysis of metrics of security and for the quantification of the overall security of Electronic Business Applications. It uses a table that represents "overall security" and divides it to smaller parts. These parts correspond to

sites, potential targets and mechanisms of application security and are connected with the participating parts of an Electronic Business Application (customer, tradesman, means of communication). This process aims to the calculation of a quantifier of an Electronic Business Application, which functions as a means for the analysis, planning and comparison tool of similar applications.

Another approach that falls under this category is the one by (Serreli, 2007) that aims to offer the foundation for a model that could help security analysts to quantify and measure security. Comparing to the requirements that were initially set, the suggested model has supported the consistency requirement throughout the document by the use of questions with objective answers. The questions also aimed to answers which would be cheap to gather, since the answers could come from automated systems such as IPSs. Additionally the model has managed to express security as a number (percentage). Security calculation has used at least one unit of measure (such as blocked spam emails) satisfying another requirement of a proper quantification model. The last requirement has also been covered since the level of security of the overall enterprise or the individual services makes a lot of sense to management people. Thus the model can also claim to be contextually specific. On the downside, it should be pointed out that the model is not considered so much with the new products, but with the existing services. Other type of questions should be posed to calculate the security level of new services and/or products.

4.7 Combining factors that are related to security

The second category of security measurement methodologies is concerned with the combination of factors that are related indirectly to security. Approaches of this type, even if of limited number, can be grouped together if they measure or calculate elements which are not directly related to security unlike the previous category. They aim to describe the relation between security and factors that are easily and objectively measured.

A typical example of this category is the approach that is presented in (Campbell, 2003), where the economic impact of incidents of security it is examined. The economic impact is translated in fall of the stock prices of an organisation, as a result of the negative image that is created in the investment public. Thus the stock price constitutes a factor which can indirectly be related to the level of security of organisation.

This approach has been an important factor for the development of the approach proposed and presented in paragraph 6, which also aims in the development of methodology for the objective measurement of security using factors that are related indirectly to security.

5. From measurement to calculation

As it is defined by Webster dictionary, calculation is “deliberate process for transforming one or more inputs into one or more results, with variable change”. The term calculation is used in numerous sciences, from the precisely defined arithmetic calculation to the calculation of abstract concepts which is implemented with the use of special algorithms and suitable combination of factors.

Another, alternative way of calculation of sizes is also the statistical analysis, eg. the calculation of likely results of an electoral result.

In every case the calculation of a factor is advisable in cases where his direct measurement or the quantification of its size is not feasible. These cases mainly include abstract concepts that are not straightly measurable, with security being a very representative example.

The calculation of security is differentiated from the measurement of security. While the measurement is based on the collection and representation of primary data, the calculation uses primary data with a combinational way so that it produces a result which represents security.

Because of the abstract nature of IT security, a direct measurement will not have real and usable results since it will be based on the limitations of the methods reported in the previous chapter. The calculation of however of security can be more efficient by overcoming these limitations using measurable sizes.

There are various methods for the calculation of security. These can depend on the judgments and estimates of researchers thus they are labelled as classifications or gradations. A second category of methods can be based on statistical methods, which could lead to an estimate of the level of security. Other methods can combine the measurement of values with the co-calculation of which the value of security could be deduced.

The optimum method of calculation of security is differentiated depending on the specific needs of each organisation, service, infrastructure or system. However, in all cases, it should be selected with a process that will be based on well defined factors that should also have well defined relations between them.

It should be also clarified that the methods of calculation and the methods of measurement of security do not constitute alternative solutions from each other, but complementary. Precise measurement of security should not be considered feasible, due to its abstract nature, but due to the fact that calculation of security cannot be reliable if it is not based on measurable factors. The researcher of security should therefore use measurement methodologies that would result measurable factors. From the combination of these sizes the value of security will also be deduced.

It becomes easily understood that the calculation of security can be realised with two ways. These are the quantification of non measurable factors and abstract significances, such as security and the use of measurable factors that are measured with the appropriate security measurement methodologies. The latter involve measurable factors and are applied as a type of quantitative analysis. Also, the level of security can result as an estimate of the researchers which is applied as a type of qualitative analysis.

6. A new quantification methodology

Within the framework of the current research, a new approach of calculation of security was also created, which manages security quantification methodology as a value which is calculated with the combination of certain easily measurable factors. These factors are

related indirectly with security as well as with the level of security of specific business services.

This approach is based on the principle that the abstract concepts can be calculated with the combination of factors that is related indirectly with them and themselves can be easily measured. At the same time it seeks the satisfaction of following general requirements:

- Appreciation of security as factor that is an important part of the business production environment and not a collateral issue with minimal or no operational interest.
- Usage and combination of factors that can be measured or be quantified with objective ways.
- Connection of results of calculation of security with the business decisions.

For the application of the particular methodology of calculation of security, the sources of primary data should be determined in order to be combined for security calculation. This particular methodology considers that the value of security can result from the combination of parameters that are related with it. The factors that were selected are five and they all concern certain business, functional or commercial value. These five factors are mentioned as CARLS from initial their names in the English language. These are:

- Compliance: It expresses the percentage of conformity with the legal and regulatory framework that is applicable to the business service.
- Availability: It expresses the percentage of uptime of service in comparison its mission time.
- Return: It expresses the size of profits that results from the particular service.
- Liabilities: It expresses the size of economic losses due to the particular service.
- Stock price: It expresses the price of stock of the company that offers the service.

All factors are essentially different aspects of business services and should be available in order be composed in a value that would represent the level of security for each of the business services. The usage of these specific factors has two basic advantages, which are the main reasons for their choice in this model.

The first advantage is the fact that all factors are already have been measured by organisations for reasons of operational evaluation. This means that there is no need for additional effort in order to assemble the information required. The second advantage is the fact that each factor provides an objective image of the organisation and its particular business services, which objectivity can not disputed.

The measurement and monitoring of all factors is considered essential for the development, viability and proper operation of each organisation. It is exceptionally usual, in the great majority of organisations, to monitor and collect all the above factors. In many cases these factors are also measured for each business service separately. This fact makes the collection

of the necessary elements a simple, easy and feasible process. Moreover, the CARLS factors have a lot more meaning for the persons that are found in not technical positions and their mentality is directed from the operational needs of their organisations. Compared to other approaches this can be seen as an advantage because using this methodology the understanding but also the usability of the value of security is facilitated which is based on familiar notions and not on technical terms as “integrity” and “confidentiality”.

The main objective remains the calculation of the value of security of a specific service. This is implemented with the determination, the quantification and the combination of factors that can be measured easily and that is related indirectly with the security. This value portrays the level of security of specific service. The following paragraphs present the arguments in favour the choice of the particular factors, stating why CARLS factors are considered suitable for the calculation of security as well as an analysis for each one from them.

Compliance: The impact of non-compliance is profound. While many small issuers can operate with inconsistent compliance processes, problems eventually arise. Instead of focusing on the regulatory and punitive aspects of incomplete or ineffective compliance, this white paper will examine the functional impact of not remaining compliant with security regulations. Compliance can impact liquidity, which can affect your ability to raise funds for growth. The difficult aspect of compliance is knowing everything you have to do, when and how. Ongoing compliance requires an investment – for the same reasons as the initial compliance work you did when going public. The liquidity opportunities that initially attracted your company to the publicly traded arena are the same reasons for remaining compliant.

Availability: Government organizations and businesses of all sizes need to create and implement comprehensive business and operations continuity plans. Most organizations understand that they need to protect their data and systems -an activity known as disaster recovery. The disaster recovery is only half the battle-enterprises need also be prepared to quickly and seamlessly restart business processes in order to continue operations.

Return: The owners of a company and the company's creditors share a similar goal: to increase wealth. They are thus very concerned about profitability in all phases of operations. Creditors are specifically concerned that the company use its resources profitably so that it can pay interest and principal on its debt. Owners are concerned that the company be profitable so that stock values will increase. Company managers must show they can manage the owners' investment and produce the profits that owners and creditors demand. Because top management must meet the profit expectations of company owners, it passes down to the lower levels of management those profitability goals, which are then spread throughout the company. All managers, therefore, are expected to meet profitability goals, which are often increased and tightened as each level of management seeks a margin of security.

Liabilities: Most health related businesses would agree that securing insurance is one of the basic costs of doing business. As responsible business owners, they have budgeted for the appropriate coverages as a precautionary measure in the event of a loss— especially a catastrophic loss. However, there are a few optimistic souls who think of business insurance

as an option. These risk takers appear perfectly content to operate their salons with little or no coverage in place. Unfortunately, most financial experts agree that this is a very dangerous practice, as those who gamble and lose usually pay a much higher price in the long run. The bottom line is, while none of us ever expects to get sued, we've all got to accept that even the most adept operator may face litigation for any number of reasons. While there are a wide variety of insurance coverages available for protecting yourself and your business, one of the most essential is liability. Liability is an especially important issue for those in the tanning industry, whose business is providing customers a service which may pose some risk of injury to them. Liability risks come in many more forms than might be expected. In addition to liability arising specifically from the use of tanning equipment, salon owners also may be held accountable for a variety of other kinds of business liabilities, such as a customer slipping and falling.

Stock Market: By using a stock market return framework to examine the economic implications of information security breaches, [3] study contributes to the literature examining the economic effects of information security breaches. We find there evidence of an overall negative stock market reaction to announcements of information security. The economic cost of publicly announced information security breaches 445 breaches in major newspapers, although this finding is not robust across all specifications. Nevertheless, these results provide some support for the argument that information security breaches adversely affect the future economic performance of affected firms.

The choice of the above factors satisfies the two of the three basic requirements that had been placed within framework of the current approach of for the calculation of security, that is to say to appreciate security as a factor that is part of the business production environment as well as to use and combine measurable factors that can be quantified with objectively.

Based on the ideas described in the previous paragraph, a figure that would represent the level of security for a specific service should take into account all security factors presented. In order to mathematically express a formula that calculates security, several assumptions have been made.

Firstly, the notion of Target level for each factor is introduced. The target level is set by the upper management who is responsible for the overall operating constrains, such as security, of each business service offered. Within this context, the target level of Compliance, Availability, Returns and Stock Price are set. The target level of compliance can be the compliance to a specific industry directive or governmental law or even an international standard. Similarly, the target level of Availability should be defined taking into account the business needs of each service.

The Current level of Compliance will be represented as a "Yes" or "No" factor in order to keep the model as simple as possible. A future extension to that model can express the current compliance level as a percentage, signifying that a service does or does not cover all compliance needs (e.g. covers a law but not a specific international standard).

So, the Security figure for each service can expressed as a function of independent variables by the use of following formula:

$$S_s = \frac{C}{C_T} \times \frac{A}{A_T} \times \frac{R-L}{R_T} \times \frac{S}{S_T} \quad (1)$$

Where:

S_s = Security level of a Specific Business Service

C = Current Level of Compliance [0 | 1]

C_T = Target Level of Compliance (0-1)

A = Current Level of Availability [0...1]

A_T = Target Level of Availability [0...1]

R = Current Return of the Service (in a monetary value)

L = Current Liabilities of the Service (in a monetary value)

R_T = Target Return of the Service (in a monetary value)

S = Current Stock Price (Represents the company brand) (in a monetary value)

S_T = Target Stock Price (in a monetary value)

Having expressed security using the formula above, a proper usage of the results could be to a financial motivated one. Our target is to balance the spending in Security for each Business Service in order to maximize the organisation's Return on Investment:

$$\max ROI(S_1, S_2, \dots, S_n) \quad (2)$$

subject to

$$\sum_{i=1}^n (R_i - I_i - L_i) > 0 \quad (3)$$

Where:

S_n = Security level of a Business Service

R_i = Revenues of a Business Service

I_i = Total Investments in a Business Service (Part of which is Security spending-investments)

L_i = Total Liabilities of a Business Service

Using the last formula another objective is achieved. This is the connection of results of calculation of security with the business decisions that are related to security investments. In other words this formula aims to answer the question whether a security investment would cost more than the expected benefits.

7. Conclusion

The question that was analyzed in this chapter is whether and how the principles of the security measurement methodologies can be applied so that the objective measurement of security of business services can be achieved. The motives that support this question are focused in the justification of expenses and investments that are related with to security. Thus, although the management of security is closely related to technical and organisational

level it is often difficult to define a quantified “version” of security that would be more comprehensible and usable in operational level.

This chapter also presents a critical evaluation and categorisation of the requirements of measurement and calculation methods of security, on which the restrictions of approaches that exist are based. Additionally a new security calculation approach is developed that attempts the quantification of security with the use of factors that are related indirectly to security.

The basic principle that was followed is that the research focus should be moved from the measurement of security to the calculation of security. Other principles were:

- Appreciation of security as factor that is an important part of the business production environment and not a collateral issue with minimal or no operational interest.
- Usage and combination of factors that can be measured or be quantified with objective ways.
- Connection of results of calculation of security with the business decisions.

The approach for the quantification of security is implemented via calculation. The variables that are used for the calculation are the CARLS factors, that is to say Compliance C with the legal and regulatory framework, Availability A of business services, Return R of each business service, Liabilities L due to specific services and the Stock price S of the organisation that reflects its fame and public image. The methodology supports that the security is mirrored in each one from these factors and hence the factors are related indirectly to that. This connection is expressed with a mathematic formula through the use of which the factors are considered equivalent.

An important advantage of the methodology is that through its use the management executives can comprehend more immediately the values that are produced by this and evaluate better where they should focus and support the security investments. This is particularly important because Security is an abstract concept which it is not easy to be expressed as a measurable value.

One of the restrictions of method is the fact that the all factors they are considered equivalent during the calculation of security. A future research could investigate further the degree that security is influenced from every parameter as well as how this is altered in terms of service, organisation or market type.

8. References

- Danahy, J. (2004) *The Need for Metrics and Measurement in Application Security*, OWASP Metrics and Measurement Standards
- Fisher, D. (2009) Experts call for better measurement of security, Blog, recovered: 14/6/09, <http://www.threatpost.com/blogs/experts-call-better-measurement-security>
- Sonnenreich, W.; Albanese, J. & Stout, B. (2006) Return On Security Investment (ROSI) – A Practical Quantitative Model, *Journal of Research and Practice in Information Technology*, Vol. 38, No. 1, February 2006
- Maizlits, B. & Handler, R. (2005) *IT Portfolio Management: Step by Step*, John Wiley & Sons, ISBN: 978-0-471-64984-7, US
- Sommerville, I. (1996) *Software Engineering*, Fifth Edition, Addison-Westley, ISBN: 978-0201427653, UK
- Tipton, H. & Krause, M. (2008) *Information Security Management Handbook*, Sixth edition, Auerbach Publications, ISBN 978-1420067088
- Olzak, T. (2007) The Pros and Cons of Security Risk Management, Tech Republic, recovered: 14/6/09, <http://blogs.techrepublic.com.com/security/?p=180>
- Parker, D. (2007) Risks of risk-based security, *Communications of the ACM*, Volume 50, Issue 3, March 2007, pp 120.
- Jaquith, A. (2007) *Security Metrics: Replacing Fear, Uncertainty, and Doubt*, Addison-Wesley Professional, ISBN 978-0321349989
- Campbell, K.; Gordon, L. A.; Loeb, M. P. & Zhou L. (2003) The economic cost of publicly announced information security breaches: empirical evidence from the stock market, *Journal of Computer Security*, Volume 11, Issue 3 (March 2003) pp 431–448
- Serrelis, Em. & Alexandris, N (2007) An Empirical Model for Quantifying Security Based on Services, IEEE Computer Society, *Proceedings of the International Multi-Conference on Computing in the Global Information Technology*, pp 30
- Knorr, K.; Rohrig, S. (2000) Security of Electronic Business Applications: Structure and Quantification, *Proceedings of the 1st International Conference on Electronic Commerce and Web Technologies EC-Web 2000*, pp 25-37

A testing process for Interoperability and Conformance of secure Web Services

Spyridon Papastergiou and Despina Polemi
*University Department of Informatics, University of Piraeus
80 Karaoli & Dimitriou Str, 185 34 Piraeus, Hellas*

1. Introduction

The design, development and implementation of electronic (e-) services relying on XML and Web Service (WS)-based technologies is the current trend in achieving interoperability. E-services can be offered either as autonomous Web Services or embedded in Service Oriented Architectures (SOAs) (High et al., 2005).

In this context, despite the fact that applications with similar business goals adopt the same technical standards, quite often their interactions capabilities are extremely limited. Thus, application developers show an increasing concern for evaluating interoperability between common services which are offered either autonomously or through a SOA. The creation of a proper framework (EIF) has a significant importance in the evaluation of interoperability of such services and is accomplished by the precise definition of the applied standards and guidelines which guarantee the interaction of the services. Existing testing methodologies developed by various organizations (e.g ISO/IEC 9646, ESTI) treat the interoperability of services as a generic problem. They merely provide guidelines and describe high level testing procedures that can be applied to test interoperability of various telecommunication as well as software and data communication systems. Most Web Service-oriented methodologies (i.e. WS-I, ebXML IIC framework) demonstrate weaknesses as they are not capable of testing all the required aspects that compose an interoperability framework and mostly the security aspects of the message content.

Additionally, in literature, specific testing types (Saglietti et al., 2008) have been presented defining diverse testing approaches that treat the applications under test either as white boxes having full knowledge of the software or as black boxes without any understanding of their internal behaviour or even as grey boxes with limited knowledge of their internal architecture. The nature the WSs (e.g. geographic distribution of the examined WSs and dependencies with external trusted third parties) plays an important role in the adoption of the most appropriate testing type as they raise specific challenges that should be underlined and taken into account.

Therefore, there is a specific need for targeted methodologies and frameworks that check and guarantee the end-to-end application interaction capabilities of common Web Services and follow and deploy the most appropriate testing strategies covering all WSs aspects. Identifying this need, this paper proposes a well-formed grey box testing methodology

entitled ICoM, able to test whether various services achieve communication effectively based on the adopted standards. It defines the precise structure of the involved parties, specifies distinct steps to follow, describes concrete tests that should be applied, and enables the execution of specific testing suites. ICoM has been applied in order to evaluate the interoperability of the existing autonomous SELIS e-invoicing service (Kaliontzoglou et al., 2006) and the SWEB e-invoicing service embedded in a SOA-based platform (SWEB), (Karantjias et al., 2008).

2. Prior Work

This section presents the existing testing methodologies and frameworks illustrating their weaknesses and indicating the need for a more holistic methodological framework. Widely used types of testing are also described identifying the most appropriate method that should be applied to WSs due to their inherent characteristics.

2.1 Existing Testing Methodologies and Frameworks

Traditionally, interoperability testing methodologies for the Internet Protocols have been used extensively in the telecommunication industry and in the Internet world. For example the Open Systems Interconnection - Conformance Testing Methodology and Framework (ISO/IEC 9646) (OSI, 1997) is a widely spread and successfully applied conformance testing methodology which has evolved over the years. Nevertheless, it is considered as overly generic framework that allows a high degree of freedom and gives little practical guidance (ETSI, 1998).

The European Telecommunications Standards Institute (ETSI), acknowledging the importance of the testing methodologies, has also contributed towards this direction. ETSI defined a more integrated framework that consists of two primitive types of tests, the conformance and the interoperability testing. The ETSI conformance testing (Moseley et al., 2003), (ETSI, 1995) is based on the ISO/IEC 9646 using its principles as a basis, but not as strict guidelines. It focuses mostly on making easier, more applicable and more readable the use of test suites in the proposed methodology. Therefore, ETSI defined a core language the Testing and Test Control Notation TTCN-3 (ETSI, 2002), (Dibuz & Kremer, 2003) that can be used for the specification of test suites which are independent of test methods, layers and protocols.

On the other hand, ETSI interoperability testing (ETSI, 2007) constitutes a generic approach merely providing guidance on the specification and execution of the interoperability tests. These guidelines are in the form of recommendations rather than strict rules. The TTCN-3 can also be used in this kind of testing offering a higher level of flexibility.

Despite the independent operation of both types of testing, they are closely connected satisfying different objectives (Kulvatunyou et al., 2003). Conformance Testing checks to what extent a solution conforms to the corresponding specification or standard. Interoperability Testing proves the end-to-end functionality between the solutions. It should be noted that the use of either type of test does not guarantee interoperability.

A significant limitation of both abovementioned methodologies is that they treat the interoperability of the WS-based services as a generic problem providing generic testing practices.

Currently, there are only two widely used WSs testing methodologies, WS-I and ebXML: The Web Service Interoperability Standardization Organization (WS-I) (Seely & Lauzon, 2005), (Ehnebuske, 2003) is an industry consortium chartered to promote WS interoperability across platforms, operating systems, and programming languages. It has released a number of profiles and testing tools that compose a scalable testing environment. The shortcomings of this methodology are the following:

1. WS-I does not support the definition of specific and discrete test cases that should be followed during the tests.
2. WS-I tools achieve to monitor only the message flow, without being able to control the testing execution.
3. WS-I tools do not achieve to support a wide range of evaluation criteria (interoperability areas or multiple types of testing). They achieve to test only the conformance of the exchanged messages and the defined services' descriptions against the appropriate standards. They fail to cover criteria regarding the interoperability and the conformance of the applied security features of the message content and the transformation of the exchanged documents between different formats.

OASIS ebXML (ebXML, 2001) is an end-to-end B2B XML framework that provides concrete specifications for dynamic B2B collaborations. It has specified the Implementation, Interoperability, and Conformance (IIC) test framework (OASIS, 2003), (Lee, 2005), (Kim & Yun, 2003) describing the required architecture and providing the necessary test material to be processed by the architecture, a mark-up language and format for representing test requirements, and test suites (set of Test Cases). This approach has the following shortcomings:

1. ebXML IIC is intended to support conformance and interoperability testing only for ebXML specifications and implementations.
2. ebXML IIC does not impute the responsibility of a communication failure to the corresponding system.
3. ebXML IIC does not cover criteria regarding the interoperability and the conformance of the applied security features of the message content and the transformation of the exchanged documents between different formats.

Despite their weaknesses, the above frameworks can be used as the basis for the development of enhanced and more targeted methodologies to test and guarantee the interworking of common WS-based applications.

2.2 Testing Types Existing Testing

Software testing, including interoperability testing, may take place at different levels of depth, depending on the actual known technical details of the application under test. The bibliography acknowledges three main types of testing: black box, white box and grey box testing (Peyton et al., 2008).

- Black box testing treats the software as a black-box where only the inputs and outputs of the black box are tested without any understanding of the internal behaviour or the adopted specifications. It aims to test the functionality according to the requirements. Thus, the tester inputs data and only sees the output from the test object based on the objects published and known interfaces.

- In White box testing, the tester has access and knowledge of the internal data structures, code, and algorithms of the software. A White Box tester typically analyzes source code, derives the corresponding test cases and targets specific code paths to achieve a certain level of code coverage.
- In recent years the term Grey Box testing has come into common usage and it refers to a technique of testing the system with limited knowledge of its internal architecture. The tester usually has access to specification documents (beyond simple requirements) and generates tests based on information such as state-based models or architecture diagrams.

The selection of the most appropriate testing type relies on the nature of the application under test. Web Services (WSs) are composed by a set of related and integrated services (High et al., 2005). The main characteristics of these services are the following (Rizwan & Mamoon, 2007):

- They can be distributed in the sense that they are located in different geographic areas, having independent capabilities and are accessible only via specific interfaces that are described by WSDL documents.
- They can be implemented in diverse programming languages and support independent operating systems.
- They can be chained with dependencies on other trusted third parties such as PKI and timestamp authorities.

The business design of WSs is essentially a composition of these repeatable services which represents and forms the desired business logic. The above features of the WSs introduce several challenges with respect to testing. White box testing especially in a SOA environment is quite impractical to perform due to the nature of the WSs that make access to source code or binaries very difficult. On the other hand, the WS interoperability testing methodologies (WS-I, ebXML IIC) adopted in the frameworks analyzed in the previous paragraph only enable black box testing of WS, leading to limited and inefficient test coverage due to the “blind” nature of that type of testing.

The distributed nature of Web Services makes Grey Box testing ideal for detecting interoperability flaws on the communication channels between WSs, mostly by leveraging the rich information contained in the descriptions of the services interfaces (WSDL documents). A Grey Box Tester is able to identify at a high level the internal structure of the tested WSs, defining the composed services and accessing the provided interfaces having limited or even no access to the actual code. Additionally, several test cases regarding the deployment of the security features and the communication protocols can be performed covering all the aspects of the WSs. Therefore grey box testing has been adopted as the most appropriate type to use in the methodology proposed in this paper.

3. The ICoM methodology

In this Section, we will describe the structure of the proposed methodology. Before presenting its main features and implementation steps, it is imperative to present the requirements it satisfies which overpass the weaknesses of the existing frameworks. The requirements that ICoM satisfies are:

- **Clarity:** the methodology specifies concisely the evaluated entities and the required information items for the testing process (e.g. test cases and test data).
- **Adaptability & Extensibility:** the methodology is extensible in the sense it may easily evaluate new aspects of the WSs and adopt and integrate new testing tools and libraries.
- **Flexibility:** the methodology is parameterizable in the sense that different parts of the methodology can be adopted for the realization of specific sets of test suites.
- **Structural:** the methodology is structured and comprises a definite and precise set of implementation steps.
- **Independency & scalability:** the methodology offers a high level of independency from testing technologies, the number of entities involved and the platforms hosting systems under test. This fact will offer the possibility of testing interoperability on several different WSs.

In order for the methodology to accomplish its objective goals comprises of four distinct phases:

- *Phase 1 “Entity identification and setting”:* the entities involved are identified and their specific setting and order for the tests is defined. The involved entities are the Systems under Test (SuTs) being evaluated for interoperability and conformance and the Test Coordination Infrastructure (TCI) which monitors the testing suites applied by the SuTs in order to identify any erroneous behaviour.
- *Phase 2 “Entity structure definition”:* the structure of the involved entities is identified. This includes the SuTs, which consist of the actual WS under evaluation and the specific structural additions of the testing infrastructure that are required in the testing procedure. The internal structure of the WS is immutable and an analysis of the available services is carried out. This phase also includes the specification of the precise structure of the TCI.
- *Phase 3 “Conformance testing”:* the definition and generation of the executed conformance test cases for each SuT, based on which the parameterization of the SuTs structural additions and the TCI is completed. Each SuT is evaluated against the adopted standards.
- *Phase 4 “Interoperability testing”:* the formulation and the derivation of the interoperability test cases, performing the necessary parameterization of the the SuTs structural additions and the TCI. The actual interoperability testing between all SuTs’ communications takes place during this phase.

The following section describes these four phases in detail.

3.1 Phase 1: Entity Identification and Setting

As shown in the Figure 1, the initial step is the identification of the exact number, order and setting of all entities participating in the testing suite. The full methodology deployment demands the execution of the following process:

- *Entities (SuTs) definition:* defining the entities (SuT 1... SuT N (Figure 1), TCI) that participate in the test.
- *Web Service (WS) declaration:* declaring the Web Service that will be tested.

- *Testing Scenario*: specifying a testing scenario. A testing scenario consists of a sequence of actions described at a high level which must be performed by the SuTs in order to execute a specific electronic transaction.
- *Role determination*: defining the entities' role based on the specified testing scenario. Specifically, the entity that initializes the execution of the testing suite and the sequence of the other entities, as this is described in the testing scenario.

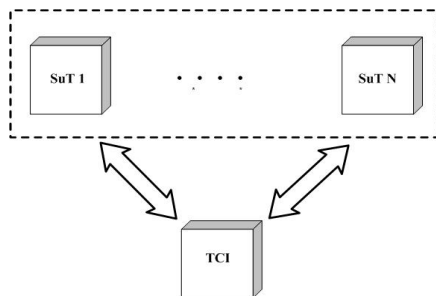


Fig. 1. General Entity Setting

The TCI, as presented in Phases 3 and 4 of the proposed methodology, acting as the initiator and an intermediate entity of the testing suite analyzes the testing results and indicates the conformance and the interoperability of the entities under test.

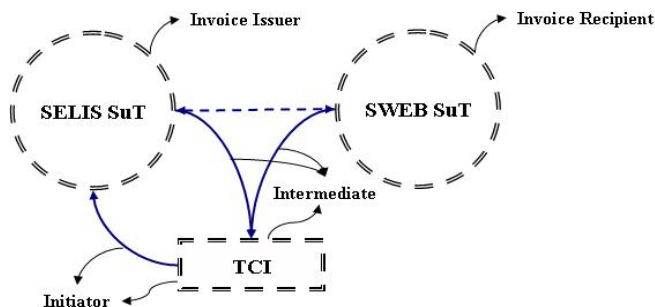


Fig. 2. Example Entities Identification

In this paper, we will use ICoM to test the interaction capabilities of two existing and fully operational WS-based solutions for e-invoicing, the autonomous SELIS service (Papastergiou et al., 2007b) and the SOA-based SWEB service (Papastergiou et al., 2007a). These services shall be used as demonstration material (case study) throughout the methodology presentation to better show its capabilities with specific examples. A case study example that imprints the main steps of the current phase is the following:

Example 1

The involved entities that participate in the test, Figure 2, are the SELIS SuT and the SWEB SuT while the provided invoicing service constitutes the Web Service that will be tested. The testing scenario we shall be following involves a communication initiated by SELIS SuT,

acting as the Invoice Issuer (*role: Issuer*) that invokes SWEB SuT, which acts as the Invoice Recipient (*role: Recipient*) handling the dispatched invoice and responding with an acknowledgment. This scenario covers specific aspects of an invoicing transaction such as the issuance, dispatch and receipt of an invoice document. The TCI triggers, monitors the execution of the applied testing suites.

3.2 Phase 2: Entity Structure Definition

During this phase, the actual structure of the involved entities, TCI and SuTs, should be defined and shaped precisely by the responsible operators, indicating their constituent parts. The final form of each entity depends upon the tests that will be performed and the suites that will be followed.

3.2.1 Definition System under Test (SuT)

A SuT primarily contains the WS to be tested (WSuT). A common WSuT, as depicted in Figure 3, includes a set of primary services $\{S_1, \dots, S_n\}$ that are combined to form and execute the WS's business logic. Each service encapsulates an explicit function having a number of Inputs and producing a specific Output. It may also interact with other services in order to complete its objective goal.

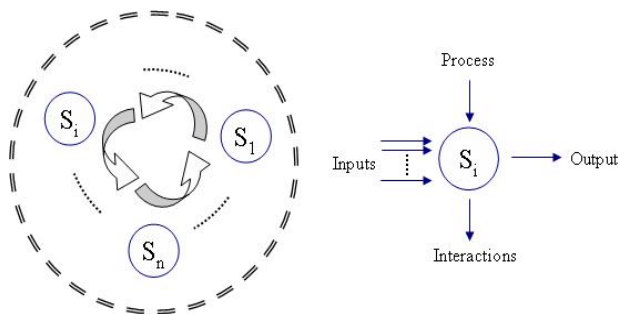


Fig. 3. Web Service under Test (WSuT)

A representative example of an advanced WSuT (Papastergiou et al., 2009), (Karantjias et al., 2008) may include a set of primary services such as the following:

- communication with legacy systems and proper data transformation,
- document management and application of digital signatures and / or encryption as a secure mechanism on the business level, and
- messaging formulation, processing and application of security mechanisms on the messaging level.

All these services, following concrete procedures, manage and derive data that are based on specific standards. The factual conformance of these data to the individual requirements specified by the corresponding specifications and the ability of other systems to handle these data correctly constitute factors that are able to leverage and infer the interoperability and the conformance capabilities of a WSuT. Take for example the case where the security service of a "WSuT A" signs a XML document according to the W3C XML Digital Signature

standard. Initially, the produced signed document should be tested for compliance with the principles of this standard and then there should be test that another “WSuT B” is able to successfully verify that signature. The conclusion that comes from this example is that the “WSuT A” produces signed documents that conform to the W3C XML Digital Signature standard and is able to interoperate successfully with the “WSuT B” on the application of signature at the business level.

Thus, in order for a WSuT to derive the required evaluation data, it should perform a certain number of activities as represented and indicated by specific test cases. A test case comprises a set of conditions and variable and depicts the test control logic which is sufficient for the execution of a testing suite.

ICoM deals with the execution of test cases and the collection of the respective data defining two structural components the *Service Orchestration Engine* and the *Report Engine* as seen on the following figure. These components are part of the testing infrastructure and along with the WSuT compose an integrated SuT.

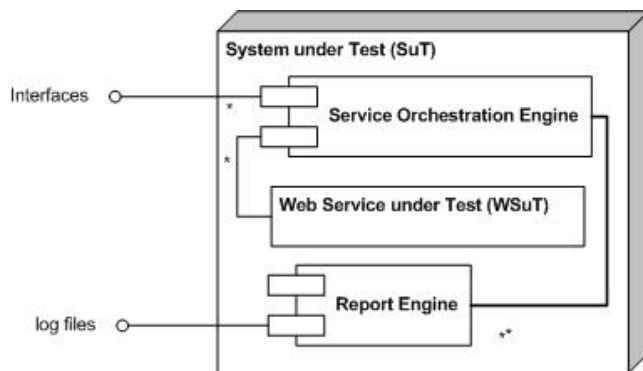


Fig. 4. System under Test (SuT)

The Service Orchestration Engine is responsible to coordinate the available WSuT’s services to perform specific workflows. Each workflow is represented by a predefined BPEL process which imprints the logic of an executed test case. The BPEL test cases that will be adopted and executed are designed during phases 3 and 4 where the actual conformance and interoperability testing is performed. In these phases, the Engine undergoes the appropriate parameterization in order to run the defined BPEL processes. The parameterization includes also the provision of specific interfaces that can be invoked by the TCI in order to initiate the execution of a BPEL test case promoting the concept of automation of the testing procedure. Thus, the Service Orchestration Engine provides a thorough control of the testing suite and enables the effective accumulation of the evaluated data by adopting BPEL processes as forms to represent and execute test cases. The nature of the accumulated data depends on the type and the purpose of the executed test case and varies from simple XML documents and documents with security features to secure or not SOAP messages and supported XSLT files used for document transformation.

The second structural component of the SuT, the Report Engine acts as the collector of the evaluated data. Its main responsible focuses on gathering and consolidating them in log files.

Concluding, ICoM defines four major steps that lead to the definition of a SuT.

1. *service identification*: Identification of the primary services of the WSuT.
2. *interface analysis*: Analysis of the services' interfaces and data types used per service, extracting the messages types of the inputs and outputs as described in the WSDL documents. This may include an extensive analysis of several WSDL documents.
3. *standards definition*: definition of the standards that each WSuT adopts and implements.
4. *testing components adaptation*: adaptation of the two structural testing components, the Service Orchestration Engine and the Report Engine to the WSuT.

It should be noted that the above steps are completely independent of the underlying infrastructure and the WSuT implementation language and do not bind ICoM to a specific technological solution. The application of these steps in our demonstrated case study is depicted in the following example:

Example 2

The SELIS and SWEB SuTs are setup following the aforementioned four basic steps. In SELIS, we have identified three specific areas of services with diverse functions. These services are divided into:

- *Basic services*, which provide the basic functions used to perform primitive tasks. SELIS includes document management services, message and document transformation services, message forwarding services, publication and query services and notification services.
- *Security mechanisms and services* that address the security requirements of SELIS-invoicing. It supports the following security mechanisms: digital signatures, advanced electronic signatures, encryption, timestamping and credential management.
- *Infrastructure support services*, which manage the connection with the back-office systems such as databases, wrapper software on top of legacy systems, existing ERPs etc.

Each of these services has specific known interfaces as described by the corresponding WSDL documents having concise inputs and giving a concrete output. The standards that adopted in the services implementations are the following:

- the XML common business library version 4.0 (xCBL 4.0) and the Exact ERP XML schema, for the representation of the invoice information,
- the Extensible Stylesheet Language (XSL) Transformations, for the transformation of the invoice documents,
- several security standards such as W3C XML Encryption, W3C XML Digital Signature and XML Advanced Electronic Signature (XAdES) for the application of the appropriate security features on the business level and
- SOAP and the WS-Security standard for Web Services invocation and
- the Web Services Description Language (WSDL) for the description of invoked services.

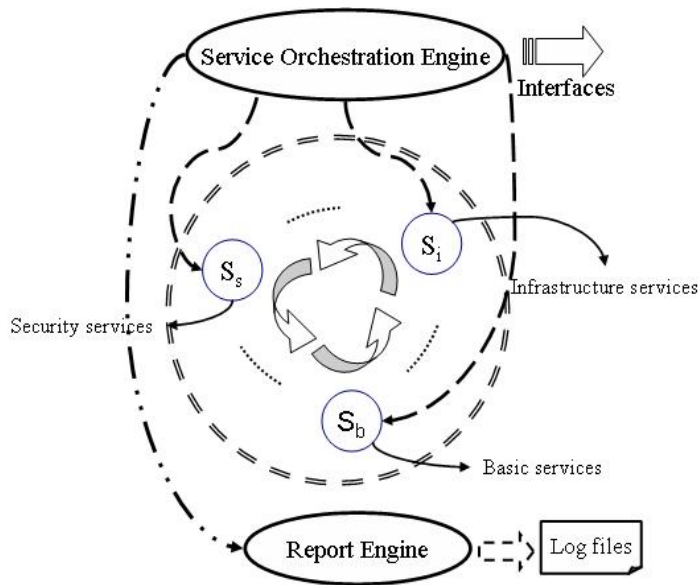


Fig. 5. SELIS SuT

Figure 5 depicts the overall form of the defined SELIS SuT which is composed by the two proposed structural testing components, the Service Orchestration Engine and the Report Engine and the SELIS WSuT. The process is similar in the case of SWEB.

3.2.2 Test Coordination Infrastructure (TCI)

The logical and organizational structure of the TCI is independent of the nature of the SuTs. On the contrary, technologically, it offers significant flexibility and scalability allowing the upgrade of already adopted testing tools and libraries and the integration of new more advanced ones. This upgrade enables the testing of different systems in various aspects.

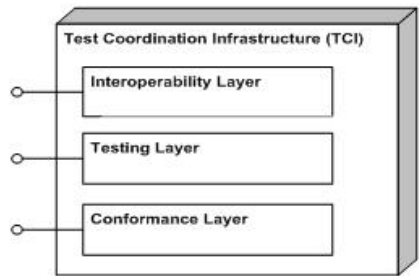


Fig. 6. Test Coordination Infrastructure

The general structure and functionality of the TCI is depicted in the Figure 6. Three fundamental layers, the *Testing Layer*, the *Interoperability Layer* and the *Conformance Layer* compose the TCI at a high level.

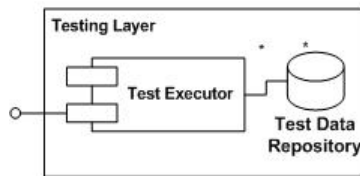


Fig. 7. Testing Layer

The Testing Layer (Figure 7) consists of the *Test Data Repository* and the *Test Executor*. The Test Executor orchestrates the execution of the deployed BPEL test cases in ICoM phases 3 and 4, based on a predefined schedule formed during these phases. Following this schedule, the Executor invokes sequentially the interfaces provided by the SuT Service Orchestration Engine in order to initiate the execution of the corresponding BPEL process. The invocation of the interfaces may require as input specific parameters (test data) which are retrieved from the Test Data Repository. As we will see in the following Section, the required test data rely on the nature of the deployed BPEL processes and are created along with the definition of the test cases.

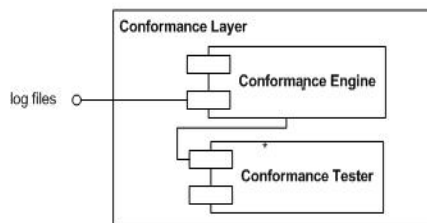


Fig. 8. Conformance Layer

The next layer of the TCI is the Conformance Layer (Figure 8) its main responsibility is to evaluate the conformance of each SuT against the adopted standards in phase 3. The evaluation process is performed based on the data produced by the SuT during the execution of the BPEL test cases. This layer contains two sub-components, the *Conformance Engine* and the *Conformance Tester*.

The Conformance Tester consists of a set of testing tools and libraries that actually assess the conformance of these data to the corresponding standards. The exact tools and libraries that should be adopted depend on the SuT's standards that have been defined during the formulation of a SuT (step 3). These range from tools that verify the conformance of a XML document to the corresponding XML schema and libraries that validate the security features of the XML documents to tools that check the conformance of the exchanged SOAP messages and message descriptions against the respective standards. Therefore, ICoM as methodology is quite extensible allowing the integration of new more updated tools/libraries in the Conformance Tester.

In the current phase, the steps that should be performed are the following:

- ✓ the identification and the integration of the required tools in the Conformance Tester,
- ✓ the implementation of the appropriate tools' interfaces that enable Conformance Engine to invoke them. In case that the required interfaces can not be implemented due to the

nature of the testing tools, a manual process is performed for the communication of these components.

In this layer, the Conformance Engine acts as the recipient and the analyzer of the log files that are produced by the SuTs in order to specify the tools that should be used for the evaluated data respectively.

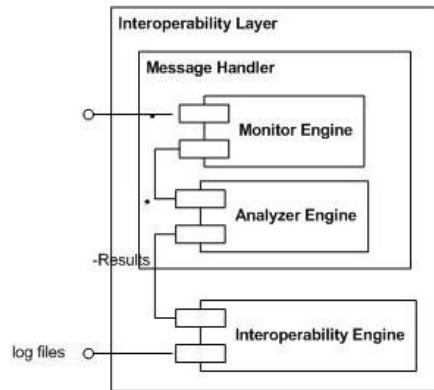


Fig. 9. Interoperability Layer

The Interoperability Layer (Figure 9) is the last but equally important layer comprising the *Message Handler* and *Interoperability Engine*. The *Message Handler* operates as the intermediate node during Interoperability testing (phase 4) that intercepts and delegates the exchanged messages to the SuTs. It comprises the *Monitor Engine* which intercepts the SOAP messages exchanged among the SuTs and the *Analyzer Engine* that determines whether these messages conform to their corresponding message descriptions.

The second subcomponent of the layer, the *Interoperability Engine*, is the actual interoperability consolidation point. It obtains as input the log files that are produced by the SuTs and the *Analyzer Engine*'s conformance results providing them to the test operator. Based on these results, the test operator is able to do the following:

- ✓ notify the SuTs' interaction possibilities,
- ✓ detect the inaccurate points and
- ✓ impute the responsibility of the communication failure to the corresponding SuT.

Deductions are made by carrying out a process that consists of three main steps. The first one is to verify that the SuTs possess and handle the evaluated data without the detection of any erroneous behaviour. This means that the evaluated data are valid for all the SuTs. The second step includes a comparison process. During this process the exchanged data (e.g. exchanged SOAP messages and documents) are compared in order to acknowledge that the SuTs possess the same data. In the last step, the test operator confirms that the exchanged messages are part of the business processes that the SuTs have defined in their service descriptions. These steps enable the operator to identify whether the SuTs are interoperable and to what extent during the last phase of the proposed methodology.

In the following example we present the form of the TCI in our working example.

Example 3

The TCI has been shaped as presented in Figure 10. All the layer's components, Conformance Engine, Test Executor and Interoperability Engine have been implemented in Java technology. In the Testing Layer, a native XML Database, eXist has been adopted as the Test Data Repository enabling the storage of XML-based test data. The Conformance Tester of the Conformance Layer is shaped based on the SELIS and SWEB standards adopting testing tools that perform the conformance evaluation. Representative tools include testing environment and libraries for XML signature validation (IBM XML security suite (IBM XML), IAIK XML signature library (IAIK XML), IAIK XAdES toolkit (IAIK XAdES)) and tools for XML Schema document conformance (Altova XMLSpy).

The WS-I Tools (Brittenham, 2003), monitor (Brittenham et al., 2005) and analyzer (Brittenham, 2005), have been used in order to operate as the two subcomponents of the Interoperability Layer's Message Handler, Monitor and Analyzer Engine correspondingly. These tools provide an unobtrusive and automated way to log and analyze Web Service messages producing the respective conformance results that will be used for the interoperability evaluation.

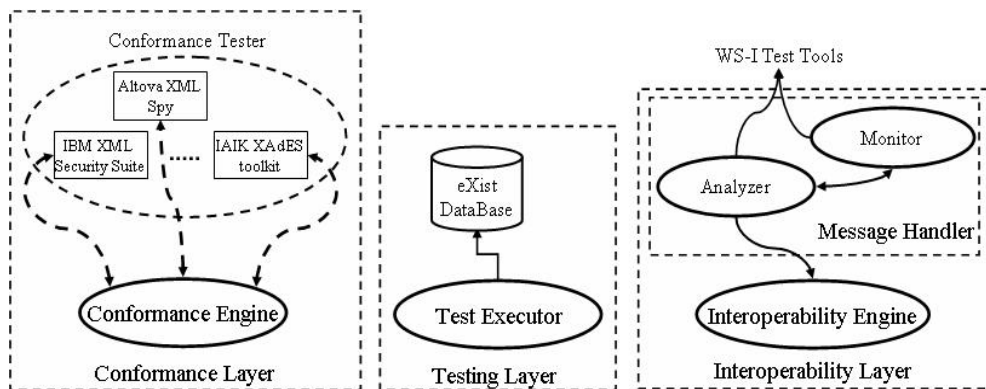


Fig. 10. TCI Working Example

Based on the above mentioned form of the TCI, the SELIS and SWEB SuT are able to be assessed taking into account the corresponding evaluation results that are derived from the execution of the defined test case in phase 3 and 4 of the proposed methodology.

3.3 Phase 3: Conformance Testing

Conformance testing has the primary goal of testing a WS against the standard it implements. This testing type involves a single SuT plus the TCI. The main steps that ICoM proposes are:

- Definition of Conformance Test Cases:** In Section 0, we specified that ICoM adopts a BPEL representation of the deployed test cases. Our decision was based on the identification of a number of limitations (Pentafronimos et al., 2008) that the existing test case languages present and the nature and features of the BPEL language. Generally, BPEL allows the definition and the representation of specific business flows in a XML format. It has unique features in both syntax (e.g. flow with activity

synchronization, join condition) and semantics (e.g. dead-path-elimination) that make BPEL a highly compact and expressive language. In literature, there is intensive research on providing precise semantics for BPEL and verification of BPEL models (Fostre et al., 2006), (Xu et al., 2006). Additionally, there exist frameworks and models (Zheng et al., 2007), (Sinha & Paradkar, 2006), (Yuan et al., 2006), (Yan et al., 2006) that enable the automatic generation and definition of BPEL-based test cases.

Currently, most of the derived BPEL test cases are generated to test whether the implementation of a WS conforms to the BPEL behaviour and WSDL interface models. In ICoM, the BPEL processes specified and adopted depict activity flows that produce the appropriate evaluated data based on which the conformance and the interoperability capabilities of a WSuT are assessed.

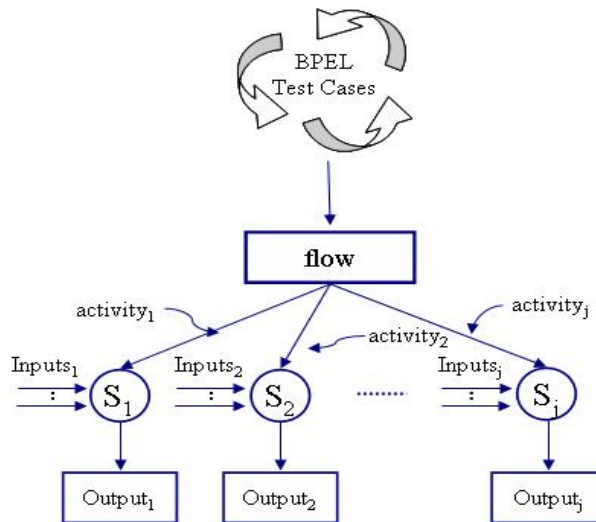


Fig. 11. BPEL Test Case

As depicted in Figure 11, a BPEL test case is a partially-ordered list of basic activities that that should be executed during a specific test run. The structural definition of a BPEL test case is as follows:

BPEL test case = {Activity_j, S_j, Input_j, Output_j}.

- Activity_j is a set of activities that should be performed.
- S_j is a set of primary WSuT services that are invoked and orchestrated for the execution of a test case. Each service is associated with the realization of a specific activity.
- Output_j is the result that each S_j derives. These results are gathered as the evaluated data of the SuTs.
- Input_j is set of parameters that each service S_j requires according to the provided functionality in order to complete a defined process. The inputs can be:

- test data that are fed by the TCI during the initiation of a BPEL process,
- outputs of other primary services,
- parameters that have been defined and are included in the BPEL process.

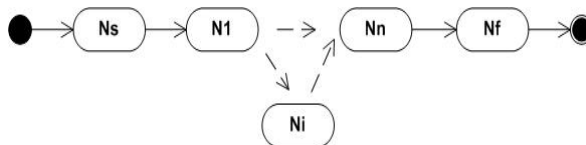


Fig. 12. Test Case Activity Graph Diagram

An activity graph diagram is also able to provide a visualization of the BPEL test cases (Figure 12). It illustrates and reflects the activity flow of the BPEL process in an effective manner. A test case activity diagram is composed by $\{N, E, N_s, N_f\}$ where N is a set of nodes $\{N_1, \dots, N_n\}$ that correspond to the applied activities, E is a set of edges $\{E_1, \dots, E_n\}$ that are the implicit sequence concatenations as defined by the performed workflow, N_s is the Start Node and N_f is the Final Node of the workflow.

In conformance testing, the designed (conformance) test cases include only internal activities. This means that merely the services of the examined SuT interact with each other in order to derive the evaluated data. Usually, in this type of test these data come from the final node of the test case.

The main objective of the current step focuses on the definition of the deployed BPEL processes based on the standards evaluated against. In this sense, the ICoM is quite adaptable since it is able to evaluate different aspects of the WSuT.

The test operator is able to create these processes utilizing existing BPEL test case generation frameworks or any other BPEL creation engine (ActiveBPEL, 2006). The processes are embedded in the SuT Service Orchestration Engine and the appropriate interfaces are implemented. During the BPEL generation phase, the required test data for each test case are specified and produced by the conformance evaluated tools of the TCI or by any other process.

Additionally, the test operator constructs a test case execution schedule taking into account the complexity of the applied processes. The schedule is a XML document that consists of the URLs of the BPEL processes interfaces and the corresponding required test data. This schedule along with the test data is stored in the Test Data Repository of the TCI. The steps that follow present and include the actual conformance testing sequence of the proposed methodology.

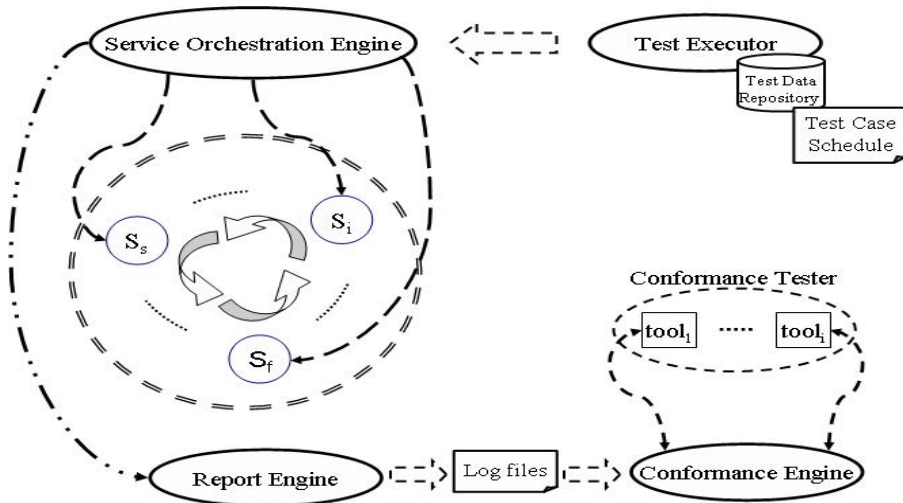


Fig. 13. Conformance Testing suite

- *Execution of test cases:* In this step, the TCI Test Executor automatically initiates the testing procedure by invoking the interface of the Service Orchestration Engine following the execution schedule. This action triggers the execution of the corresponding BPEL process running the defined test case. In Figure 13, the S_s is the initial service that is invoked and S_i is a set of services that contribute to the derivation of the evaluated data by the S_f .
- *Collection of results:* At the end of the test case, the Service Orchestration Engine has accumulated the evaluated data that are produced as outputs by the involved service of the WSuT. The Report Engine collects and consolidates these data in log files which are prepared for evaluation.
- *Results analysis.* Initially, the Conformance Engine analyzes the log files and extracts the produced data. Then, the Engine specifies the tools ($tool_1, \dots, tool_i$) of the Conformance Tester that should be used based on the nature of these data. The Engine feeds the data via the implemented interfaces to the corresponding tools, which in turn infer the conformance of the implementation to the respective standard. Suggestions concerning any corrections to the implementation are also provided to the SuT, enabling the deployment of a fine-tuning process that will correct discrepancies.
- *Corrective actions:* Corrective actions by the WSuTs' developers may include both re-design and re-implementation or only updates of specific services within the WSuT.
- *Re-execution of failed tests:* When the corrective actions are complete, the SuT undergoes a new test round to check if the previously failed tests are now successfully passed.

At the end of the last step, the process begins again from "Execution of test cases" step in subsequent iterations until all tests are successful.

In the example that follows, it is illustrated an instance of the execution of the current phase's steps in our demonstrated case study.

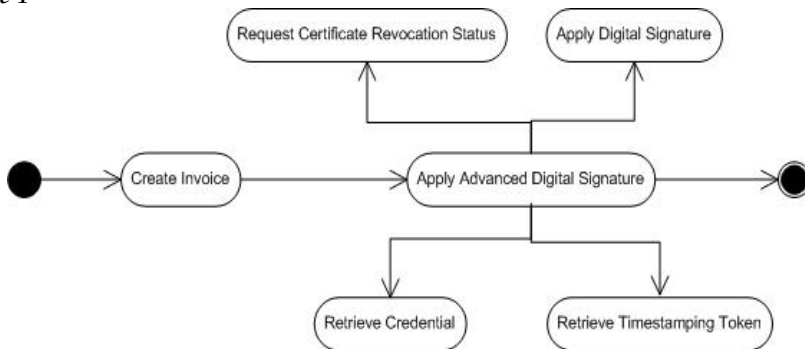
Example 4

Fig. 14. Activity Diagram of the XAdES conformance test case

Figure 14 depicts an activity diagram representing an example test case for the conformance testing of the SELIS XAdES implementation against the XAdES standard. The actions that compose the test case are the following:

- ✓ creation of a XML Invoice Document based on the xCBL standard,
- ✓ application of Advanced Digital Signature on the produced document. This process includes four separate sub-processes which occur transparently:
 - certificate retrieval of the certificate that will be used for the signature process,
 - digital signing of the invoice and certain other properties according to the W3C XML Digital Signature standard,
 - time stamping by requesting, and embedding a time stamp token on the generated signature according to the IETF 3161 standard, and finally
 - revocation information inclusion by embedding certificate revocation status data after requesting them on-line from the server of the Certification Authority that has issued the signer's certificate.

The above test case is to be used as follows: initially the TCI Test Executor initiates the execution of the corresponding BPEL process of the SELIS Service Orchestration Engine invoking the appropriate interface. The choreography of the required WSuT services derives an xCBL invoice document signed according to the XAdES standard as illustrated in Figure 15. The Report Engine embeds the signed document to a log file which is retrieved by the TCI's Conformance Engine. The Engine delegates the document to the integrated IAIK toolkit to validate the applied signature.

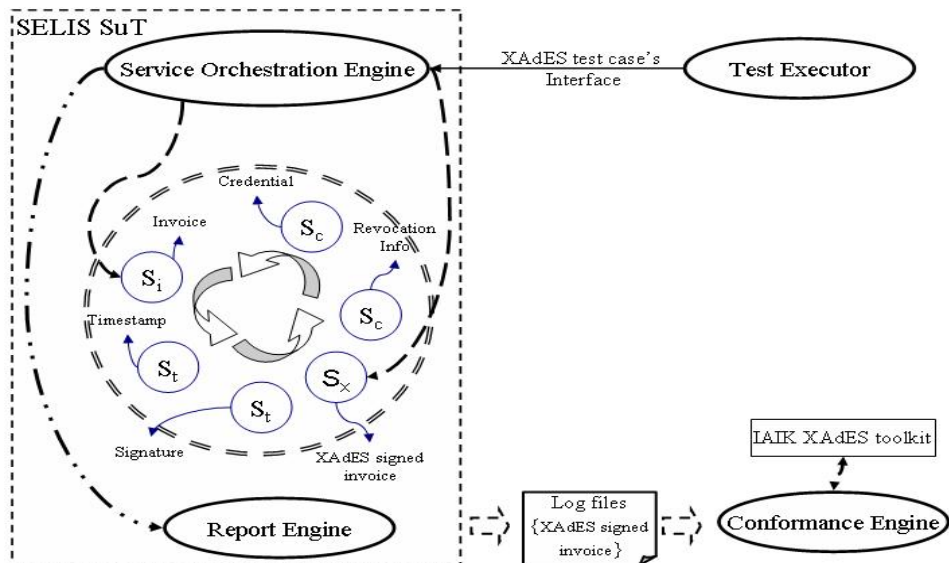


Fig. 15. SELIS Conformance XAdES Testing Suite

During our testing effort, in the actual first execution of this test case the toolkit indicated an incompatibility of the produced signature to the corresponding XML schema. This discrepancy was pointed out to the developers of the SELIS-invoicing which performed the necessary corrections. The re-execution of the test case denoted the conformance of SELIS with the XAdES standard.

3.4 Phase 4: Interoperability Testing

Interoperability testing is a more complex process than conformance testing because it usually involves at least two SuTs wishing to intercommunicate (Figure 16). The interoperability steps proposed by ICoM are not different from the corresponding conformance steps described in the previous section with regards to the logic and the sequence of the performed steps. On the contrary, the objective goals and the operation of these steps present significant differentiations. Thus, interoperability testing includes:

- *Definition of Conformance Test Cases:* The nature and the structure of the (interoperability) test cases that are designed and used in this phase are almost similar with the conformance ones, presented in Section 0. These test cases apart from internal activities include also and external activities. This means that the primary services of a SuT do not interact only with each other, but also with the services of other SuTs (Figure 16). Thus, the complexity of the BPEL processes that should be adopted to represent the logic of a test case is increased at a significant level.

The definition of the interoperability test cases takes into consideration two parameters. The first one is the testing scenario and the role assignment to the SuTs which was defined in the phase I of ICoM. The second parameter is the standards that the SuTs have adopted and are evaluated against. The specified interoperability test cases should

constitute instance of this scenario taking into account the variants of this scenario based on the adopted standards.

BPEL processes should be formulated merely for the SuTs that initiate the execution of a testing suite. For example, the specified scenario, as depicted in Figure 16, is that the “SuT A” interacts with the “SuT B” retrieving a response. Therefore, based on a defined test case, the “WSuT A” orchestrates its corresponding services e.g. As and Ak to execute a specific process that enables service Ai to communicate with the service Bj of the “WSuT B” expecting a reply which is handled it appropriately. The BPEL process that corresponds to the above test case includes merely the actions which are executed by the “WSuT A”. “WSuT B” reacts to this interaction creating the reply according to the logic that is included in the implementation.

Identically to the conformance testing, the BPEL processes are created via BPEL generation framework or any other available BPEL engine and are embedded in the Service Orchestration Engine of the respective SuT. The execution of the test cases is defined by a testing schedule that is prepared.

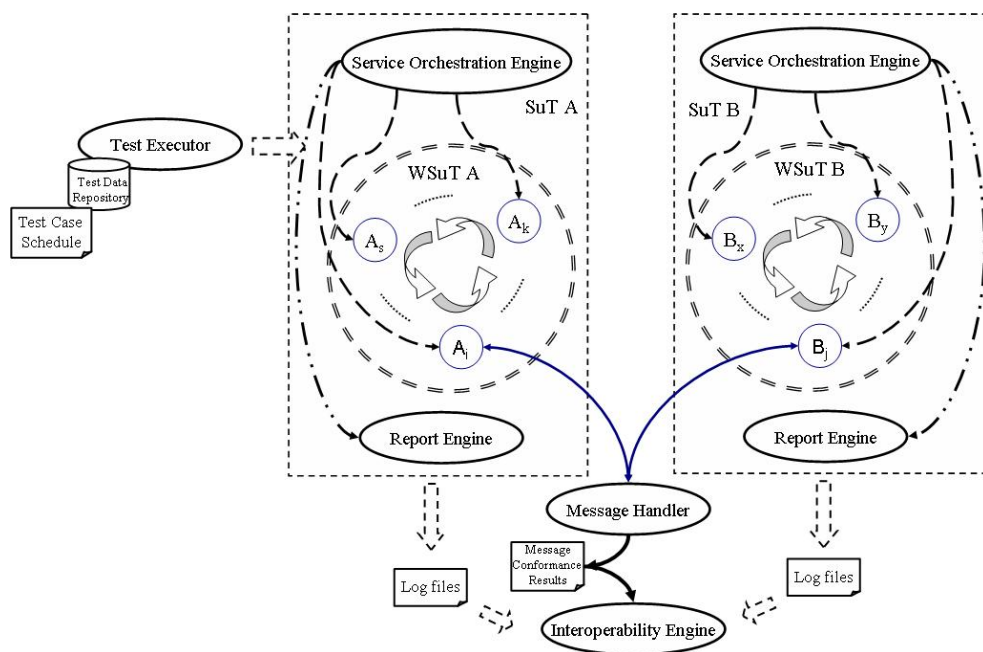


Fig. 16. Interoperability Testing Suite

- *Execution of test cases:* As presented in Figure 16, this step is completed following a similar process with the corresponding of the conformance testing. The execution of the external interactions among the involved WSuTs introduces an additional factor that should be taken into account and concerns the conformance of the exchanged messages to the defined business process. Thus, the TCI's Message Handler participates in the external interactions intercepting and analyzing the exchanged messages, and checking that they belong to the proper message domain.

- *Collection of results:* The Service Orchestration Engines of the SuTs taken part in the test case accumulate the whole of the evaluated data that are produced by all the involved WSuTs' services. Then the corresponding Report Engines collect and consolidate them in log files.
- *Results analysis.* The test results are analyzed and conclusions are made on the interoperability capabilities of the SuTs, as presented in Section 0.
- *Corrective actions:* Corrective actions may include both re-design and re-implementation or only updates of specific areas within the SuT.
- *Re-execution of failed tests:* As soon as the corrective actions have finished, the SuTs undergo a new test round to check if the previously failed tests are now successfully passed.

At the end of last step, the process begins again from second one in subsequent iterations until all tests are successful. In the Example 5 that follows the operation of the interoperability testing is described in detail.

Example 5

Figure 17 represents the activity diagram of an interoperability test case that is an instance of our working demonstration scenario, as implemented by the SELIS and SWEB SuTs. This test case includes only the actions performed by SELIS which is the actual initiator of the testing procedure.

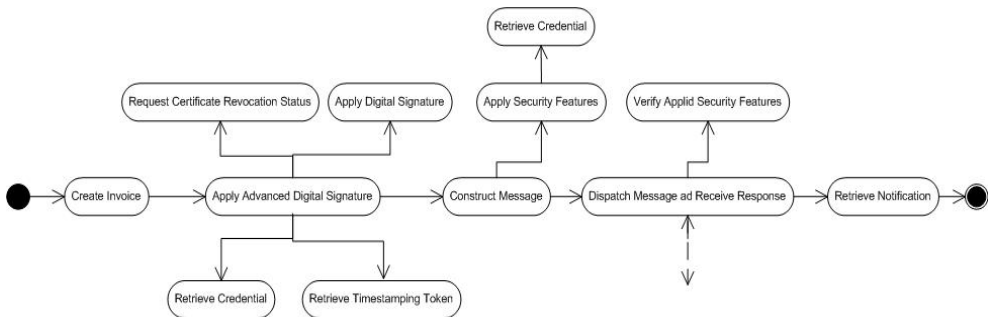


Fig. 17. Activity diagram of an Interoperability Test Case

As depicted in Figure 18 the process begins from the service A_1 of SELIS WSuT that creates a xCBL invoicing document (I_1). The document is signed by service A_2 according to the XAdES standard (SI_1) gathering the required time stamps and related certificate revocation status information data from their respective sources. Service A_4 packages the signed invoice in a SOAP message (M_1) where the WS Security features are applied (SM_1) using the service A_3 . Finally, the service A_5 dispatches the message to the to the SWEB.

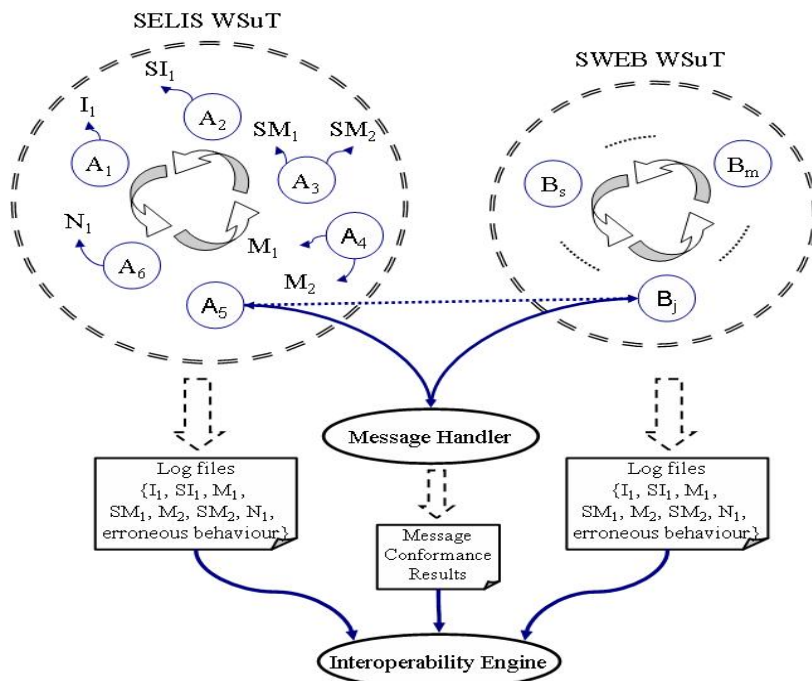


Fig. 18. SELIS and SWEB-invoicing Interoperability Testing Suite

SWEB handles the received SOAP message according to its predefined internal processes without following any steps specified by the test operator. Thus, SWEB receives the SOAP message, decrypts it and verifies the WS Security features. Then, the invoice document is extracted and the XAdES signature is verified (along with the rest of the cryptographic information it encompasses like any time stamps). A notification (N_1) is created that is packaged in a new SOAP message (M_2) where the WS-Security extensions are applied (M_2) and is sent to SELIS.

The latter as soon as receives the response, handles it according to the test case actions. Service A_5 retrieves the SOAP message and using the service A_3 validates the applied security features. Finally, the service A_6 manages the notification.

It should be noted that, the exchanged SOAP messages (SM_1 and SM_2) are intercepted by the TCI Message Handler. The Handler analyzes them checking that these messages conform to the defined SuTs' WSDL documents and produces the appropriate conformance results. The SuT Report Engines of both SuTs collect all the evaluation results and import them in log files. These consist of the exchanged SOAP messages (M_1 , SM_1 , M_2 , and SM_2), the exchanged documents (I_1 , SI_1 , N_1) and any erroneous behaviour that was reported during the whole process e.g. the application and validation of the applied security features of the invoice document.

The TCI Interoperability Engine receives the SuTs' log files and the Message Handler conformance results that are provided to the test operator which is responsible to analyze

them and indicate the SuTs communication capability. The analysis we have performed in our demonstration systems verified that:

- ✓ the SuTs possess the same documents (I_1 , SI_1 , N_1) and messages (M_1 , SM_1 , M_2 , SM_2) that were created either from the one or from the other system,
- ✓ no erroneous behaviour has been detected during the interaction e.g. the validation of all message and document security features was successful,
- ✓ the exchanged messages (SM_1 , SM_2) are part of the SuTs' business processes.

Therefore, according to the abovementioned analysis the two systems are interoperable.

4. Conclusions and Future Work

The creation of an interoperability framework is based on two factors, the precise definition of the used standards and the specification of a strict set of guidelines that should be followed. The testing methodologies are able to guarantee the development of this framework. Nevertheless, the existing testing methodologies and frameworks are not able to cover all interoperability aspects of a Web Services-based environment.

This paper presents an enhanced well-formed testing methodology named ICoM, which is able to test and guarantee the end-to-end application interaction capabilities of WS-based services offered autonomously or via a SOA. ICoM achieves to overcome the weaknesses of the existing methodologies being able to satisfy various requirements. ICoM is based on the principles of existing interoperability methodologies and frameworks adopting widely used testing types, covering a set of interoperability and conformance aspects of the WSs and finally formulating an adaptable, extensible and flexible framework. ICoM was demonstrated using an e-invoicing service which is offered as an autonomous WS (SELIS) and via the SOA-based platform SWEB.

Future work in this area includes the integration of new types of testing such as integration, performance and stress testing in the proposed methodology. These new testing types will enrich ICoM allowing to test various dimensions of the evaluated services and enabling the creation of a more integrated and enhanced testing framework.

5. Acknowledgment

This work has been supported by the GSRT (PENED) programme and the IST project SWEB (IST-2006-2.6.5). The authors would like to thank all the participants for valuable discussions and the European Union for funding the SWEB project.

6. References

- High, R.; Kinder, S. & Graham, S. (2005). *IBM's SOA Foundation – An architectural Introduction and Overview*.
- European Interoperability Framework for Pan-European eGovernment Services (EIF), available at <http://europa.eu.int/idabc/>
- Kaliontzoglou, A.; Boutsi, P. & Polemi, D. (2006). eInvoke: Secure e-Invoicing based on Web Services. *Electronic Commerce Research*, vol 6, Numbers 3-4, pp. 337-353, Springer.

- Papastergiou, S.; Karantjias, A. & Polemi, D. (2007a). Innovative, Secure and Interoperable E/M-Governmental Invoicing. *18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Athens.
- Secure, interoperable, cross border m-services contributing towards a trustful European cooperation with the non-EU member Western Balkan countries (SWEB), IST-2006-2.6.5, 6th Framework Programme, Priority 2, Information Society Technologies (2007), www.sweb-project.org/.
- OSI - Open System Interconnection, Conformance testing methodology and framework (1997), ISO/IEC 9646.
- European Telecommunications Standards Institute (ETSI), available at <http://portal.etsi.org/mbs/Testing/testing.htm>
- Moseley, S.; Randall, S. & Wiles, A. (2003). Experience within ETSI of the combined roles of conformance testing and interoperability testing. *Standardization and Innovation in Information Technology*, pp 177- 189, IEEE Press.
- ETSI 300 406 (1995). Methods for testing and Specification (MTS); Protocol and profile conformance testing specifications; Standardization methodology.
- ETSI ES 201 873 (2002): The Testing and Test Control Notation version 3, v2.2.0.
- Dibuz, S. & Kremer, P. (2003). Interoperability Testing - Framework and Model for Automated Interoperability Test and Its Application to ROHC. *TestCom 2003*, LNCS vol 2644, pp. 243-257, Computer Science.
- ETSI EG 202 237 V1.1.2 (2007-04): Methods for Testing and Specification (MTS); Internet Protocol Testing (IPT); Generic approach to interoperability testing.
- Kulvatunyou, B.; Ivezic, N.; Martin, M. & Jones, A.T. (2003). A Business-to-Business Interoperability Testbed: An Overview. *The Fifth International Conference on Electronic Commerce*, Pittsburgh, PA October.
- Seely, S. & Lauzon, D. (2005). *WS-I monitor tool functional specification*, ver. 1.1. Technical report, WS-I.
- Ehnebuske, D. et al. (2003). *WS-I Overview*. Available at www.ws-i.org/docs/20021003.wsi.introduction.pdf
- ebXML Technical Architecture Project Team, (2001). ebXML Technical Architecture Specification v1.0.4, February 16.
- OASIS ebXML Implementation, Interoperability and Conformance Technical Committee:, (2003). ebXML Test Framework v0.3, 07 March.
- Lee, Y. (2005). ebXML Test Framework on Dually Coupled Asynchronous MSH. *Fourth Annual ACIS International Conference on Computer and Information Science (ICIS'05)*, pp. 198-203, IEEE Press.
- Kim, D. & Yun, J.H. (2003): Development -of an ebXML Conformance Test System for e-Business Solutions. *EC-Web 2003*, LNCS, vol 2738, pp. 145-154, Springer-Verlag Berlin Heidelberg
- IBM XML security suite, available at <http://www.alphaworks.ibm.com/tech/xmlsecuritysuite>.
- IAIK XML signature library, available at http://jce.iaik.tugraz.at/sic/products/xml_security.
- IAIK XAdES library, http://jce.iaik.tugraz.at/sic/products/xml_security/xades

- Papastergiou, S.; Kaliontzoglou, A. & Polemi, D. (2007b) Interoperability issues of a Secure ELectronic Invoicing Service (SELIS). *18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2007)*, 3-7 September, Athens.
- ETSI Technical Report 266 (1996): Methods for testing and Specification (MTS); Test Purpose style guide.
- Saglietti, F.; Oster, N. & Pinte, F. (2008). White and grey-box verification and validation approaches for safety- and security-critical software systems. *Information Security Technical Report*, Vol. 13, *Elsevier Advanced Technology*, ISSN 1363-4127, p. 10 – 16.
- Peyton L.; Stepien, B. & Seguin, P. (2008). Integration Testing of Composite Applications, *Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, IEEE Computer Society.
- Rizwan M.; Mamoon Y., (2007). SOA Testing using Black, White and Gray Box Techniques, White paper Crosscheck Networks, available at <http://www.crosschecknet.com/resources/articles/soatesting-v2.htm>.
- Brittenham, P. (2003). Understanding the WS-I Test Tool, available at <http://www-128.ibm.com/developerworks/webservices/library/ws-wsitest/>.
- Brittenham, P.; Durand, J.; Kleijkers, L. & Stobie, K. (2005a). WS-I Monitor Tool Functional Specification, WS-I Testing Work Group.
- Brittenham, P., et al. (2005b). WS-I Analyzer Tool Functional Specification WS-I Testing Work Group.
- Pentafronimos G.; Papastergiou S. & Polemi D. (2008). Definition and representation of test cases for e-government Web Services, *2nd International Conference on Theory and Practice of Electronic Governance (ICEGOV2008)*, 1 - 4 December, 2008, Cairo, Egypt.
- Fostre, H.; Uchitel, S.; Magee, J. & Kramer, J. (2006). Model based analysis of obligations in web service choreography, *Proc. of AICT-ICIW*. IEEE Computer Society.
- Xu, K.; Liu, Y. & Pu, G., (2006). Formalization, verification and restructuring of bpel models with pi calculus and model checking, IBM, Tech. Rep.
- Zheng, Y.; Zhou, J. & Krause, P., (2007). An Automatic Test Case Generation Framework for Web Services. *Journal of Software*, vol. 2(3), pp. 64-77.
- Sinha, A. & Paradkar, A. (2006). "Model based functional conformance testing of web services operating on persistent data," in *Proc. of TAV-WEB*. ACM Press, pp. 17-22.
- Yuan, Y.; Li, Z. & Sun, W., (2006). A graph-search based approach to bpel4ws test generation, *Proc. of ICSEA*. IEEE Computer Society, p. 14.
- Yan, J.; Li, Z.; Yuan, Y.; Sun, W & Zhang, J. (2006). Bpel4ws unit testing: Test case generation using a concurrent path analysis approach, *Proc. of ISSRE*. IEEE Computer Society, 2006, pp. 75-84.
- ActiveBPEL Engine Architecture (2006), available at <http://www.activebpel.org/docs/architecture.html>.
- Karantjias A.; Papastergiou S. & Polemi D. (2008). Holistic Electronic & Mobile Government Platform Architecture. *8th Joint Conference on Knowledge - Based Software Engineering 2008 (JCKBSE 08)*, IOS Press, August 25-28, Athens, Greece.
- Papastergiou, S.; Polemi, D. & Douligieris, C., (2009). SWEB: An advanced mobile Residence Certificate Service, *3rd International Conference on e-Democracy "Next Generation Society: Technological and Legal Issues"*, Lecture Notes of ICST (LNICST), Springer, ISBN 978-963-9799-54-7, Athens, Greece.