
WIRELESS COMMUNICATIONS AND NETWORKS – RECENT ADVANCES

Edited by **Ali Eksim**

INTECHWEB.ORG

Wireless Communications and Networks – Recent Advances

Edited by Ali Eksim

Published by InTech

Janeza Trdine 9, 51000 Rijeka, Croatia

Copyright © 2012 InTech

All chapters are Open Access distributed under the Creative Commons Attribution 3.0 license, which allows users to download, copy and build upon published articles even for commercial purposes, as long as the author and publisher are properly credited, which ensures maximum dissemination and a wider impact of our publications. After this work has been published by InTech, authors have the right to republish it, in whole or part, in any publication of which they are the author, and to make other personal use of the work. Any republication, referencing or personal use of the work must explicitly identify the original source.

As for readers, this license allows users to download, copy and build upon published chapters even for commercial purposes, as long as the author and publisher are properly credited, which ensures maximum dissemination and a wider impact of our publications.

Notice

Statements and opinions expressed in the chapters are those 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 chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

Publishing Process Manager Ivana Zec

Technical Editor Teodora Smiljanic

Cover Designer InTech Design Team

First published March, 2012

Printed in Croatia

A free online edition of this book is available at www.intechopen.com

Additional hard copies can be obtained from orders@intechweb.org

Wireless Communications and Networks – Recent Advances, Edited by Ali Eksim

p. cm.

ISBN 978-953-51-0189-5

INTECH

open science | open minds

free online editions of InTech
Books and Journals can be found at
www.intechopen.com

Contents

Preface IX

Part 1 Wireless Communication Antennas 1

- Chapter 1 **Latest Progress in MIMO Antennas Design 3**
Yue Li, Jianfeng Zheng and Zhenghe Feng
- Chapter 2 **Review of the Wireless Capsule
Transmitting and Receiving Antennas 27**
Zhao Wang, Eng Gee Lim, Tammam Tillo
and Fangzhou Yu
- Chapter 3 **Travelling Planar Wave Antenna
for Wireless Communications 47**
Onofrio Losito and Vincenzo Dimicoli
- Chapter 4 **Superstrate Antennas for Wide
Bandwidth and High Efficiency
for 60 GHz Indoor Communications 93**
Hamsakutty Vettikalladi, Olivier Lafond
and Mohamed Himdi
- ### **Part 2 Wireless Communication Hardware 123**
- Chapter 5 **Hardware Implementation of Wireless
Communications Algorithms: A Practical Approach 125**
Antonio F. Mondragon-Torres
- Chapter 6 **Gallium Nitride-Based Power Amplifiers for Future
Wireless Communication Infrastructure 157**
Suramate Chalermwisutkul
- Chapter 7 **Analysis of Platform Noise Effect on Performance
of Wireless Communication Devices 177**
Han-Nien Lin

Part 3 Channel Estimation and Capacity 227

Chapter 8 **Indoor Channel Measurement
for Wireless Communication 229**
Hui Yu and Xi Chen

Chapter 9 **Superimposed Training-Aided Channel
Estimation for Multiple Input Multiple
Output-Orthogonal Frequency Division Multiplexing
Systems over High-Mobility Environment 255**
Han Zhang, Xianhua Dai, Daru Pan and Shan Gao

Chapter 10 **Channel Capacity Analysis Under Various Adaptation
Policies and Diversity Techniques over Fading Channels 281**
Mihajlo Stefanović, Jelena Anastasov, Stefan Panić, Petar Spalević
and Ćemal Dolićanin

**Part 4 Wireless Communication Performance
Analysis Tools and Methods 303**

Chapter 11 **Generalized Approach to Signal Processing
in Wireless Communications:
The Main Aspects and some Examples 305**
Vyacheslav Tuzlukov

Chapter 12 **Engineering of Communication Systems and Protocols 339**
Pero Latkoski and Borislav Popovski

Chapter 13 **Cell Dwell Time and Channel Holding Time
Relationship in Mobile Cellular Networks 357**
Anum L. Enlil Corral-Ruiz, Felipe A. Cruz-Pérez
and Genaro Hernández-Valdez

**Part 5 Next Generation Wireless
Communication Technologies 379**

Chapter 14 **Automatic Modulation Classification
for Adaptive Wireless OFDM Systems 381**
Lars Häring

Chapter 15 **User Oriented Quality of Service
Framework for WiMAX 403**
Niharika Kumar, Siddu P. Algur and Amitkeerti M. Lagare

Chapter 16 **Introduction to the Retransmission Scheme Under
Cooperative Diversity in Wireless Networks 429**
Yao-Liang Chung and Zsehong Tsai

- Chapter 17 **Intelligent Transport Systems:
Co-Operative Systems (Vehicular Communications) 447**
Panagiotis Lytrivis and Angelos Amditis
- Chapter 18 **Wireless Technologies in the Railway:
Train-to-Earth Wireless Communications 469**
Itziar Salaberria, Roberto Carballado and Asier Perallos
- Chapter 19 **Super-Broadband Wireless Access Network 493**
Seyed Reza Abdollahi, H.S. Al-Raweshidy and T.J. Owens
- Part 6 Biological Effects of Wireless
Communication Technologies 521**
- Chapter 20 **Evaluations of International Expert Group Reports
on the Biological Effects of Radiofrequency Fields 523**
Verschaeve Luc
- Part 7 Wireless Sensor Networks and MANETS 547**
- Chapter 21 **Power Management in Sensing Subsystem
of Wireless Multimedia Sensor Networks 549**
Mohammad Alaei and Jose Maria Barcelo-Ordinas
- Chapter 22 **Multimedia Applications for MANETs over
Homogeneous and Heterogeneous Mobile Devices 571**
Saleh Ali Alomari and Putra Sumari

Preface

Wireless communications and networks have been one of the major revolutions of the last three decades. We are witnessing a very fast growth in these technologies where wireless communications and networks have become so ubiquitous in our society and indispensable in our daily lives. The demand for new services to support high speed wideband Internet access and advanced high quality real-time video applications push the researchers to investigate new technologies in wireless communications and networks.

Progress in wireless communications and networks continues as this book is being written. Although there have been many journal and conference publications regarding wireless communication, they are often in the context of academic research or theoretical derivations and sometimes omit practical considerations. Although the literature has many conference and journal papers, technical reports, and standard contributions, they are often fragmental engineering works and thus are not easy to follow up. The objective of this book is to accelerate research and development by serving as a forum in which both academia and industry can share experiences and report original studies and works regarding all aspects of wireless communications. In addition, this book has great educational value because it aims to serve as a virtual, but nonetheless effective bridge between academic research in theory and engineering development in practice, and as a messenger between the technical pioneers and the researchers who followed in their footsteps.

This book which is titled "Wireless Communications and Networks - Recent Advances", focuses on the current research topics from a wide range of wireless communications and networks and provides "on-going" research progress on these issues. During the preparation of this book, I emphasized to the authors to add recent research findings and future works in this area and to cite latest references in the chapter. For this reason, a variety of novel techniques in wireless communications and networks are investigated in this book. The authors attempt to present these topics in detail. Insightful and reader-friendly descriptions are presented to nourish readers of any level, from practicing and knowledgeable communication engineers to beginning or professional researchers. All interested readers can easily find noteworthy materials

in much greater detail than in previous publications and in the references cited in these chapters.

This book includes twenty two chapters that were authored by the well-known researchers in the world. Each chapter was written in an introductory style beginning with the fundamentals, describing approaches to the hottest issues and concluding with a comprehensive discussion. The content in each chapter is taken from many publications in prestigious journals and conferences and followed by fruitful insights. The chapters in this book also provide many recent references for relevant topics, and interested readers will find these references helpful when they explore these topics further.

This book was divided into seven parts. Part 1 consists of four chapters which are dedicated to wireless communication antennas. Part 2 consists of three chapters which are dedicated to wireless communication hardware. Part 3 consists of three chapters which are dedicated to channel estimation and capacity. Part 4 consists of three chapters which are dedicated to wireless communication performance analysis tools and methods. Part 5 consists of six chapters which are dedicated to next generation wireless communication technologies. Part 6 consists of only one chapter which is dedicated to biological effects of wireless communication technologies. Finally, Part 7 consists of two chapters which are dedicated to wireless sensor networks & Mobile Ad Hoc Networks (MANETs).

Chapter 1 provides a comprehensive discussion on the latest technologies of antenna design for space-limited Multi-Input Multi-Output (MIMO) applications, such as minimized base station, portable access point and mobile terminals. solve the contradiction of system volume and antenna performance, two basic methods are proposed in this chapter to maintain the channel capacity in a reduced system volume. The first method is to reduce the volume each antenna occupied without decreasing the number of antenna elements. Another is to antenna performance in space-limited MIMO system, without increasing the antenna volume.

Chapter 2 introduces Wireless Capsule Endoscopy (WCE) system and antenna specifications. Special consideration of body characteristics for antenna design and state-of-the-art WCE transmitting and receiving antennas are also reviewed in this chapter.

Chapter 3 explains travelling planar wave antenna for wireless communications. This chapter describes the types of travelling planar wave antennas that are Wave Antenna (LWA), Meanderline antenna, taped LWA and taped composite right/left-handed transmission-line LWA. In this chapter, measurements are verified with simulations for all types of LWA.

Chapter 4 explains how to develop a wideband, high gain and high efficient antenna sufficient for 60 GHz communications using superstrate technology. This chapter also

explains the importance of different sources on antenna performance in terms of bandwidth, gain and efficiency.

Chapter 5 explains hardware implementation of wireless communications algorithms with a practical approach. This chapter navigates through the author's encounters with different technologies at different stages in his career and how different applications have been and are currently approached. This chapter also gives a summary of the author's last ten years of working with different tools, methodologies and design flows.

Chapter 6 reviews state-of-the-art research in power amplifiers for wireless communication infrastructure featuring advantages of Gallium Nitride (GaN)-based power devices including large bandwidth capability, high power density and high output impedance. Regarding the issues of power amplifier design, state-of-the-art power amplifier architectures discusses with various prospects. This chapter also discusses widespread techniques for average efficiency enhancement including Doherty power amplifier concept and envelope tracking with state-of-the-art results with examples.

Chapter 7 discusses radio frequency (RF) desensitivity analysis for components and devices on mobile products. To improve the total isotropic sensitivity performance of wireless communication on notebook computer, this chapter investigate the electromagnetic interference noise from the built-in camera display module as examples and analyzed the impact of various modes on performance with throughput measurement. This chapter discovers throughput and receiving sensitivity of wireless communications and the solutions to improve system performance. Moreover, this chapter describes how to design and implement periodic structures for isolation on the notebook computer to effectively suppress noise source-antenna coupling and improve the receiving sensitivity of wireless communication system.

Chapter 8 explains indoor channel measurement for wireless communications. This chapter firstly gives detailed information about indoor channel measurement for MIMO-Orthogonal Frequency Division Multiplexing (MIMO-OFDM) systems. Secondly, channel measurement schemes are explained. Finally, channel measurement applications are given in this chapter.

Chapter 9 addresses the problem of estimating the linearly time-varying (LTV) channel of MIMO-OFDM systems using superimposed training (ST). The LTV channel is modeled by truncated discrete Fourier bases. Based on this model, a two-step approach is adopted to estimate the LTV channel over multiple OFDM symbols. This chapter also presents a performance analysis of the channel estimation and derives a closed-form expression for the channel estimation variances. It is shown that the estimation variances, unlike that of the conventional ST-based schemes, approach to a fixed lower-bound as the training length increases, which is directly proportional to

information-pilot power ratios. For wireless communications with a limited transmission power, the authors` try to optimize the ST power allocation by maximizing the lower bound of the average channel capacity. Simulation results show that the proposed approach in this chapter outperforms the frequency-division multiplexed trainings schemes.

Chapter 10 focuses on more general and nonlinear fading distributions. An analytical study of the β - γ fading channel capacity, e.g., under the optimal power and rate adaptation (OPRA), constant power with optimal rate adaptation (ORA), channel inversion with fixed rate (CIFR), and truncated CIFR (TIFR) adaptation policies and maximum ratio combining (MRC) and selection combining diversity techniques are performed. The expressions for the proposed adaptation policies and diversity techniques are derived in this chapter. Capitalizing on them, numerically obtained results are graphically presented, in order to show the effects of various system parameters, such as diversity order and fading severity on observed performances. In a similar manner an analytical study of the Weibull fading channel capacity, under the OPRA, ORA, CIFR and TIFR adaptation policies and MRC diversity technique are performed in this chapter.

Chapter 11 explains generalized approach to signal processing (GASP) in wireless communications with examples. The used technique in this chapter, GASP, allows researchers to extend the well-known boundaries of the potential noise immunity set by classical and modern signal processing theories. Employment of wireless communication systems, the receivers of which are constructed on the basis of GASP, allows the researchers to obtain high detection of signals and high accuracy of signal parameter definition with noise components present compared with that systems, the receivers of which are constructed on the basis of classical and modern signal processing theories.

Chapter 12 emphasizes the importance of conducting an early performance evaluation of the communication protocols and systems, and to suggest an appropriate solution for carrying out such an activity. Performance evaluation activity denotes the actions to evaluate the protocol under development regarding its performance. This process can take place in different phases of the development, and can be based on modelling or measurements. If the designer can control the performance of the product, rather than just manage its functionality, the result will be a much superior creation. This problem is treated in this chapter through a tangible wireless communication protocol example.

Chapter 13 discusses statistical relationships among residual cell dwell time (CDTr), cell dwell time (CDT), and channel holding time (CHT) for new and handoff calls. In particular, under the assumption that unencumbered service time is exponentially distributed and CDT is phase-type distributed, a novel algebraic set of general equations that examine the relationships both between CDT and CDTr and between CDT and channel holding times are obtained. Also, the condition upon which the

mean channel holding time for new calls (CHT_n) is greater than the mean channel holding time for handoff calls (CHTh) is derived in this chapter. Additionally, novel mathematical expressions for determining the parameters of the resulting CHT distribution as functions of the parameters of the CDT distribution are derived in this chapter for hyper-exponentially or Coxian distributed CDT.

Chapter 14 highlights the classification of digital quadrature amplitude modulation schemes in wireless adaptive OFDM systems using the likelihood principle. The author particularly focuses on time-division duplex systems in which the channel can be regarded as reciprocal. In contrast to other research work, a lot of new constraints are taken into account. Namely, many parameters are known by the receiver that can be utilized to enhance the classification reliability.

Chapter 15 introduces a user based framework in Worldwide Interoperability for Microwave Access (WiMAX) and explores user based bandwidth allocation algorithms, user based packet classification mechanism and user based call admission control algorithm.

Chapter 16 covers the conceptual description of many representative retransmission schemes under various environments and presented a novel fast packet retransmission scheme intended for effectively transporting delay-sensitive flows in a general cooperative diversity environment.

Chapter 17 highlights the significant role of cooperative (vehicular) communications in future Intelligent Transport Systems. This chapter describes the architecture of cooperative systems, wireless technologies used within the cooperative systems framework and the applications of vehicular networks and their corresponding categories. This chapter also emphasizes on hot research topics concerning cooperative systems such as data fusion, routing, security and privacy.

Chapter 18 describes a specific wireless communications architecture developed taking into account railway communications needs and the restrictions that have to be considered in terms of broadband network features. It is based on standard communication technologies and protocols to establish a bidirectional communication channel between trains and railway control centers. In this chapter, a brief description of the state of art in railway communications, a specific train-to-earth wireless communication architecture, the main challenges concerning with the management of the quality of service in train-to-earth communications, some services that are arising as result of using this connectivity architecture and the way in which they interoperate the future lines of work oriented to improve the proposed communication channel are also explained.

Chapter 19 explains super-broadband wireless access networks. This chapter firstly discusses the evolution of Internet traffic growth in subscribing the internet and wireless network worldwide in diverse domain of services. This chapter secondly

presents the solutions for transportation with huge traffic demand, according with the expected growth in interactive video, voice communication and data traffic for providing the cost effective communication services. Finally, it describes the radio over fiber network as a future proof solution for supporting super-broadband services that is a reliable, cost-effective and environmentally friendly technology.

Chapter 20 gives evaluations of more than 30 international expert group reports on the biological effect of wireless communication systems. Evaluated reports in this chapter were published during the 2009-2011 period. The vast majority did not consider that there is a demonstrated health risk of RF exposure from mobile phones and other wireless communication devices.

Chapter 21 describes a mechanism for the management of the wireless multimedia sensor nodes. The mechanism, first, clusters nodes according to their scale of similarity in covering the environment; second, selects and schedules members of established clusters to monitor the sensing region which is divided among clusters. The members of each cluster are scheduled with an exclusive frequency based on the number of members in the cluster and the scale of overlapping among fields of view of the cluster members and thus the monitoring efficiency is increased. Moreover, because of the established intra cluster coordination and collaboration, sensing subsystem of multimedia nodes are optimized to avoid redundant and overlapped sensing. Thus, the capability of energy saving is considerably enhanced with respect to ordinary duty-cycling manners of environment monitoring by wireless multimedia sensor networks. On the other hand, optimizing the data sensed by sensing subsystem results in conservation of energy in the transmission and processing subsystems since they meet less amounts of multimedia data to be transmitted and/or processed by the network nodes. Results in this chapter show how this mechanism prolongs the network lifetime along with a better monitoring performance.

Chapter 22 explains wireless communications for over homogeneous and heterogeneous mobile devices. This chapter introduces related background and main concepts of the MANETs, existing wireless mobile network approaches, wireless ad hoc networks, wireless mobile approaches, characteristics of MANETs and types of MANETs. Second, the traffic types in ad hoc networks, ad hoc network routing protocol performance issues and the types of ad hoc protocols are given in this chapter. Third, comparison between proactive versus reactive and clustering versus hierarchical protocols are explained. Finally, mobility, Quality of Service provisioning, multicasting and security issues of MANETs are presented.

Briefly, this book will provide a comprehensive technical guide covering fundamentals, recent advances and open issues in wireless communications and networks to the readers. objective of the book is to serve as a valuable reference for

students, educators, scientists, faculty members, researchers, engineers and research strategists in these rapidly evolving fields and to encourage them to actively explore these broad, exciting and rapidly-evolving research areas.

Dr. Ali Ekşim

Chief Senior Researcher
Center of Research for Advanced Technologies of
Informatics and Information Security
(Tubitak-Bilgem)
Turkey

Part 1

Wireless Communication Antennas

Latest Progress in MIMO Antennas Design

Yue Li, Jianfeng Zheng and Zhenghe Feng
Tsinghua University
China

1. Introduction

Multiple-Input Multiple-Output (MIMO) wireless communication system, which is also called Multiple-Antenna system, is well known as one of the most important technologies and widely studied nowadays (Winters, 1987; Foschini & Gans, 1998; Marzetta & Hochwald, 1999; Raleigh & Cioffi, 1998). The main idea of MIMO wireless communication is to utilize the spatial degree of freedom of the wireless multi-path channel by adopting multiple antennas at both transmit and receive ends to improve spectrum efficiency and transmission quality of the wireless communication systems. MIMO technology is able to extremely improve the transmission data rates and alleviate the conflict between the increasing demand of wireless services and the scarce of electromagnetic spectrum. Two famous techniques of the MIMO systems are spatial multiplexing (SM) and transmit diversity (TD) (Nabar et al, 2002). In the scheme of SM, multiple data pipes between transmit and receive ends provide multiplexing gain to dramatically increase the channel capacity linearly with the number of antennas (Telatar, 1999; Bolcskei et al, 2002). The TD technologies, such as space-time coding, are adopted to improve the link reliability of wireless communication, especially in the multi-path fading channels (Marzetta & Hochwald, 1999; Tarokh et al, 1998; Bolcskei et al, 2001). The channel knowledge is not required in the transmit end for TD technologies. MIMO is the key technology for future wireless communication systems, such as 3GPP LTE, WiMAX 802.16, IEEE 802.20, IMT-Advanced and so on.

Although the spatial degree of freedom is important and has the potential to extremely increase the capacity of the MIMO systems, how to utilize the space resources is still needed to be studied. Physical layer design is the most important issue of wireless communication systems. Among all the components, the antenna is the interface of the MIMO wireless communication systems to the channel, which is the most sensitive part for the spatial degree of freedom. The system performance is directly dictated by the number of antennas adopted in transmit and receive end. The key issue to achieve high channel capacity of the MIMO system is the mutual coupling between antenna elements. In traditional MIMO systems, space-separated antenna array is adopted at the base station or mobile terminal. Nearly half of the wavelength is required to achieve acceptable isolation, about -15 dB for most of the situations. However, for the space is limited in both the base station and the mobile terminal, the mutual coupling between the adjacent antenna elements becomes more and more serious, restricts the performance of MIMO systems (Wallace & Jensen, 2004; Morris & Jensen, 2005). The design of antenna in space-limited MIMO system is still need further discussed. This chapter will focus on this topic.

In this chapter, we provide a comprehensive discussion on the latest technologies of antenna design for space-limited MIMO applications, such as minimized base stations, portable access points and mobile terminals. To solve the contradiction of system volume and antenna performance, two basic methods are proposed to maintain the channel capacity in a reduced system volume, as illustrated in Fig. 1. The first one is to reduce to volume each antenna occupied without decreasing the number of antenna elements. The polarization resource is one of the important space resources. Different from the space-separated antennas, the polarization antenna array can utilize the multiple field components to improve the spatial degree of freedom of MIMO systems within a limited space. And the antennas with different polarizations can locate in the same place to save the space occupied. The ports isolation is the challenge for antenna design. Another one is to enhance the antenna performance in the space-limited MIMO system, without increasing the antenna volume. Using switching mechanism, one more polarization or radiation pattern can be selected due to the channel conditions. Based on the adaptive antenna selection, suitable signal processing methods can be adopted alternatively to achieve better performance. The design of switching mechanism is the key issue for carefully consideration.

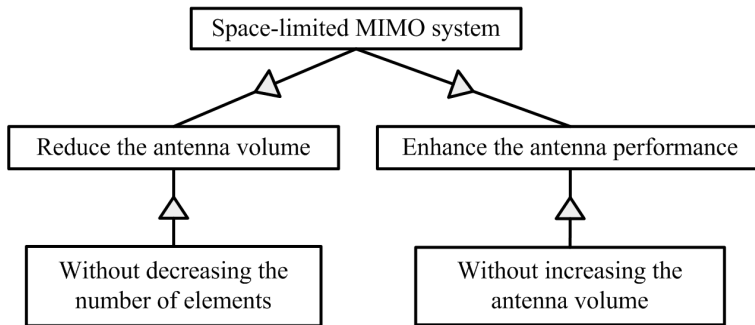


Fig. 1. Technical diagram for antenna design in space-limited MIMO system.

This chapter is organized as following. In Section 2, dual-polarized antenna solution is proposed as an example of 2-element polarization antenna array. Two practical designs are present to show the isolation enhancement between ports. Section 3 describes polarization reconfigurable antenna element based on the Section 2. Channel capacity benefit has been validated by experiment. In Section 4, another type of reconfigurable antenna, pattern reconfigurable antenna element is proposed. Section 5 will give a summary of this chapter.

2. Dual-polarized antenna

In this section, we talk about the polarization resource of antenna. The polarization antenna array has been studied in mobile communications for decades. In 1970s, the polarization characteristics of mobile wireless channel had been widely measured and discussed. The results illustrated that the correlation between feeding ports of different polarization antenna elements must be low to satisfy the requirements of diversity, and the volume occupied is much smaller than the space-separated antennas. Thus, more uncorrelated sub-channels can be obtained by using polarization antenna array. Further, the orientations of the mobile terminals are commonly not perpendicular to the ground. Polarization antenna

array is an effective solution to reduce the polarization mismatch. In traditional cellular mobile communication systems, the system with polarization diversity antennas has a 7 dB gain than the one with space diversity antennas in Line-of-Sight scenarios, and a 1 dB gain in Non-Line-of-Sight scenarios (Nakano et al, 2002).

In MIMO systems, the channel capacity of MIMO system with polarization antenna array is approximately 10%~20% higher than that with space-separated co-polarized antenna array, though the system SNR of polarization antenna array is lower (Kyritsi et al, 2002; Wallace et al, 2003). Another measurement results in micro- and pico-cell show the channel capacity of MIMO systems with dual-polarized antenna elements are 14% higher than that with twice-numbered single-polarized antennas (Sulonen et al, 2003). Similar results are also obtained (Erceg et al, 2006). Of course, the dual-polarized antenna element can be treated as a 2-element single-polarized antenna array. For this application, two important issues must be considered: one is the ports isolation, the other one is the antenna dimension. High-isolated compact-volume dual-polarized antenna is our goal of design.

In recent research, different methods of isolation enhancement are introduced. An air bridge, which is utilized in the cross part of two feedings for high isolation, was proposed in (Barba, 2008; Mak et al, 2007). Different feed mechanisms, feed by probe and coupling through aperture, were used in (Guo et al, 2002). Another isosceles triangular slot antenna is proposed for wideband dual-polarization applications in (Lee et al, 2009). TE₁₀ and TE₀₁ modes are excited by two orthogonally arranged microstrips. The above mentioned methods are difficult to be realized in a compact structure and unable to be adopted in space-limited multiple antenna systems. In this section, we introduce two compact antenna designs with good ports isolation.

2.1 Dual-polarized slot antenna

For the purpose to realize dual orthogonal polarizations, slot structure is selected as the main radiator. As shown in Fig. 2, both vertical and horizontal polarizations can exist simultaneously in a rectangular slot. The operating frequency is dictated by the widths of the slot. The slot also has the advantages of wide bandwidth, bi-directional radiation pattern and high efficiency (Lee et al, 2009). However, how to excite these two polarizations is still a question. The traditional method is to feed both polarizations in the same way through two adjacent sides of the slot. Thus, the feeding structure is simple but with large dimension, which isn't able to fulfil our requirement of compact size.

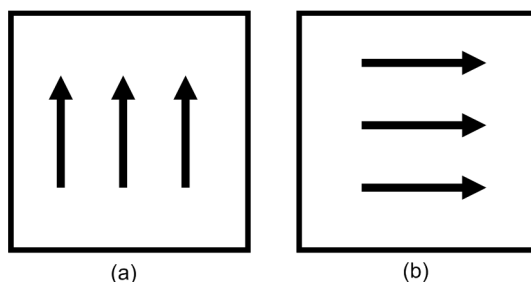


Fig. 2. Polarization mode in slot: (a) vertical polarization, (b) horizontal polarization.

In order to excite dual orthogonal polarizations in a compact structure, we utilized the dual modes of co-planar waveguide (CPW). Fig. 3 shows the geometry of the proposed antenna with CPW feeding structure. The overall dimensions of the antenna are $100 \times 80 \text{ mm}^2$. The antenna is made of the substrate of FR4 ($\epsilon_r=4.4$, $\tan\delta=0.01$), whose thickness is 1 mm. A $52 \times 50 \text{ mm}^2$ slot, etched in the front side of light region, serves as the main radiator. In the back side of dark region, an L-shaped microstrip line is fed through port 1. The CPW is fed through port 2 in the front side. As shown in Fig. 4(a), when feeding through port 1, a normal odd mode of CPW is excited to feed the vertical polarization mode. When feeding through port 2, as shown in Fig. 4(b), the mode in the CPW is the even mode as a slot line, which can excite the horizontal polarization mode.

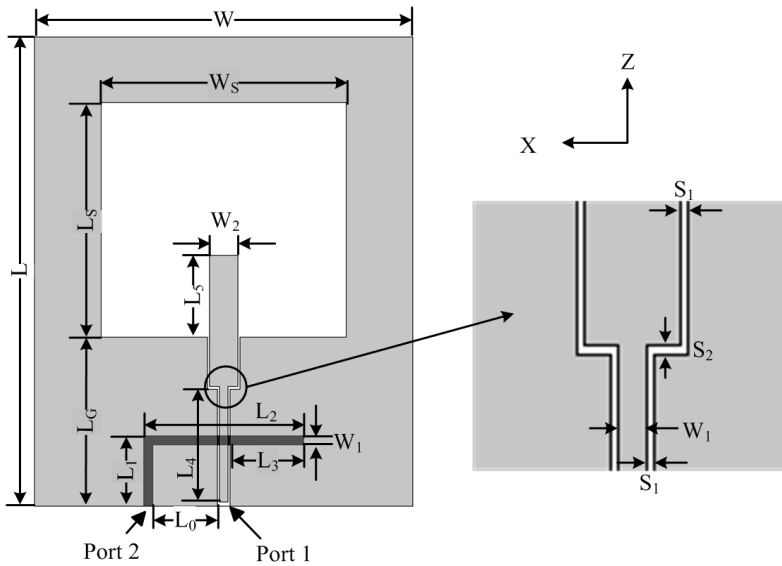


Fig. 3. The geometry of the proposed antenna. ($L=100 \text{ mm}$, $L_s=50 \text{ mm}$, $L_G=36 \text{ mm}$, $L_0=15 \text{ mm}$; $L_1=15 \text{ mm}$, $L_2=32 \text{ mm}$, $L_3=12.5 \text{ mm}$, $L_4=25.5 \text{ mm}$; $L_5=19 \text{ mm}$, $W_1=1.9 \text{ mm}$, $W_2=6 \text{ mm}$, $W_s=52 \text{ mm}$, $W=80 \text{ mm}$, $S_1=0.35 \text{ mm}$, $S_2=0.5 \text{ mm}$). Reprinted from (Li et al, 2010) by the permission of IEEE).

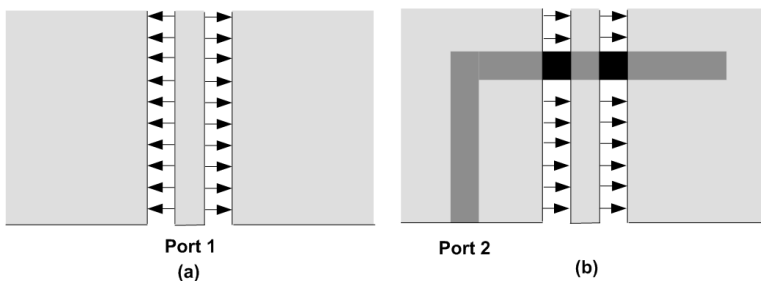


Fig. 4. Feeding modes in CPW: (a) odd mode, (b) even mode.

The current distributions of both polarizations are shown in Fig. 5 for better explanation. A half wavelength distribution appears on each side of slot. Dimensions of L_5 and W_5 determine the resonant frequencies of the vertical mode and horizontal mode respectively. The L_3 is the tuning parameter for matching port 1. To match port 2, dimensions of W_2 , L_5 and L_6 need to be optimized. Due to the symmetric and anti-symmetric characteristics of the two modes in CPW, high isolation can be achieved between two ports. As a result, the feeding structure can excite both polarization modes simultaneously and independently.

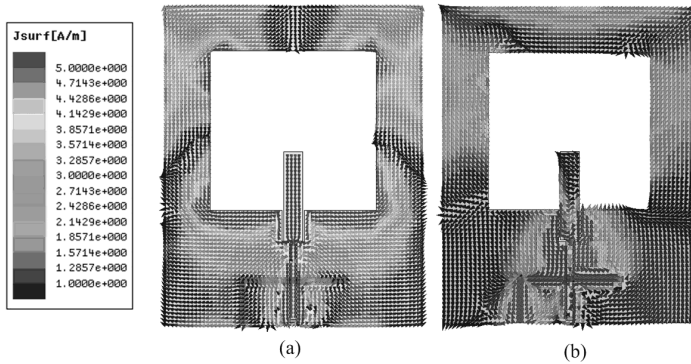


Fig. 5. Current distributions of (a) vertical polarization and (b) horizontal polarization.

To validate the design, the S parameters of the proposed antenna are simulated using Ansoft high frequency structure simulator (HFSS). The antenna has also been fabricated and measured. Fig. 6 shows the measured S parameter of the proposed antenna in solid lines, compared with the simulated ones in dash lines. The centre frequencies of the dual polarizations are both 2.4GHz. The bandwidths of -10dB reflection coefficient are 670MHz (1.96-2.63GHz, 27.9%) and 850MHz (1.93-2.75GHz, 35.4%) for horizontal polarization and vertical polarization, respectively. Throughout the WLAN frequency band (2.4-2.484GHz), the isolation between two ports in the required band is lower than -32.6dB. These results show that the proposed antenna is simpler, more compact than the references (Barba, 2008; Mak et al, 2007; Lee et al, 2009).

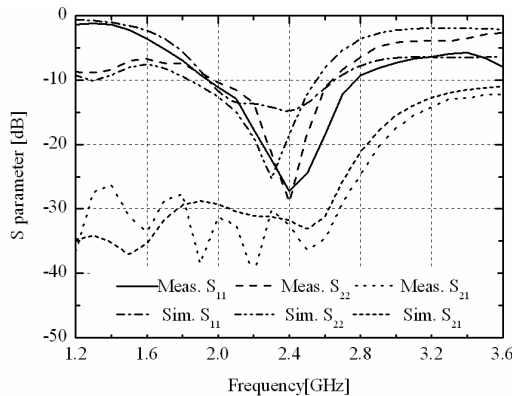


Fig. 6. Simulated and measured S parameters of the proposed antenna.

The radiation patterns of the proposed antenna when feeding through port 1 and 2 are shown in Fig. 7 and Fig. 8. For port 1, the vertical polarization case, the 3dB beam widths are

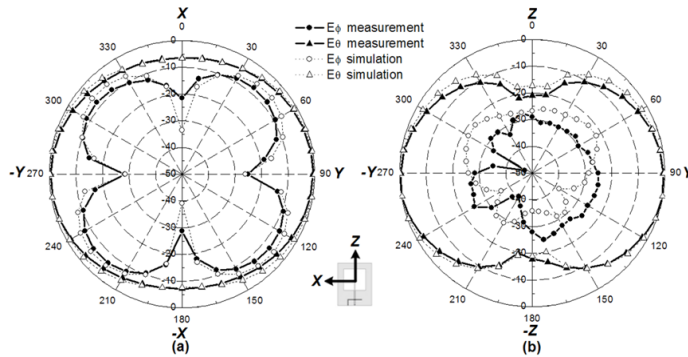


Fig. 7. Measured and simulated radiation patterns when feeding from port 1 at 2.4 GHz: (a) X-Y plane (b) Y-Z plane.

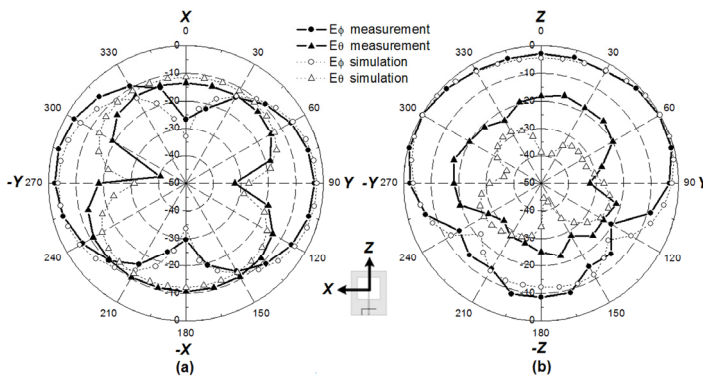


Fig. 8. Measured and simulated radiation patterns when feeding from port 2 at 2.4 GHz: (a) X-Y plane (b) Y-Z plane.

100° and 70° in E-plane (Y-Z plane) and H-plane (X-Y plane). From these results it may be noted that the cross polarization in X-Y plane is worse than what was achieved in earlier designs as values for cross polarization are not lower than -15dB. From the radiation patterns, however, we can observe that the poles of E_θ and E_ϕ are almost corresponding to the maximum of each other, which means the integration of the two patterns is close to zero. In other words, the signals of co and cross polarizations are almost uncorrelated. In the Y-Z plane, the cross polarization level is sufficiently low to be ignored. For port 2, the horizontal polarization is the dominant polarization. The 3dB beam widths are 60° and 180° in E-plane (X-Y plane) and H-plane (Y-Z plane). From the above discussion, we may conclude that the signals received by the two ports are uncorrelated, so dual-polarization in single antennas can be treated as two independent antennas. The radiation efficiency and gain of the proposed antenna are also measured. In the WLAN band of 2.4-2.484GHz, the efficiency is better than 91.2% and 84.4% for port 1 and 2; and the gain is better than 3.85 dBi and 5.21

dBi for port 1 and 2. The proposed antenna is a candidate for compact volume dual-polarized antenna application.

2.2 Dual-polarized loop antenna

The half wavelength resonant structure, such as the patch and the slot, is able to be adopted in dual-polarized antenna design. In order to realize even more compact dimension, we choose the loop antenna, whose circumference is one wavelength. The radiation patterns of the slot and the loop are almost the same. Also, the loop element can support two orthogonal polarizations using the same structure, shown in Fig. 9. Seen from these two modes, the current distribution is 90° rotated from one to another one. Good orthogonality is illustrated with high isolation. The current distribution of its one-wavelength mode is dictated by the feeding position, and feed should not be arranged at the position of the current null. However, the maximum point of one mode is the null of the other mode. It is difficult to feed the dual polarizations in one side of loop.

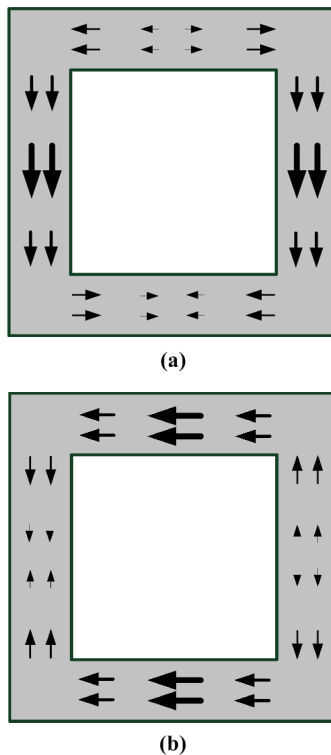


Fig. 9. Modes in loop antenna: (a) vertical polarization, (b) horizontal polarization.

The feeding method should be considered carefully. In order to excite two orthogonal one-wavelength modes, it is common to arrange two feeds at two orthogonal positions, which will make the overall dimension much larger. A compact size could be realized if such two modes of operation are fed at only one position. The compact CPW feed backed with

microstrip line adopted in the last design is an effective solution to feed the dual-mode of loop antenna. Fig. 10 shows the geometry of the loop antenna, which is quite similar as the slot design. This antenna consists of a rectangular loop, a CPW feeding and a microstrip line, and supported by the same FR4 board as last design with the thickness of 1 mm. The loop has width of 4 mm; narrower than the slot design. The loop and CPW are etched on the front side and the microstrip line is printed on the back side.

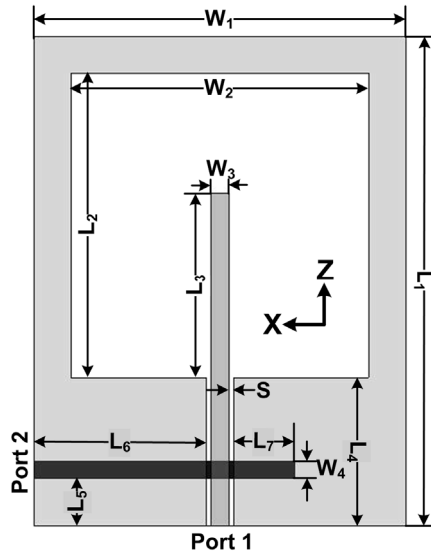


Fig. 10. Geometry of the proposed loop antenna. ($L_1=53$ mm, $L_2=33$ mm, $L_3=20$ mm, $L_4=16$ mm; $L_5=5.1$ mm, $L_6=18.5$ mm, $L_7=6.5$ mm, $W_1=40$ mm, $W_2=32$ mm, $W_3=2$ mm, $W_4=1.9$ mm, $S=0.5$ mm. Reprinted from (Li et al, 2011a) by the permission of IEEE).

When the loop fed through port 1, the CPW operates at its typical symmetrical mode. In this mode the vertical polarization is excited. The inner conductor works as a monopole with the vertical polarization. The energy is coupled from monopole to the loop, exciting the vertical polarization mode. It is a good solution to feed the one-wavelength mode at the position of current null. The radiation consists of two modes, the one-wavelength mode of the loop and a monopole mode. When the loop is fed through port 2, the horizontal polarization of the loop antenna is excited. The feed is exactly at the maximum of current, and the horizontal mode is clearly excited in this configuration.

Fig. 11 shows the current distributions of two polarizations, which are totally different from the slot antenna. For the same application of 2.4 GHz WLAN in last design, the rectangular slot is etched in a large ground. The slot's length and width are approximately half wavelength. For a typical slot mode, the width of extended ground is a quarter of wavelength or smaller. If the size of surrounded ground decreases to some level, the slot turns to be a loop mode with the frequency shift. What's more, a loop has four edges with the overall dimension of the loop antenna is 40×53 mm², including the feeding structure. The slot antenna is with the dimension of 100×80 mm². It is clear that the area of the proposed antenna is only 26.5% of the slot one. Fig. 12 (a) and (b) show the loop antenna, in front and

back views, respectively. Fig 12(c) shows the slot antenna design, which also operates in the same band. A significant size reduction is achieved using the loop design.

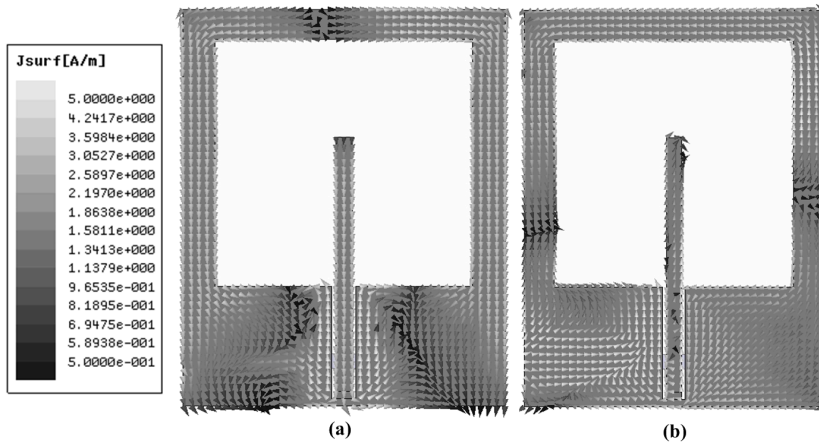


Fig. 11. Current distributions of (a) vertical polarization and (b) horizontal polarization.

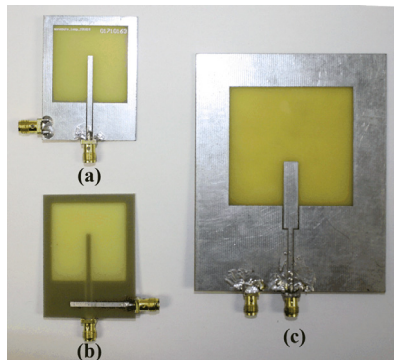


Fig. 12. Photograph of the loop antenna (a) front side, (b) back side and (c) the slot antenna. The total length is one wavelength. Therefore, the dimension of a rectangular loop antenna is much smaller than the slot design with large ground. However, the slot antenna can be adopted in the array design in the same ground for special requirements.

The measured and simulated S parameters are illustrated in Fig. 13. The -10 dB bandwidth of the reflection coefficients are 770 MHz (1.98-2.75 GHz, 32.1%) for the vertical polarization and 730 MHz (1.96-2.69 GHz, 30.4%) for the horizontal polarization, both covering the of 2.4 GHz WLAN band. The isolation in this band is better than -21.3 dB, which is lower than the slot design, as a cost of dimension reduction. The isolation deterioration is mainly contributed to the feeding structure of the vertical polarization. The feeding monopole is located at the current maximum point of the horizontal mode. The energy couples between two modes. But it still fulfils the -15 dB industrial requirement. The radiation patterns of the loop antenna is quite similar to the slot antenna, but with a lower level of cross polarization. In the 2.4 GHz WLAN band, the measured gains are better than 2.9 dBi and 4.1 dBi.

Considering the compact structure of loop, this antenna is suitable for the space-limited systems.

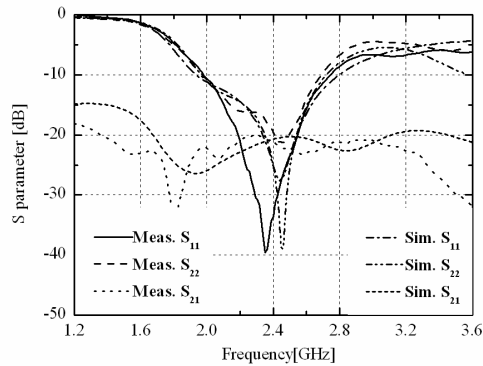


Fig. 13. Simulated and measured S parameters of the loop antenna.

3. Polarization reconfigurable antenna

As described in the introduction, reconfigurable antenna is an effective solution for the space-limited MIMO systems by adaptive antenna selection. This kind of systems is called adaptive MIMO system. The adaptive MIMO system takes the advantage of varying channel characteristics to make the best use of the improvement of channel capacity (Cetiner et al, 2004). Due to the channel condition, different antenna properties, such as polarizations and radiation patterns, are selected for better transmitting or receiving. Also, different data processing algorithms are used depending on the antenna selection. For this reason, the reconfigurable antenna is very important to the MIMO system, especially for the space-limited system. In this section, we will introduce the polarization reconfigurable antenna, based on the dual-polarized slot antenna described in the last section. In order to validate the benefit of polarization selecting, the channel capacity of a 2x2 MIMO system using the polarization reconfigurable antenna has been measured in a typical indoor scenario.

3.1 Reconfigurable mechanism

The geometry of the proposed reconfigurable slot antenna element is shown in Fig. 14, based on the design of (Li et al, 2010). The port 1 and port 2 are combined together and controlled by two PIN diodes. The port 1 is connected the microstrip line on the back side through a via hole, and controlled by PIN 1. The port 2 is connected directly to the microstrip line on the back side, and controlled by PIN 2. When PIN1 is ON and PIN2 is OFF, the antenna is fed through the port 1. The vertical polarization of the slot is excited. When PIN1 is OFF and PIN 2 is ON, the antenna is fed through the port 2, and the horizontal polarization of the slot is excited. Therefore, two ports are fed alternatively and controlled by the PIN diodes. The two PIN diodes need the bias circuit to control. Due to compact feed design, the two PIN diodes share the same bias circuit, saving the space of the antenna system and using less lumped components.

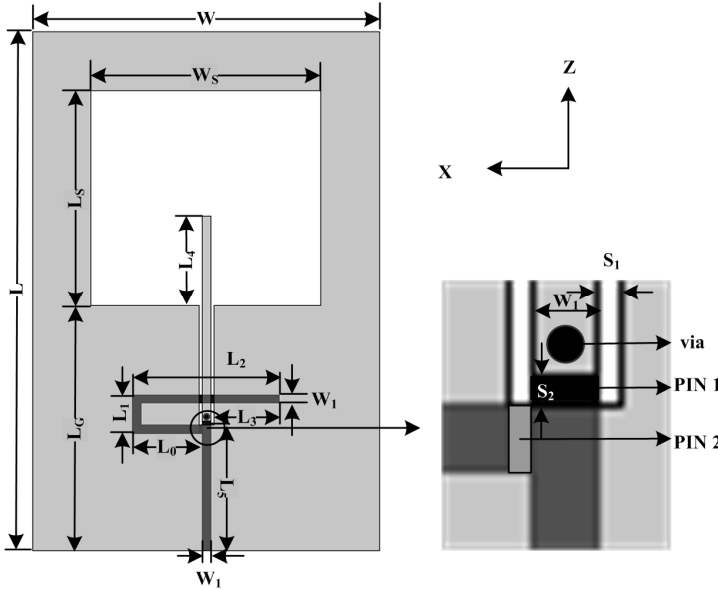


Fig. 14. Geometry of the proposed loop antenna. ($L=120$ mm, $L_S=50$ mm, $L_G=36$ mm, $L_0=16$ mm; $L_1=8.9$ mm, $L_2=33.9$ mm, $L_3=15.3$ mm, $L_4=20.1$ mm; $L_5=30$ mm, $W_1=1.9$ mm, $W_S=53$ mm, $W=80$ mm, $S_1=0.7$ mm, $S_2=1$ mm. Reprinted from (Li et al, 2011b) by the permission of John Wiley & Sons, Inc.).

A prototype of the dual-polarized slot antenna with switching mechanism is fabricated, and shown in Fig. 15. The PIN diodes with bias circuit are on the back side of the antenna. The detailed bias circuits of two PIN diodes (D1 and D2, Philips BAP64-03) are shown in Fig. 15 (c). The 'ON' and 'OFF' states of the two PIN diodes are controlled by a 1-bit single-pole 2-throw (SP2T) switch on the front side. The bias circuit consists of three RF choke inductors (L_{b1} , L_{b2} and L_{b3} , 12 nH), a DC block capacitor (C_b , 120 pF), three RF shorted capacitors (C_{s1}

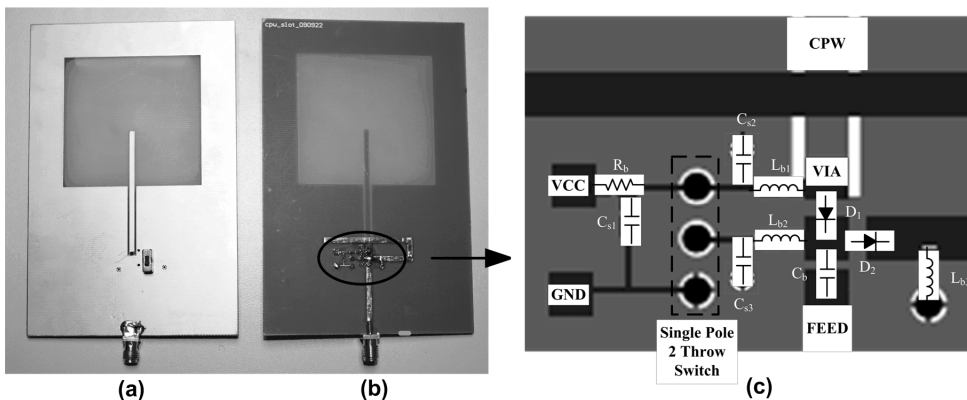


Fig. 15. Photograph of the antenna prototype (a) front side, (b) back side; (c) bias circuit of the PIN diodes.

C_{s2} and C_{s3} , 470 pF) and a bias resistor (R , 46 Ω). The bias resistor is selected depend on the value of VCC and the operating current of the PIN diode. In this application, the VCC is 3 V.

The measured reflection coefficients for both polarizations are shown in Fig. 16. Compared with results of the dual-polarized slot antenna in Fig.13, the difference is mainly contributed from the parasitic parameters of PIN diodes and the bias circuit. The -10dB bandwidths are 700MHz (2.02-2.72 GHz, 29.2%) and 940MHz (1.84-2.78 GHz, 40%) for vertical and horizontal polarizations, both covering the WLAN band (2.4-2.484 GHz). The gain decreases approximately 0.5 dB due to the insertion loss of PIN diodes.

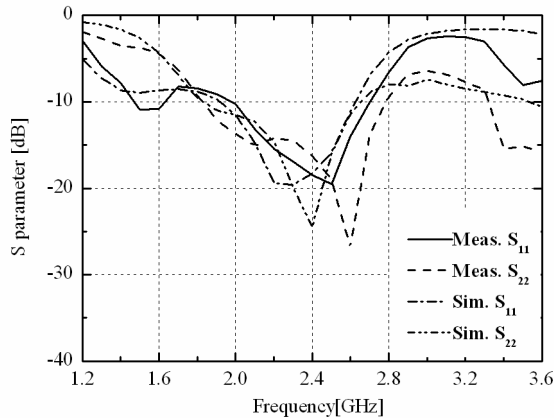


Fig. 16. Simulated and measured S parameters of the reconfigurable antenna.

3.2 Channel capacity measurement

In this section, we measured the channel capacity of a 2x2 MIMO system in a typical indoor scenario by using the proposed polarization reconfigurable antenna. The measurement setup is shown in Fig.17. The measurement system consists of an Agilent E5071B Vector network analyzer (VNA), which has 4 ports for simultaneous measurement, transmit and receive antennas, a computer and RF cables. Two standard omni-directional dipoles are utilized as the transmit antennas (TX), and arranged perpendicular to XY plane along Z axis. Two proposed reconfigurable antennas are used as the receive antennas (RX). The 2x2 antennas are connected to the 4 ports of the VNA. The computer is used to control the measurement procedures and record the measured channel responses. In order to validate the improvement in channel capacity by using reconfigurable antennas, another two reference dipoles are adopted as receive antennas for comparison. The measurement was carried out in a room of the Weiqing building, Tsinghua University, illustrated in Fig. 18. The framework of the room is reinforced concrete, the walls are mainly built by brick, and the ceiling is made with plaster plates with aluminium alloy framework. The heights of desk partition and wood cabinet are 1.4 m and 2.1 m. The transmit antennas are fixed in the middle of room (TX). The receive antennas are arranged in several typical locales which are noted as RX1-5 in Fig. 20. Here, the scenarios when the receive antennas are arranged in RX1 and RX2 are line-of sight (LOS), while that is NLOS when the receive

antennas are arranged in RX3, RX4 and RX5. In this measured, the antennas used are fixed at the height of 0.8 m. The space of antenna elements in TX or RX is 0.5λ , with the mutual coupling less than -25dB .

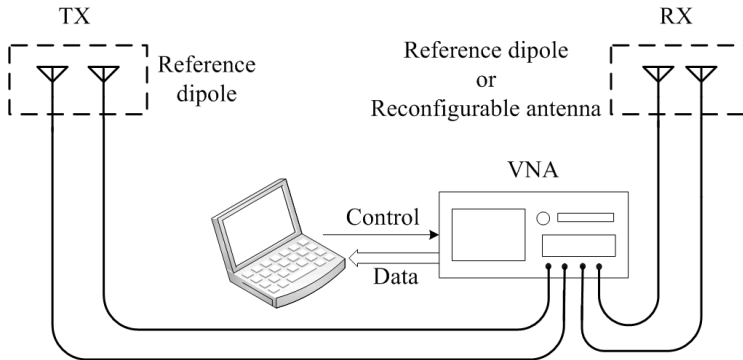


Fig. 17. Experiment setup of the measurement.

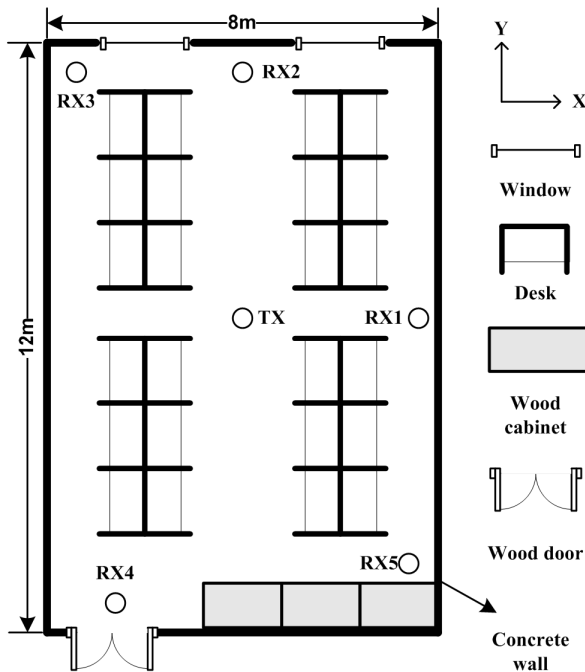


Fig. 18. Layout of measurement environment.

The measurement was carried out in the band of 2.2-2.6 GHz, with a step of 2 MHz. Three different orientations (ZZ, YY, and XX) of RX antennas were measured to simulate different operational poses of the mobile terminals. For two horizontal (H) and vertical (V) polarizations reconfigurable antennas, 4 configurations (HH, HV, VH, VV) were switched

manually for each channel capacity measurement in a quasi-static environment, and the result with the biggest value was chosen for statistics. Given the small-scale fading effect, 4x4 grid locations for each RX position were measured. Therefore, a total $201 \times 3 \times 16 \times 2 = 19296$ measured channel capacity for LOS condition was obtained, and $201 \times 3 \times 16 \times 3 = 28944$ was the measured results for NLOS condition.

The channel capacity can be calculated through following formula (Foschini & Gans, 1998):

$$C = \log_2 \det \left[I_{N_r} + \frac{SNR}{N_t} H_n H_n^H \right] \quad (1)$$

where N_r and N_t are the numbers of RX and TX antennas. I_{N_r} is a $N_r \times N_r$ identity matrix, SNR is the signal-to-noise ratio at RX position, H_n is the normalized H , and $()^H$ is the Hermitian transpose. H is normalized by the received power in the 1x1 reference dipole with identical polarization. We selected the SNR when the average channel capacity is 5 bit/s/Hz in a 1x1 reference dipole system in LOS or NLOS scenario.

The measured Complementary Cumulative Distribution Functions (CCDF) of the channel capacity for the 2x2 MIMO system using polarization reconfigurable antennas in both LOS and NLOS conditions are shown in Fig. 19 and 20. As summarized in Table 1, the average and 95% outage channel capacities are both improved, especially in NLOS scenario. For NLOS, the received signal is mainly contributed from reflection and diffraction, which vary the polarization property of the wave. However, the path loss is higher in NLOS scenario. The transmit power should be enhanced to guarantee the system performance. Considering the insertion loss introduced from non-ideal PIN diodes, better capacity can be obtained by using high quality components. The measurement results prove the benefit by using polarization reconfigurable antennas.

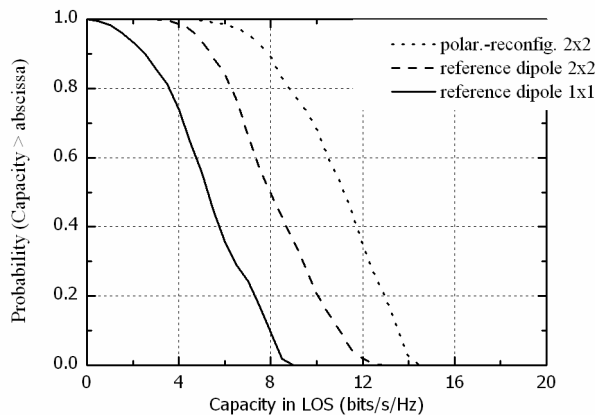


Fig. 19. CCDFs of channel capacity in LOS condition.

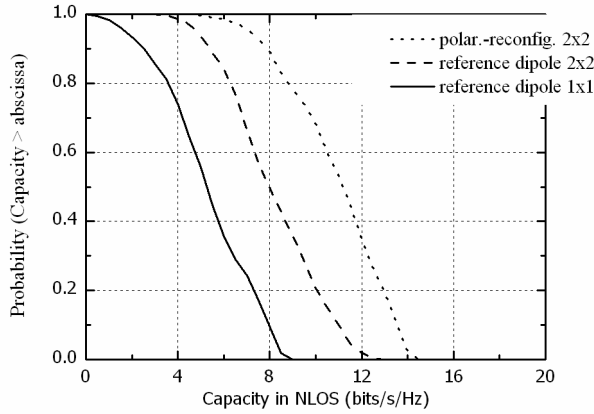


Fig. 20. CCDFs of channel capacity in NLOS condition.

Channel capacity	Condition	1x1 dipole	2x2 dipole	2x2 polar.-reconfig.
Average	LOS	5	7.86	10.62
	NLOS	5	9.9	13.18
95% outage	LOS	1.75	4.91	7.11
	NLOS	1.94	6.87	11.32

Table 1. Average and 95% Outage Channel Capacity (bit/s/Hz).

4. Pattern reconfigurable antenna

Pattern reconfigurable antenna is another type of reconfigurable antenna. Such antenna provides dynamic radiation coverage and mitigates multi-path fading. In this section, we introduce a design of pattern reconfigurable antenna with compact feeding structure. The benefit by using pattern reconfigurable antennas in the MIMO system is also proved by experiment of channel capacity measurement.

The configuration of the pattern reconfigurable antenna is shown in Fig. 21 (a). It is composed of an elliptical topped monopole, two Vivaldi notched slots and a typical CPW feed with 2 PIN diodes. The antenna is printed on the both sides of a 50 x 50 mm² Teflon substrate, with $\epsilon_r=2.65$, $\tan\delta=0.001$ and thickness is 1.5 mm. The CPW is connected to the microstrip at the back side through several via holes. A 0.2 mm wide slit is cut from the ground on the front side for DC isolation. Three curves are used to define the shape of antenna, fitted to the coordinates in Fig. 21 (a). Curve 1 is defined by equation (2) and curve 2 is defined by equation (3). Curve 3 and curve 2 are symmetrical along X axis.

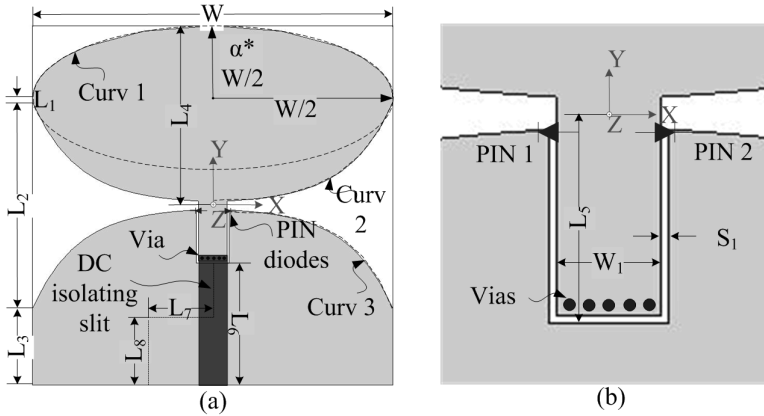


Fig. 21. Geometry of the proposed loop antenna. ($L_1=1.74$ mm, $L_2=28.52$ mm, $L_3=10.74$ mm, $L_4=25$ mm; $L_5=8$ mm, $L_6=16.8$ mm, $L_7=10$ mm, $L_8=10$ mm, $W=50$ mm, $L_p=2$ mm, $W_1=4$ mm, $S_1=0.3$ mm. Reprinted from (Li et al, 2010c) by the permission of IEEE).

$$\left(\frac{x}{W/2}\right)^2 + \left[\frac{y - (L_4 - \alpha * W/2)}{\alpha * W/2}\right]^2 = 1 \quad (2)$$

where $L_4 - \alpha * W/2 \leq y \leq L_4$, and $\alpha = 0.4$.

$$y = C_1 e^{c \cdot x} + C_2 \quad (3)$$

where $C_1=14$, $C_2=0.26$, $c=0.16$.

4.1 CPW-slot transition design

Different radiation patterns are provided by different work states of the same antenna. In order to achieve different work states, a switchable CPW-to-slotline transition with two PIN diodes is proposed and shown in Fig. 21 (b). Three feeding modes are achieved in this structure by varying the states of PIN diodes. When both PIN diodes are OFF, the elliptical topped monopole is fed through a typical CPW and a nearly omni-directional radiation pattern is achieved in XZ plane. When PIN 1 is OFF and PIN 2 is ON, the right slotline is shorted. The left Vivaldi notched slot is fed through the left slotline (LS) of the CPW, and a unidirectional radiation pattern is formed along the $-X$ axis. In the same way, when PIN 1 is ON and PIN 2 is OFF, a unidirectional beam along the $+X$ axis is obtained in the right Vivaldi notched slot through the right slot (RS). The proposed CPW-to-slotline transition is able to achieve good switching from the CPW to slotline with any other extra structures. Compared with this design, the CPW-to-slotline transition reported in (Wu et al, 2008; Kim et al, 2007; Ma et al, 1999) all required extra structures for mode convergence, including $\lambda/2$ phase shifter (Ma et al, 1999) and $\lambda/4$ matching structures (Wu et al, 2008; Kim et al, 2007), which occupy considerable space in the feed network. Such structures are not suitable for the space-limited systems. The proposed CPW-to-slotline transition here is designed to reduce the overall dimensions of the antenna.

In order to explain work principle of the feed transition, the equivalent transmission line model is utilized, illustrated in Fig.22 and 23. The PIN diode is expressed as perfect conductor for 'ON' state and open circuit for the 'OFF' state. Fig. 22 (a) shows the normal CPW structure. By tuning the L_5 , the radiation resistance R_{monopole} of monopole is matched to 50Ω at the feed port. When the right slot is shorted by PIN diode, the antenna is fed through the RS mode. The diagram and equivalent transmission line model are depicted in Fig. 22 (b). The right slotline is used to feed the Vivaldi notched slot, and the shorted left slotline works as a matching branch. The shorted branch which is less than a quarter of wavelength serves as a shunt inductance and its value is determined by its length L_5 . The position of the PIN diode is not fixed, and it is another freedom for impedance matching of RS feed. As shown in Fig. 23, the value of shunt inductance is $jZ_{\text{slot}}\tan[\beta_{\text{slot}}(L_5-L_p)]$ and used to match the radiation resistance R_{vivaldi} . The advantage of this switchable feeding structure is that no extra structure is used in the CPW and slotline transition.

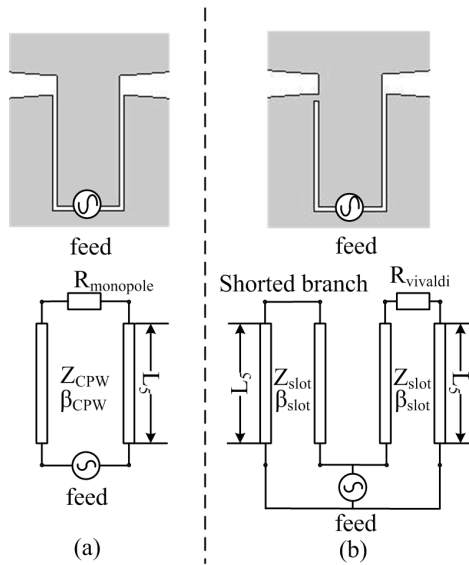


Fig. 22. Feed diagram and equivalent transmission line model. (a) CPW feed; (b) RS feed.

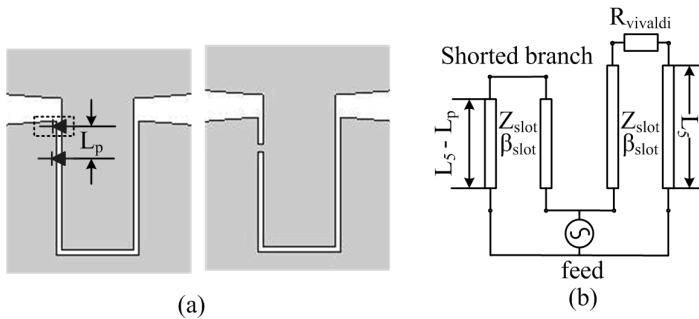


Fig. 23. Matching strategy of RS feed. (a) Feed diagram; (b) Transmission line model.

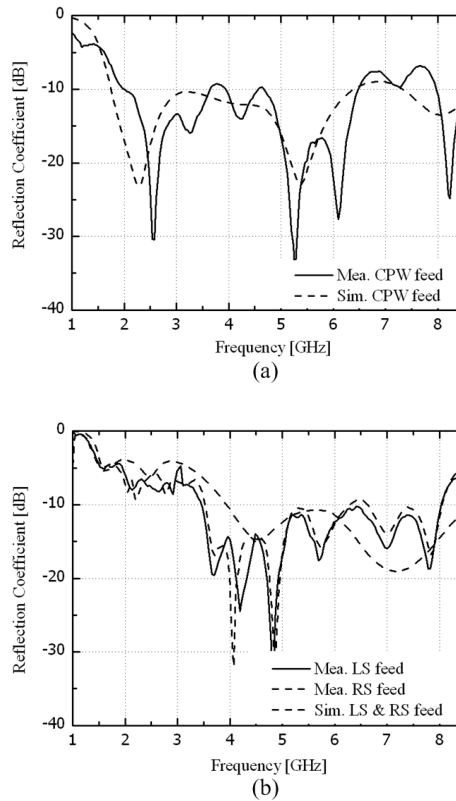


Fig. 24. Simulated and measured reflection coefficient of the reconfigurable antenna.

The selected PIN diode is Agilent HPND-4038 beam lead PIN diode, with acceptable performance in a wide 1-10 GHz bandwidth. The bias circuit is similar as the PIN diodes used in the last design in Fig. 15. The values of each component are determined by the working current of the PIN diode. The efficiency decreases approximately 0.3 dB by using this PIN diode. All the measurements were taken using an Agilent E5071B VNA. The simulated and measured reflection coefficients of CPW feed, LS and RS feeds are shown in Fig. 24. The measured -10dB bandwidths are 2.02-6.49 GHz, 3.47-8.03 GHz and 3.53-8.05 GHz for CPW feed, LS feed and RS feed, respectively. The overlap band from 3.53 GHz to 6.49 GHz is treated as the operation frequency for the reconfigurable patterns. The measured normalized radiation pattern in XZ and XY planes for CPW, LS and RS feed at 4, 5, 6 GHz are shown in Fig. 25. For the CPW feed, a nearly omni-directional radiation pattern appears in XZ plane and a doughnut shape in XY plane. For the LS or RS feed, a unidirectional beam appears along $-X$ or $+X$ axis, with acceptable front-to-back ratio better than 9.5dB. For the CPW feed, an average gain in the desired frequency range is 2.92 dBi. For the LS and RS feed, the average gains in the 4-6 GHz band are 4.29 dBi and 4.32 dBi. The improved gain is mainly contributed to the directivity of the slotline feed mode, and the diversity gain is achieved by switching the patterns.

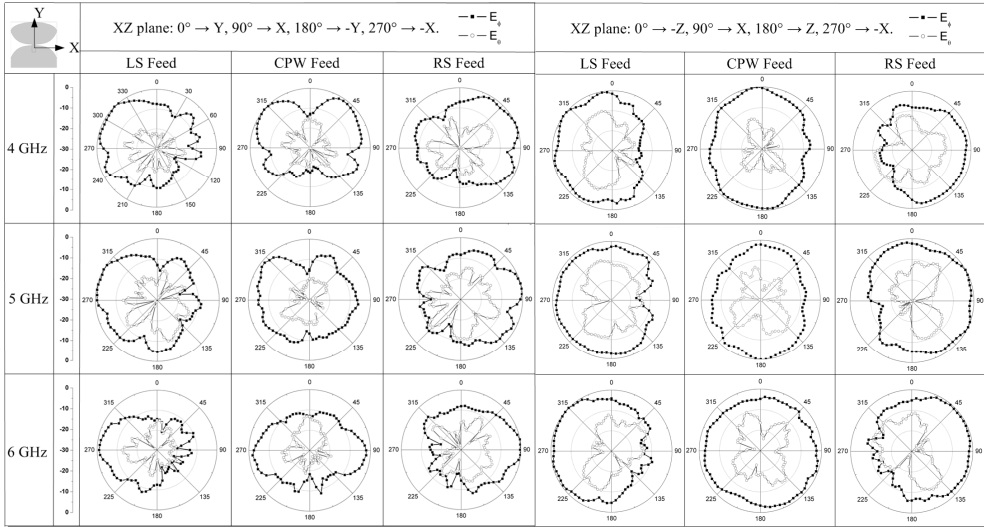


Fig. 25. Radiation patterns of the reconfigurable antenna.

4.2 Channel capacity measurement

The channel capacity of a 2x2 MIMO system by using the proposed pattern reconfigurable antenna is measured in this section. The experiment setup is as same as Fig. 17. At the TX end, two reference dipoles are arranged perpendicular to XZ plane along Y axis. Another two reference dipoles and two proposed pattern reconfigurable antennas are adopted at the RX end alternatively for comparison. Each port of the two wire dipoles has a bandwidth of 3.9-5.9 GHz with reflection coefficient better than -6 dB, and mutual coupling between the two ports is lower than -25 dB over the frequency band which is achieved by tuning the distance between two elements. Also, the isolation between two proposed pattern reconfigurable antennas is lower than -25 dB.

The measurement was also taken in the Weiqing building of Tsinghua University of Fig. 18. The locations of RX are different from last experiment. The position of RX4 is not measured. Therefore, the LOS scenario includes the RX1 and RX2, and the NLOS scenario includes the RX3 and RX5. The frequency range of measurement is 4-6 GHz, with a step of 10 MHz. A total number of 201 data points/results are obtained as samples. Three configurations (CPW, LS and RS) of each reconfigurable element of the receive end were switched together manually and the highest value signal was selected as the receiving signal. Also, considering the small-scale fading effect, 5x5 grid locations for each RX position were arranged. A total number of $2 \times 201 \times 25 = 10050$ results were measured for LOS and NLOS scenarios respectively. In the measurement, a 2x2 channel matrix H is obtained. The channel capacity is calculated by formula (1) in the last section. We also selected the SNR when the average channel capacity is 5 bit/s/Hz in a 1x1 reference dipole system in LOS or NLOS scenario.

The measured CCDFs of channel capacity of LOS and NLOS scenarios are illustrated in Fig. 26 and 27. The results consist of the channel capacity information of 2x2 multiple antenna

system using the proposed pattern reconfigurable antennas, compared with 1x1 and 2x2 systems using reference dipoles. As listed in Table 2, 2.28 bit/s/Hz and 4.13 bit/s/Hz of the average capacity enhancement are achieved in LOS and NLOS scenarios, and 2.51 bit/s/Hz and 3.75 bit/s/Hz enhancement for 95% outage capacities. In the NLOS scenario, the received signal is mainly contributed from reflection and diffraction of the concrete walls and the desk partitions, arriving at the direction of endfire. The diversity gain in the endfire increases the channel capacity. Considering the insertion loss introduced from the non-ideal PIN diodes, better performance of the proposed antenna can be achieved by using high quality switches, such as micro-electro-mechanical systems (MEMS) type switches with less insertion loss and parasitic parameters.

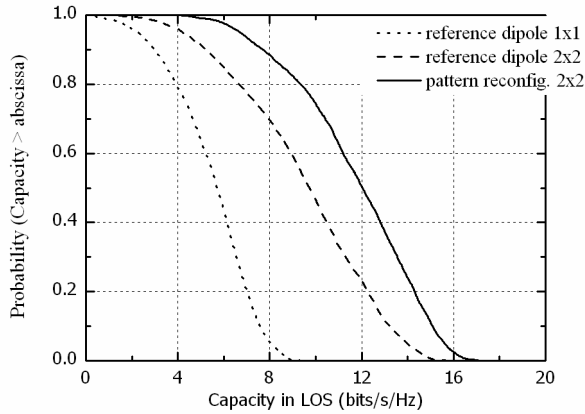


Fig. 26. CCDFs of channel capacity in LOS condition.

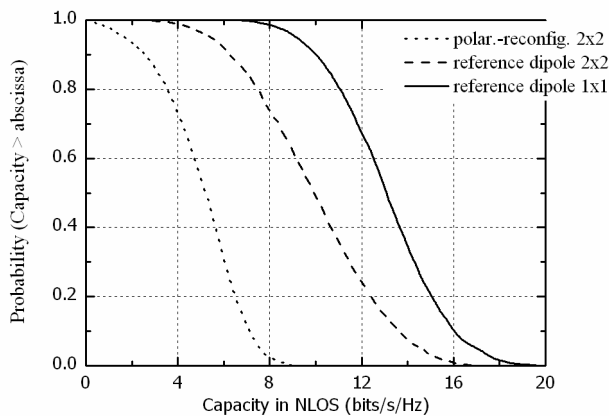


Fig. 27. CCDFs of channel capacity in NLOS condition.

Channel Capacity	Scenario	1x1 Dipole	2x2 Dipole	2x2 Pattern Reconfig.
Average	LOS	5	9.46	11.74
	NLOS	5	9.93	13.06
95% Outage	LOS	2.29	4.21	6.72
	NLOS	1.68	5.41	9.16

Table 2. Average and 95% Outage Channel Capacity (bit/s/Hz).

5. Conclusion

This chapter has introduced the now trend of antenna design in MIMO systems. From a view of antenna design, it is difficult to achieve good performance by using traditional antennas in space-limited MIMO systems. The mutual coupling between the antenna elements deteriorates the independence among multiple channels in MIMO systems. For both techniques of TD and SM, the benefit of multiple channels is difficult to be obtained due to the space limitation. However, the usage of miniaturized mobile terminals is popular and their size is getting much smaller. Here comes the contradiction between the antenna performance and the ministration of mobile handsets.

In this chapter, we are aimed to solve the space problem of antennas in MIMO systems. Two effective solutions are introduced here. The first one is to use polarization, an important spatial resource, to take the place of antenna element. Two orthogonal polarized antenna elements can be arranged together with acceptable isolation. In this way, the space between antenna elements is saved, making the overall antenna system more compact. As an important practical application, two types of dual-polarized antennas are presented and analyzed. Isolation enhancement methods are proposed, such as the feed design and operation modes design. The proposed antennas show the advantages of compact structure, high ports isolation and easy fabrication, and are suitable to be adopted in the space-limited MIMO systems.

Considering in the opposite way, the antennas with better performance in the original space is another solution in space-limited MIMO systems. The reconfigurable antenna is a prevalent type of antenna nowadays. Switching mechanism is added to achieve selectable polarizations, radiation patterns and other property. Different antenna configurations and corresponding signal processing methods are selected due to the channel information. The switching mechanism is the most important issue. Based on the dual-polarized slot antenna design, the PIN diodes are used to achieve polarization selection. A pattern reconfigurable antenna is design by using a switchable CPW-to-slotline feeding structure. In order to prove the benefit of the antenna selection, we design an experiment of channel capacity in a typical indoor environment. The results show that the channel capacity improves in both LOS and NLOS scenarios, especially in NLOS scenario. The reconfigurable antenna shows the potential application in space-limited MIMO systems.

6. Acknowledgment

This work is supported by the National Basic Research Program of China under Contract 2009CB320205, in part by the National High Technology Research and Development

Program of China (863 Program) under Contract 2009AA011503, the National Science and Technology Major Project of the Ministry of Science and Technology of China 2010ZX03007-001-01, and Qualcomm Inc..

7. References

- Barba, M. (2008). A High-Isolation, Wideband and Dual-Linear Polarization Patch Antenna. *IEEE Transactions on Antenna and propagation*, vol.56, No.5, (May 2008), pp. 1472-1476, ISSN 0018-926X.
- Bolcskei, H.; Nabar, R.; Erceg, V.; Gesbert, D. & Paulraj A. (2001). Performance of spatial multiplexing in the presence of polarization diversity. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2437-2440, ISBN 0-7803-7041-4, Salt Lake City, Utah, USA, May 7-11, 2001
- Bolcskei, H.; Gesbert, D. & Paulraj A. (2002). On the capacity of OFDM based spatial multiplexing systems. *IEEE Transactions on Communications*, vol.50, No.2, (February 2002), pp. 225-234, ISSN 0090-6778.
- Cetiner, B.; Jafarkhani, H.; Qian, J.; Hui, J.; Grau, A. & De Flaviis, F. (2004). Multifunctional reconfigurable MEMS integrated antennas for adaptive MIMO systems. *IEEE Communications Magazine*, vol.42, (December 2004), pp. 62-70, ISSN 0163-6804.
- Erceg, V.; Sampath, H. & Catreux-Erceg, S. (2006). Dual-Polarization versus single-polarization MIMO channel measurement results and modeling. *IEEE Transactions on Wireless Communication*, vol.5, No.1, (January 2006), pp. 28-33, ISSN 1536-1276.
- Foschini, G. & Gans, M. (1998). On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Communications*, vol.6, No.3, (Mar 1998), pp. 311-335, ISBN 0201634708.
- Guo, Y.; Luk, K. & Lee K. (2002). Broadband dual polarization patch element for cellular-phone base stations. *IEEE Transactions on Antenna and propagations*, vol.50, No.2, (February 2002), pp. 251-253, ISSN 0018-926X.
- Kim, H.; Chung, D.; Erceg, V.; Anagnostou, D. & Papapolymerou, J. (2007). Hardwired Design of Ultra-Wideband Reconfigurable MEMS Antenna. *Proceedings of IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1-4, ISBN 1-4244-1144-0, Athens, Greece, September 3-7, 2007.
- Kyritsi, P.; Cox, D.; Valenzuela, R. and Wolniansky, P. (2002). Effect of antenna polarization on the capacity of a multiple element system in an indoor environment. *IEEE Journal on Selected Areas in Communications*, vol.20, No.6, (August 2002), pp. 1227-1239, ISSN 0733-8716.
- Lee, C.; Chen, S. & Hsu, P. (2009). Isosceles Triangular Slot Antenna for Broadband Dual Polarization Applications. *IEEE Transactions on Antenna and propagations*, vol.57, No.10, (October 2009), pp. 3347-3351, ISSN 0018-926X.
- Li, Y.; Zhang, Z.; Chen, W.; Feng, Z. & Iskander, M. (2010). A dual-polarization slot antenna using a compact CPW feeding structure. *IEEE Antennas and Wireless Propagations Letter*, vol.9, (December 2009), pp. 191-194, ISSN 1536-1225.
- Li, Y.; Zhang, Z.; Feng, Z. & Iskander, M. (2011). Dual-mode Loop Antenna with Compact Feed for Polarization Diversity. *IEEE Antennas and Wireless Propagations Letter*, vol.10, (December 2010), pp. 95-98, ISSN 1536-1225.

- Li, Y.; Zhang, Z.; Zheng, J. & Feng, Z. (2011). Channel capacity study of polarization reconfigurable slot antenna for indoor MIMO system. *Microwave and Optical Technology Letters*, vol.53, No.6, (March 2011), pp. 1029-1213, ISSN 1098-2760.
- Li, Y.; Zhang, Z.; Zheng, J.; Feng, Z. & Iskander, M. (2011). Experimental Analysis of a Wideband Pattern Diversity Antenna with Compact Reconfigurable CPW-to-Slotline Transition Feed. *IEEE Transactions on Antenna and propagations*, accepted for publication, ISSN 0018-926X.
- Lin, Y. & Chen, C. (2000). Analysis and applications of a new CPW-slotline transition. *IEEE Transactions on Microwave Theory and Techniques*, vol.48, No.3, (March 2000), pp. 463-466, ISSN 0018-9480.
- Ma, K.; Qian, Y. & Itoh, T. (1999). Analysis and applications of a new CPW-slotline transition. *IEEE Transactions on Microwave Theory and Techniques*, vol.47, No.4, (April 1999), pp. 426-432, ISSN 0018-9480.
- Mak, K.; Wong, H. & Luk, K. (2007). A Shorted Bowtie Patch Antenna with a Cross Dipole for Dual Polarization. *IEEE Antennas and Wireless Propagations Letter*, vol.6, (June 2007), pp. 126-129, ISSN 1536-1225.
- Marzetta, T. & Hochwald, B. (1999). Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading. *IEEE Transactions on Information Theory*, vol.45, No.1, (January 1999), pp. 139-157, ISSN 0018-9448.
- Morris, M. & Jensen, M. (2005). Superdirectivity in MIMO systems. *IEEE Transactions on Antenna and propagation*, vol.53, No.9, (September 2005), pp. 2850-2857, ISSN 0018-926X.
- Nabar, R.; Bolcskei, H.; Erceg, V.; Gesbert, D. & Paulraj A. (2002). Performance of multi-antenna signaling techniques in the presence of polarization diversity. *IEEE Transactions on Signal Processing*, vol.50, No.10, (October 2002), pp. 2553-2562, ISSN 1053-587X.
- Nakano, M.; Satoh, T.; & Arai, H. (2002). Uplink polarization diversity measurement with human body effect at 900 MHz. *Electronics and Communications in Japan*, vol.85, No.7, (July 2002), pp. 32-44, ISSN 1520-6424.
- Raleigh, G. & Cioffi, J. (1998). Spatio-temporal coding for wireless communication. *IEEE Transactions on Communications*, vol.46, No.3, (March 1998), pp. 357-366, ISSN 0090-6778.
- Shin, J.; Chen, S. & Schaubert, D. (1999). A parameter study of stripline-fed Vivaldi notch-antenna arrays. *IEEE Transactions on Antenna and propagations*, vol.47, No.5, (May 1999), pp. 879-886, ISSN 0018-926X.
- Sulonen, K.; Suvikmnas, P.; Vuokko, L.; Kivinen, J. and Vainikainen, P. (2003). Comparison of MIMO antenna configurations in Picocell and Microcell environments. *IEEE Journal on Selected Areas in Communications*, vol.21, No.5, (June 2003), pp. 703-712, ISSN 0733-8716.
- Tarokh, V.; Seshadri, N.; & Calderbank, A. (1998). Space-time codes for high data rate wireless communication: performance criterion and code construction. *IEEE Transactions on Information Theory*, vol.44, No.2, (March 1998), pp. 744-765, ISSN 0018-9448.
- Telatar, I. (1999). Capacity of multi-antenna Gaussian channels. *European Transactions on Telecommunications*, vol.10, No.6, (December 1999), pp. 585-595, ISSN 1541-8251.

- Wallace, J.; Jensen, M. Swindlehurst, A. & Jeffs, B. (2003). Experimental characterization of the MIMO wireless channel: Data acquisition and analysis. *IEEE Transactions on Wireless Communication*, vol.2, No.2, (March 2003), pp. 335-343, ISSN 1536-1276.
- Wallace, J. & Jensen, M. (2004). Mutual coupling in MIMO wireless systems: A rigorous network theory analysis. *IEEE Transactions on Wireless Communication*, vol.3, No.4, (July 2004), pp. 2437-2440, ISSN 1536-1276.
- Winters, J. (1987). On the capacity of radio communication systems with diversity in a Rayleigh fading environment. *IEEE Journal on Selected Areas in Communications*, vol.5, No.5, (June 1987), pp. 871-878, ISSN 0733-8716.
- Wu, S.; Chen, S. & Ma, T. (2008). A Wideband Slotted Bow-Tie Antenna with Reconfigurable CPW-to-Slotline Transition for Pattern Diversity, *IEEE Transactions on Antenna and propagations*, vol.56, No.2, (February 2008), pp. 327-334, ISSN 0018-926X.

Review of the Wireless Capsule Transmitting and Receiving Antennas

Zhao Wang, Eng Gee Lim, Tammam Tillo and Fangzhou Yu
Xi'an Jiaotong - Liverpool University
P.R. China

1. Introduction

The organization of American Cancer Society reported that the total number of cancer related to GI track is about 149,530 in the United State only for 2010 (American Cancer Society, 2010). Timely detection and diagnoses are extremely important since the majority of the GI related cancers at early-stage are curable.

However, the particularity of the alimentary track restricts the utilization of the current available examine techniques. The upper gastrointestinal tract can be examined by Gastroscopy. The bottom 2 meters makes up the colon and rectum, and can be examined by Colonoscopy. In between, lays the rest of the digestive tract, which is the small intestine characterised by being very long (average 7 meters) and very convoluted. However, this part of the digestive tract lies beyond the reach of the two previously indicated techniques. To diagnose the small intestine diseases, the special imaging techniques like CT scan or MRI are less useful in this circumstance.

Therefore, the non-invasive technique Wireless Capsule Endoscopy (WCE) has been proposed to enable the visualisation of the whole GI track cable freely. The WCE is a sensor device that contains a colour video camera and wireless radiofrequency transmitter, and battery to take nearly 55,000 colour images during an 8-hour journey through the digestive tract.

The most popular WCEs, are developed and manufactured by Olympus (Olympus, 2010), IntroMedic (IntroMedic, 2010) and Given Imaging (Given Imaging, 2010). However, there are still several drawbacks limiting the application of WCE. Recently, there are two main directions to develop the WCE. One is for enlarging the advantages of current wireless capsule, for example they are trying to make the capsule smaller and smaller, to enhance the propagation efficiency of the antenna or to reduce the radiated effects on human body. While, others are working on minimizing the disadvantages of capsule endoscope, for instance, they use internal and external magnetic field to control the capsule and use technology to reduce the power consumption.

The role of the WCE embedded antenna is for sending out the detected signals; hence the signal transmission efficiency of the antenna will directly decide the quality of received real-time images and the rate of power consumption (proportional to battery life). The human

body as a lossy dielectric material absorbs a number of waves and decreases the power of receiving signals, presenting strong negative effects on the microwave propagation. Therefore, the antenna elements should ideally possess these features: first, the ideal antenna for the wireless capsule endoscope should be less sensitive to human tissue influence; second, the antenna should have enough bandwidth to transmit high resolution images and huge number of data; third, the enhancement of the antenna efficiency would facilitate the battery power saving and high data rate transmission.

In this chapter the WCE system and antenna specifications is first introduced and described. Next, the special consideration of body characteristics for antenna design (in body) is summarized. State-of-the-art WCE transmitting and receiving antennas are also reviewed. Finally, concise statements with a conclusion will summarize the chapter.

2. Wireless Capsule Endoscopy (WCE) system

In May of 2000, a short paper appeared in the journal *Nature* describing a new form of gastrointestinal endoscopy that was performed with a miniaturized, swallowable camera that was able to transmit color, high-fidelity images of the gastrointestinal tract to a portable recording device (Iddan et al., 2000). The newer technology that expands the diagnostic capabilities in the GI tract is capsule endoscopes also known as wireless capsule endoscopy. One example of the capsule is shown in Figure 1.



Fig. 1. Physical layout of the WCE (Olympus, 2010).

The capsule endoscopy system is composed of several key parts (shown in Figure 2): image sensor and lighting, control unit, wireless communication unit, power source, and mechanical actuator. The imaging capsule is pill-shaped and contains these miniaturized elements: a battery, a lens, LEDs and an antenna/transmitter. The physical layout and conceptual diagram of the WCE are depicted in Figure 1 and Figure 2, respectively. The capsule is activated on removal from a holding assembly, which contains a magnet that keeps the capsule inactive until use. When it is used, capsule record images and transmit them to the belt-pack receiver. The capsule continues to record images at a rate over the course of the 7 to 8 hour image acquisition period, yielding a total of approximately 55,000 images per examination. Receiver/Recorder Unit receives and records the images through an antenna array consisting of several leads that connected by wires to the recording unit, worn in standard locations over the abdomen, as dictated by a template for lead placement. The antenna array and battery pack can be worn under regular clothing. The recording device to which the leads are attached is capable of recording the thousands of images

transmitted by the capsule and received by the antenna array. Once the patient has completed the endoscopy examination, the antenna array and image recording device are returned to the health care provider. The recording device is then attached to a specially modified computer workstation (Gavriel, 2000). The software shows the viewer to watch the video at varying rates of speed, to view it in both forward and reverse directions, and to capture and label individual frames as well as brief video clips.

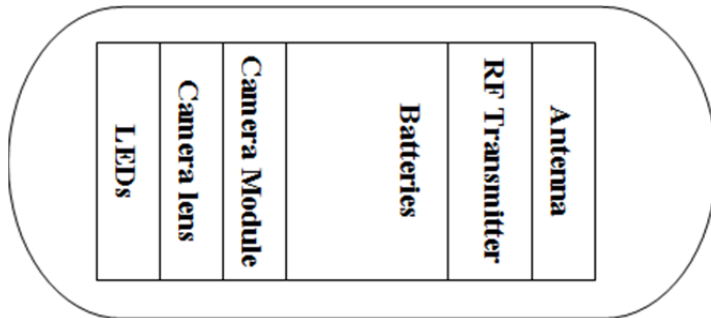


Fig. 2. Conceptual diagram of the WCE.

Since the device received FDA (American Food and Drug Administration) clearance in August 2001, over 1,000,000 examinations have been conducted globally. The 11mm by 26mm M2A capsule is propelled passively, one end of the capsule contains an optical dome with six white Light Emitting Diodes and a CMOS camera that captures 2 images a second (Given Imaging, 2010). These images relayed via a transmitter using a radio frequency signal to an array of aerials from where they are transferred over the wires to a data-recorder. The sensor array allows for continuous triangulation of the position of the capsule inside the body of the patient. The accuracy of the capsule location provided by this method was reported to be ± 3 cm (Ravens & Swain, 2002). In December 2004, FDA approved a second type of capsule developed by Given Imaging—the PillCam ESO, which allows the evaluation of esophageal disease. The response to this demand materialized in the development of the PillCam ESO which has the higher frame rate and CMOS cameras positioned at both ends of the capsule. This capsule acquires and transmits seven frames per second from each camera, giving a total of 14 frames per second (Mishkin et al. 2006). Due to the increased frame rate, the capsule battery life is only 20 minutes. In October 2005, Olympus launched a competitor system called EndoCapsule in Europe. The difference lies in the use of a different imaging technology—CCD, which the manufacturers claim is of higher quality (Fuyono, I. 2005). Another feature of EndoCapsule is the Automatic Brightness Control (ABC), which provides an automatic illumination adjustment as the conditions in the GI tract vary. In October 2006, Given Imaging received the CE Mark to market a third capsule—the PillCam COLON throughout the European Union. This capsule measures 11mm by 31mm, that is slightly larger than previous products. It captures 4 images a second for up to 10 hours. A new feature in Given Imaging capsules is an automatic lighting control (Eliakim et al. 2006; Schoofs et al., 2006). In 2007, PillCam SB2 was cleared for marketing in the US. According to the manufacturers, it offers advanced optics and a wider field of view. PillCam SB2 also captures nearly twice the mucosal area per image. It also provides Automatic Light Control for optimal illumination of each image. In 2009, the

second-generation capsule, PillCam COLON2, was cleared by the European Union. The capsule has the ability to adjust the frame rate in real time to maximize colon tissue coverage. To present, Olympus is working on the development of a new generation capsule endoscope, which features magnetic propulsion. Apart from the novel propulsion and guidance system, the capsule designers aim to provide a drug delivery system, a body fluid sampling system and also the ultrasound scan capability. RF System Lab Company announced the design of the new Sayaka capsule (RF System Lab, 2010), which acquires images at a rate of 30 frames per second and generate about 870,000 over an eight hour period of operation. Also, further applications of magnetic fields are presented (Lenaertes & puers, 2006).

3. Antenna specifications for WCE

Wireless capsule transmitting and receiving antennas belong to wireless communication unit. The transceiver in conjugation with an antenna was utilised. A bidirectional communication between the capsule and the external communication unit at recommended frequency for industrial, scientific and medical usage was established. Wireless capsule endoscopy transmitting antenna is for sending out the detected signal and receiving antenna receive the signal outside human body. The signal transmission efficiency of the antenna will directly decide the quality of the received real-images and rate of power consumption. Because a lossy dielectric material absorbs a number of waves and decreases the power of receiving signal, it presents strong negative effects on the microwave propagation (Johnson, & Guy, 1972). Therefore, some features to ideally possess are required. The WCE antenna should be less sensitive to human tissue influence. Enough bandwidth to transmit high resolution images and huge number of data is a requirement for antenna. Also, power saving and high data rate transmission can be obtained with enhancement of antenna efficiency.

In addition to the standard constraints in electronic design, a number of main challenges arise for systems that operate inside the human body. The size of the capsule endoscope system should be small because small-sized capsules are easier to swallow. Therefore, the foremost challenge is miniaturization to obtain an ingestible device (the volume should be smaller than endoscopy). The availability of small-scale devices can place severe constraints on a design, and the interconnection between them must be optimized. The size constraints lead to another challenge, noise. The coexistence of digital integrated circuits, switching converters for the power supply, and communication circuits in close vicinity of the analog signal conditioning could result in a high level of noise affecting the input signal. Therefore, capsule designers must take great care when selecting and placing components, to optimize the isolation of the front end.

The next vital challenge is to reduce power consumption. In particular, the generated wireless signal must not interfere with standard hospital equipment but still be sufficiently robust to overcome external interferences. On the basis of Friis's formula, the total loss between transmitter and receiver increases with the distance between the transmitting and the receiving antennas increasing. As the result of the dispersive properties of human body materials, the transmitting power absorbed by body varies according to the antenna's operating frequency. The radiated field intensity inside and outside the torso or gut area is determined for FCC regulated medical and Industrial Scientific Medical (ISM) bands,

including the 402MHz to 405MHz for Medical Implant Communications Service (MICS), 608MHz to 614 MHz for Wireless Medical Telemetry Service (WMTS), and the 902MHz-928MHz ISM frequency band. Moreover FCC has allocated new bands at higher frequencies such as 1395MHz-1400 MHz wireless medical telemetry services (WMTS) band. Carefully selection of target frequency is important during the antenna design.

The effective data rate was estimated to be about 500 Kbps (Rasouli et al. 2010). The transmit power must be low enough to minimize interference with users of the same band while being strong enough to ensure a reliable link with the receiver module. Lower frequencies are used for ultrasound (100 kHz to 5 MHz) and inductive coupling (125 kHz to 20 MHz). The human body is no place for operational obscurity, so the control software must enforce specific rules to ensure that all devices operate as expected. For that reason, key programs must be developed in a low-level (often assembly) language. The last challenge concern encapsulating the circuitry in appropriate biocompatible materials is to protect the patient from potentially harmful substances and to protect the device from the GI's hostile environment. The encapsulation of contactless sensors (image, temperature, and so on) is relatively simple compared to the packaging of chemical sensors that need direct access to the GI fluids. Obtaining FDA (Food and Drug Administration) approval for the US market or CE (European Conformity) marking in Europe involves additional requirements. Capsules must undergo extensive material-toxicity and reliability tests to ensure that ingesting them causes no harm. The maximal data rate of this transmitter is limited by the RC time constant of the Rdata resistor and the capacitance seen at the base. It is clear that formal frequency higher than $1/(R_{data} \cdot C_{base})$, the modulation index decreases, because the injected base current is shorted in the base capacitance. Although the occupied bandwidth decreases, the S/N ratio decreases too, and robust demodulation becomes more difficult at faster modulation rates. From experiments, the limit was found to be at 2Mbps [22]. Considering the sensitivity of small receivers for biotelemetry, the designed antenna should have a gain that exceeds -20 dB (Chi et al. 2007; Zhou et al. 2009).

4. Special consideration of body characteristics for antenna design

The antenna designed for biomedical telemetry is based on the study of the materials and the propagation characteristics in the body. Because of the different environment, the wave radio propagation becomes different in free space. The human body consists of many tissues with different permittivity and conductivity, which leads to different dielectric properties.

The same radio wave propagating through different media may exhibit different features. From an electromagnetic point of view, materials can be classified as conductive, semi conductive or dielectric media. The electromagnetic properties of materials are normally functions of the frequency, so are the propagation characteristic. Loss tangent defined as the ratio of the imaginary to the real parts of the permittivity, which is equation (Kraus & Fleisch, 1999).

$$\tan\delta = \frac{\sigma}{\omega\epsilon} \quad (1)$$

With the specific classification are given in (Kraus & Fleisch, 1999), the body material is dielectric material. The loss tangent is just a term in the bracket. The attenuation constant is

actually proportional to the frequency if the loss tangent is fixed; where the attenuation constant is

$$a = \omega\sqrt{\mu\epsilon} \left[\frac{1}{2} \left(\sqrt{1 + \frac{\sigma^2}{\epsilon^2\omega^2}} - 1 \right) \right]^{1/2}. \quad (2)$$

The dominant feature of radio wave propagation in media is that the attenuation increases with the frequency. With the formula

$$\gamma = \sqrt{j\omega\mu(\sigma + j\omega\epsilon)}, \quad (3)$$

$$E = E_0 e^{j\omega t - \gamma z}, \quad (4)$$

$$H = \frac{j}{\omega\mu} \nabla \times E, \quad (5)$$

It can find out that the power of E plane and H plane reduce with high dielectric constant and conductivity. The total power is consumed easily in human body. The efficiency of antenna becomes lower than free space. With the formula

$$v = \frac{1}{\sqrt{\mu\epsilon}} \text{ and } \beta = \omega\sqrt{\mu\epsilon} \left[\frac{1}{2} \left(\sqrt{1 + \frac{\sigma^2}{\epsilon^2\omega^2}} + 1 \right) \right]^{1/2} = \frac{2\pi}{\lambda}, \quad (6)$$

in a high dielectric material, the electrical length of the antenna is elongated. Compare dipole antenna in the air and in the body material, they have same physical length but electrical lengths are not same. Because of the high permittivity, the antenna in the body material has longer electrical length. The time-averaged power density of an EM wave is

$$S_{av} = \frac{1}{2} \sqrt{\frac{\epsilon}{\mu}} E_0^2, \quad (7)$$

which leads to high power density in human body. The intrinsic impedance of the material and is determined by ratio of the electric field to the magnetic field (Huang & Boyle, 2008).

$$\eta = \sqrt{\frac{j\omega\mu}{\sigma + j\omega\epsilon}}. \quad (8)$$

Based on wave equation $\nabla^2 E - \gamma^2 E = 0$, A and B in the wave propagating trigonometric form $E = xA\cos(\omega t - \beta z) + yB\sin(\omega t - \beta z)$ can be determined. With the relationship of A and B, it can confirm shape of polarization.

The multi-layered human body characteristic can be simplified as one equivalent layer with dielectric constant of 56 and the conductivity of 0.8 (Kim & Rahmat-Samii, 2004). So, with

the change from free space to body materials, dielectric constant changes from 1 to 56 and conductivity changes from 0 to 0.8. What's more, to detect the transmitted signal independent of transmitter a position, the antenna is required the omni-directional radiation pattern (Kim & Rahmat-Samii, 2004; Chirwa et al., 2003). To investigate the characteristics of antennas for capsule endoscope, the human body is considered as an averaged homogeneous medium as described by the Federal Communications Commission (FCC) and measured using a human phantom (Kwak et al., 2005; Haga et al., 2009).

5. State-of-the-art WCE transmitting and receiving antennas

An antenna plays a very crucial role in WCE systems. Wireless capsule transmitting and receiving antennas belong to wireless communication unit, which provides a bidirectional communication between the capsule and the external communication unit at recommended frequency at which industrial, scientific and medical band was established. Wireless capsule endoscopy transmitting antenna is for sending out the detected signal and receiving antenna receive the signal outside human body. This section is to discuss the current performance of both WCE transmitting and receiving antennas.

5.1 Transmitting antennas

The capsule camera system is shown in Figure 2. One of the key challenges for ingestible devices is to find an efficient way to achieve RF signal transmission with minimum power consumption. This requires the use of an ultra-low power transmitter with a miniaturized antenna that is optimized for signal transmission through the body. The design of an antenna for such a system is a challenging task (Norris et al., 2007). The design must fulfill several requirements to be an effective capsule antenna, including: miniaturization to achieve matching at the desired bio-telemetric frequency; omni-directional pattern very congruent to that of a dipole in order to provide transmission regardless of the location of the capsule or receiver; polarization diversity that enables the capsule to transmit efficiently regardless of its orientation in the body; easy and understandable tuning adjustment to compensate for body effects. Types of transmitting antenna are used such as the spiral antennas, the printed microstrip antennas, and conformal antennas as shown in following subsections.

5.1.1 Spiral antennas

A research group from Yonsei University, South Korea, proposed a series of spiral and helical antennas providing ultra-wide bandwidth at hundreds of megahertz.

Single arm spiral antenna

The first design is a miniaturized normal mode helical antenna with the conical structure (Kwak et al., 2005). To encase in the capsule module, the conical helical antenna is reduced only in height with the maintenance of the ultra-wide band characteristics. Thus, the spiral shaped antenna is designed with the total spiral arm length of a quarter-wavelength. The configuration of the designed antenna is shown in Figure 3(a). It is composed of a radiator and probe feeding structure. The proposed antenna is fabricated on the substrate with 0.5-oz copper, 3 mm substrate height, and dielectric constant of 2.17. The diameter of the antenna is 10.5 mm and 0.5 mm width conductor.

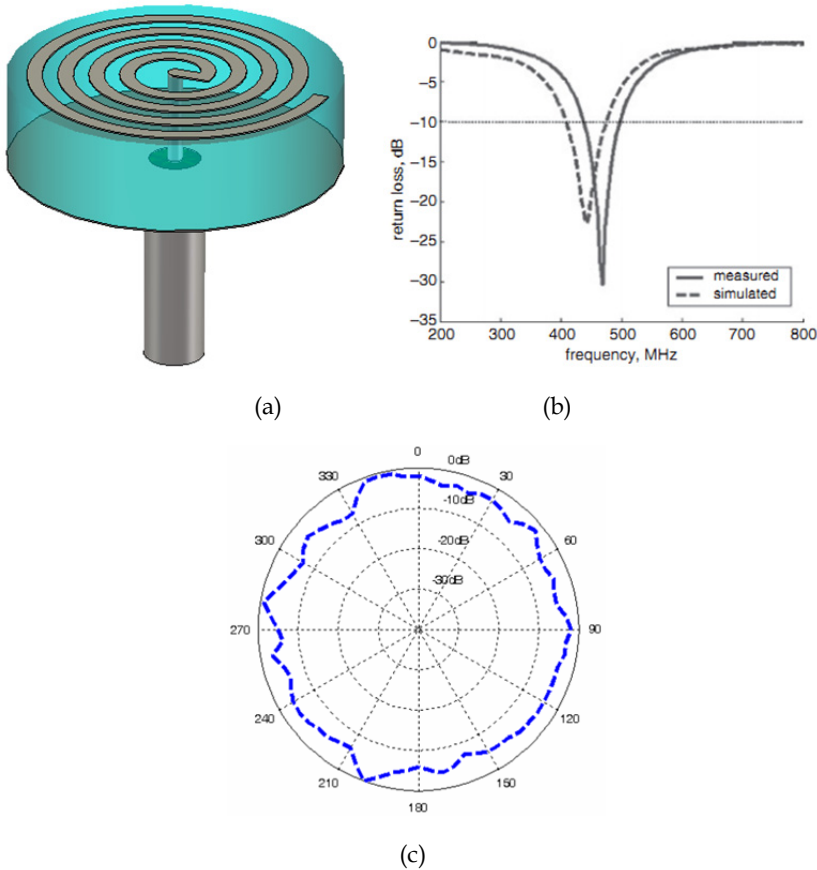


Fig. 3. Single arm spiral antenna (Kwak et al. 2005): (a) the geometric structure; (b) simulated and measured return losses; (c) azimuth pattern at 430MHz.

The simulated and the measured return losses of the antenna surrounded by human body equivalent material are shown in Figure 3(b). It can be observed that the bandwidth of the proposed spiral shaped antenna for $S_{11} < -10\text{dB}$ is 110 MHz of 400-510 MHz and the fractional bandwidth is 24.1 %, which is larger than 20%, the reference of the UWB fractional bandwidth. The measurement result of the azimuth radiation pattern is shown in Figure 3(c). The normalized received power level is varying between 0dB to -7dB, which can be considered as an omni-directional radiation pattern.

Dual arm spiral antenna

The dispersive properties of human body suggested that signals are less vulnerable when they are transmitted at lower frequency range. Therefore, a modified design is proposed to provide ultra-wide bandwidth at lower frequency range (Lee et al. 2007). Figure 4(a) shows the geometry of a dual spiral antenna. The newly proposed antenna is composed of two spirals connected by the single feeding line. The radius of designed antenna is 10.1mm and

its height is about 3.5mm. To design a dual spiral antenna, two substrate layers are used. The upper and lower substrate layers have the same dielectric constant of 3.5 and the thicknesses of them are both 1.524mm. Two spirals with the same width of 0.5mm and the same gap of 0.25mm have different overall length. The lower spiral antenna is a 5.25 turn structure and the upper spiral is 5 turns.

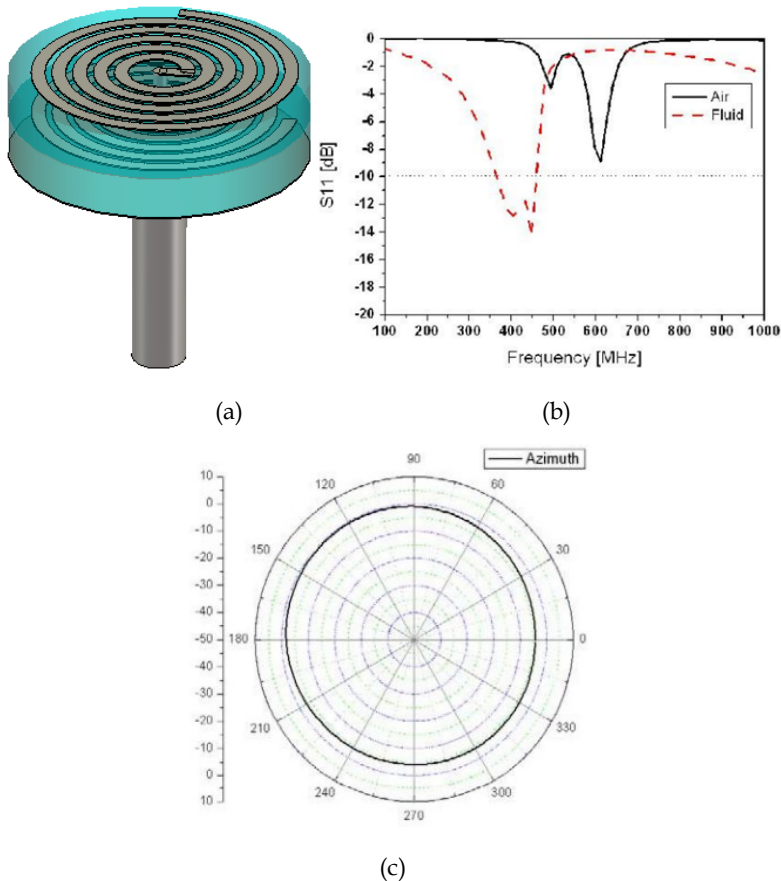


Fig. 4. Dual arm spiral antenna (Lee et al. 2007): (a) the geometric structure; (b) measured return losses; (c) azimuth pattern at 400MHz.

The return loss of the proposed antenna was measured in the air and in the simulating fluid of the human tissue as shown in Figure 4(b). Because of considering electrical properties of equivalent material of human body, return loss characteristic in the air is not good but dual resonant characteristic is shown in the air. However, the proposed antenna has low return loss value at operating frequency in the fluid and its bandwidth is 98MHz (from 360MHz to 458MHz) in the fluid, with the fractional bandwidth of about 25%. The simulated radiation pattern as shown in Figure 4(c) is omni-directional at the azimuth plane with 5dB variation.

Conical helix antenna

Extensive studies of the helical and spiral antennas were conducted with modified geometric structures. For example, a conical helix antenna fed through a 50 ohm coaxial cable is shown in Figure 5. Compared to small spiral antenna, conical spiral takes up much space. However, additional space is not necessary because a conical spiral can use the end space of the capsule as shown in Figure 5(a). The radius of the designed antenna is 10mm and the total height is 5 mm. This size is enough to be encased in small capsule.

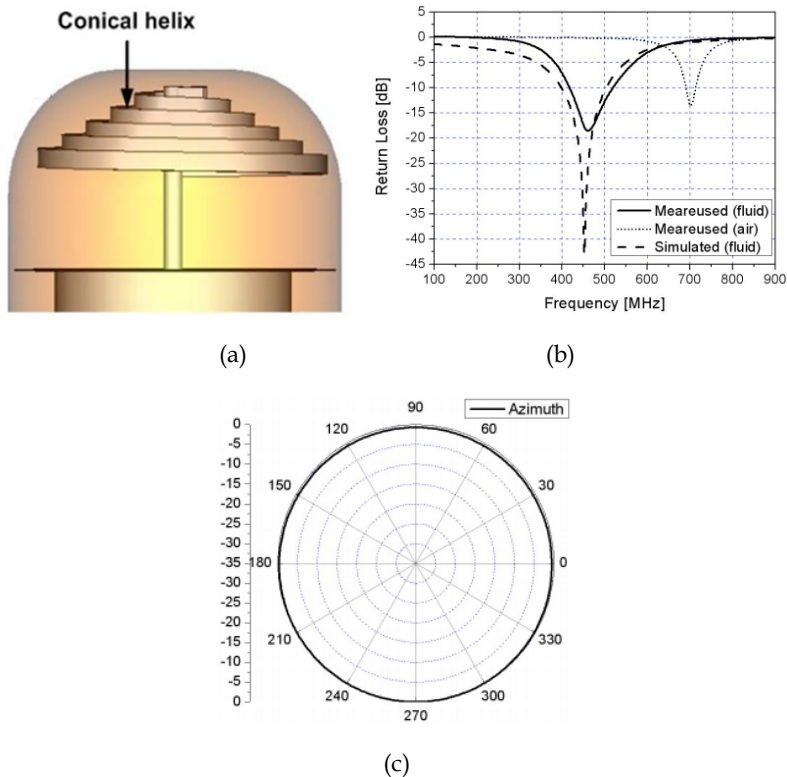


Fig. 5. Conical helix antenna (Lee et al. 2008): (a) the geometric structure; (b) simulated and measured return losses; (c) azimuth pattern at 450MHz.

The proposed antenna provides a bandwidth of 101MHz (from 418MHz to 519MHz) in the human body equivalent material as shown in Figure 5(b). Its center frequency is 450MHz, so the fractional bandwidth is about 22%. The normalized simulated radiation pattern is shown in Figure 5(c). The proposed antenna has omni-directional radiation pattern with less than 1dB variation.

Fat arm spiral antenna

Another modified design is the fat arm spiral antenna as shown in Figure 6(a). The spiral arm is 3mm wide and separated from ground plane with a 1mm air gap. The antenna is

simulationally investigated in the air, in the air with capsule shell and in the human body equivalent material.

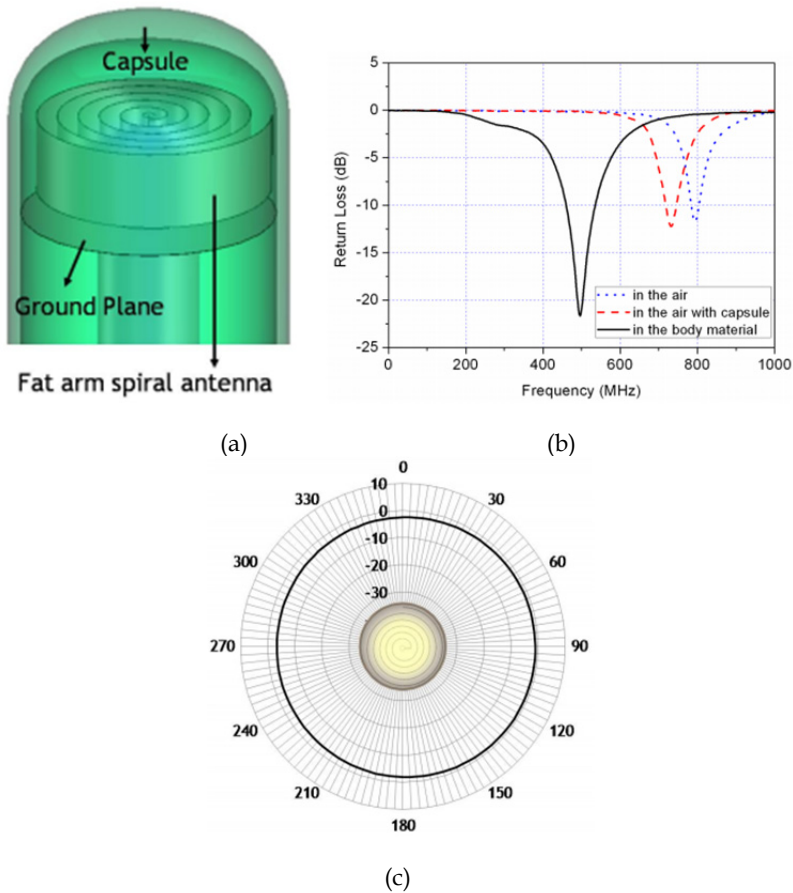


Fig. 6. Fat arm spiral antenna (Lee et al. 2010): (a) the geometric structure; (b) return losses; (c) azimuth pattern at 450MHz.

The return losses of the antenna in free space, with dielectric capsule shell and in the liquid tissue phantom are plotted in Figure 6(b). The resonant frequency is observed about 800 MHz in the air, and reduced to 730 MHz due to the capsule effects on the effective dielectric constant and matching characteristic. When the proposed antenna is emerged in the equivalent liquid, it shows good matching at a resonant frequency and its bandwidth is 75 MHz (460 ~ 535 MHz) for S_{11} less than -10dB. The radiation pattern illustrated in Figure 6(c) presents that this antenna also provides omni-directional feature at azimuth plane.

Square microstrip loop antenna

A square microstrip loop antenna (Shirvante et al. 2010) is designed to operate on the Medical Implant Communication Service (MICS) band (402MHz -405MHz). The antenna is

patterned on a Duroid 5880 substrate with a relative permittivity ϵ_r of 2.2 and a thickness of $500\mu\text{m}$ as shown in Figure 7(a). The area of the antenna is approximately 25 mm^2 which is smaller enough to be encased in a swallowable capsule for children.

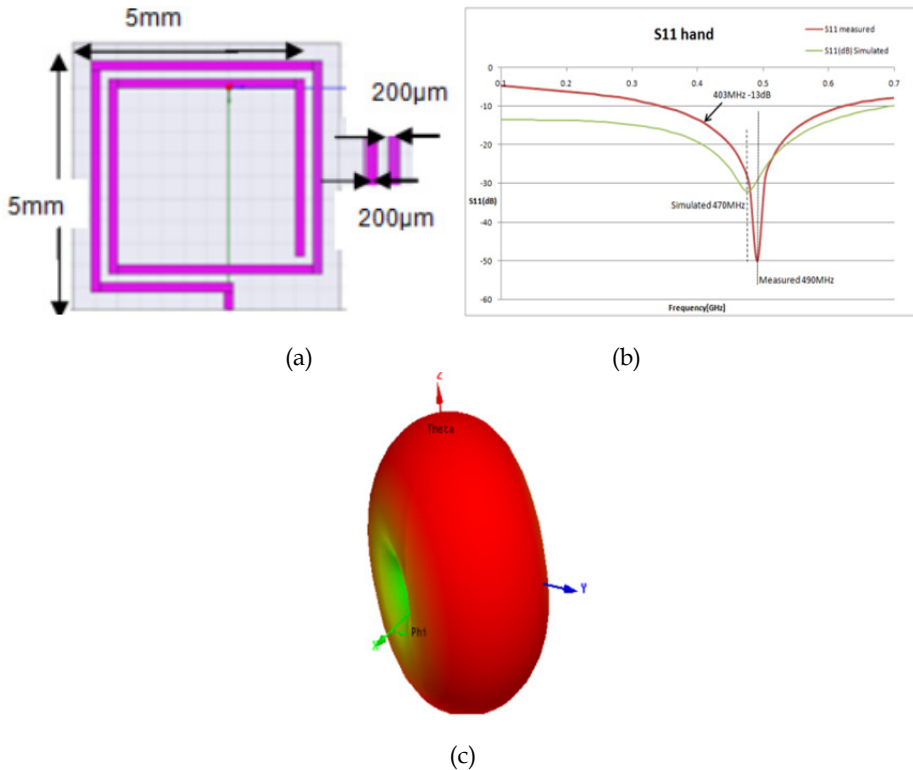


Fig. 7. square microstrip loop antenna (Shirvante et al. 2010): (a) the geometric structure; (b) simulated and measured return losses; (c) azimuth pattern at 403MHz.

The simulated and measured return losses as shown in Figure 7(b) presents that the antenna provides enough bandwidth to cover the 402MHz to 405MHz band. At the FSK operating frequency 403MHz, the measured return loss is -13dB. Moreover, the designed antenna shows a large tolerance to impedance variation at the MICS band, in correspondance to ϵ_r variation. The designed antenna also has an omni-directional radiation pattern at azimuth plane.

5.1.2 Conformal antennas

A conformal geometry exploits the surface of the capsule and leaves the interior open for electrical components including the camera system. Several designs made efficient usage of the capsule shell area are selected as examples and introduced in this subsection.

Conformal chandelier meandered dipole antenna

The conformal chandelier meandered dipole antenna is investigated as a suitable candidate for wireless capsule endoscopy (Izdebski et al., 2009). The uniqueness of the design is its

miniaturization process, conformal structure, polarization diversity, dipole-like omnidirectional pattern and simple tunable parameters (as shown in Figure 8(a)). The antenna is offset fed in such a way that there is an additional series resonance excited in addition to the parallel resonance (as shown in Figure 8(b)). The two arms with different lengths generate the dual resonances. This additional series resonance provides better matching at the frequency of interest. This antenna is designed to operate around 1395MHz - 1400 MHz wireless medical telemetry services (WMTS) band.

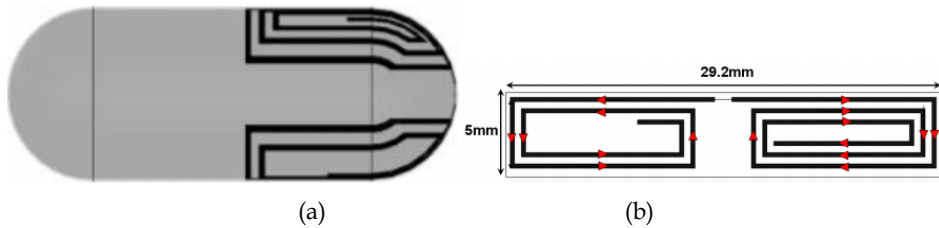


Fig. 8. Conformal chandelier meandered dipole antenna (Izdebski et al., 2009): (a) the geometric structure of the conformal chandelier meandered dipole antenna; (b) Offset Planar Meandered Dipole Antenna with current alignment vectors.

The offset planar meandered dipole antenna is simulated on a 0.127 mm thick substrate with a dielectric constant of 2.2. The antenna is placed in the small intestine and it is observed that there is a lot of detuning due to the body conductivity and the dielectric constant (average body composition has a relative permittivity of 58.8 and a conductivity of 0.84S/m). The series resonance shifts closer to 600 MHz. The antenna is then retuned to the operational frequency of 1.4 GHz by reducing the length of the dipole antenna. The return losses of both the detuned and tuned antenna are shown Figure 9(a). Figure 9(b) shows the radiation pattern of the tuned antenna inside the human body at 1.4 GHz.

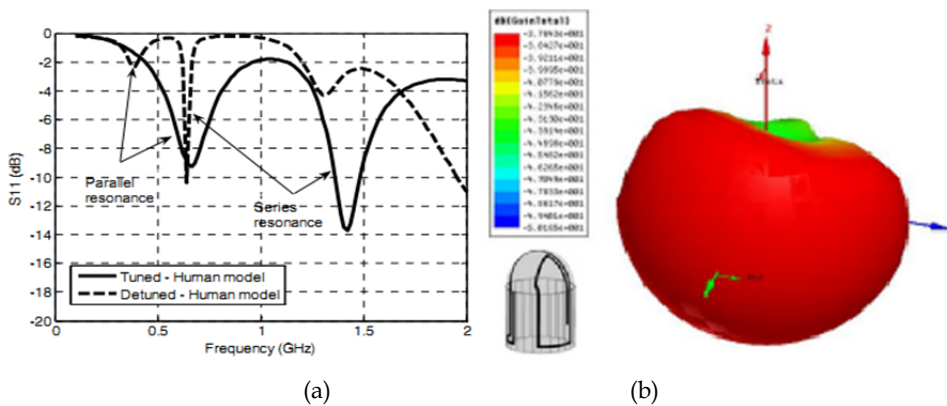


Fig. 9. Conformal chandelier meandered dipole antenna (Izdebski et al., 2009): (a) the return losses of detuned and tuned structure in human model; (b) azimuth pattern at 1.4GHz.

The radiation pattern is dipole-like but tilted due to the conformity of the structure. The axial ratio (dB) for the conformal chandelier meandered dipole antenna is about 7dB

(elliptical polarization). It possesses all the characteristics of planar structure along with polarization diversity.

Outer-wall loop antenna

The proposed outer-wall loop antenna (Yun et al., 2010.) makes maximal use of the capsule's outer surface, enabling the antenna to be larger than inner antennas. As shown in Figure 10(a), the antenna is part of the outer wall of the capsule, thus decreasing volume and increasing performance, and uses a meandered line for resonance in an electrically small area. The capsule shell with the relative permittivity of 3.15 has the outer and the inner radius of the capsule as 5.5mm and 5mm, respectively. Its length is 24 mm. The height of the meander line and gap between meander patterns are set to 7mm and 2.8mm, respectively. The opposite side of the loop line is meander patterns in the same way. Although capsule size is reduced, the radius of sphere enclosing the entire structure of the antenna is increased.

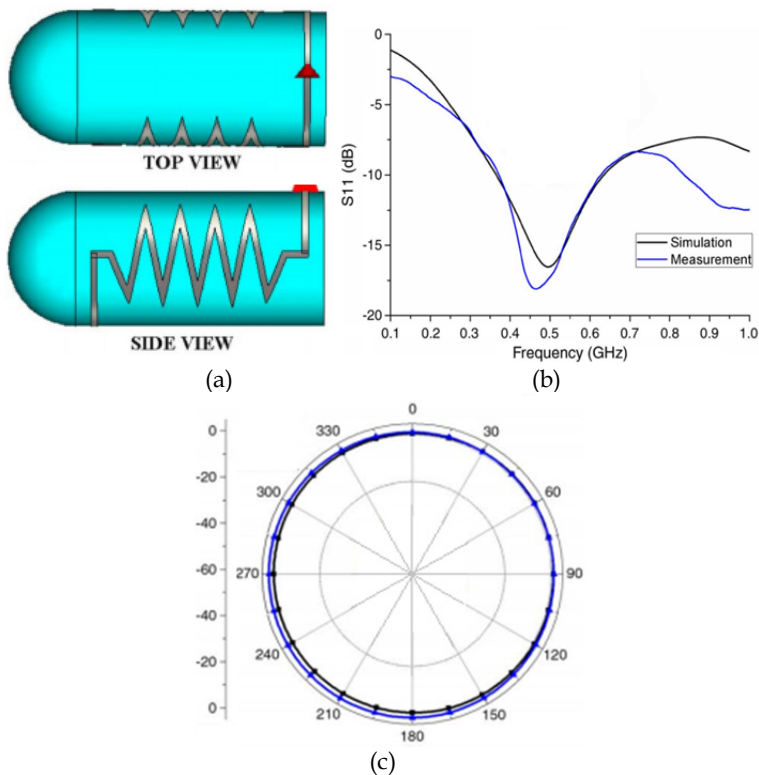


Fig. 10. Outer-wall loop antenna (Yun et al., 2010.): (a) the geometric structure; (b) simulated and measured return losses; (c) azimuth pattern at 500MHz.

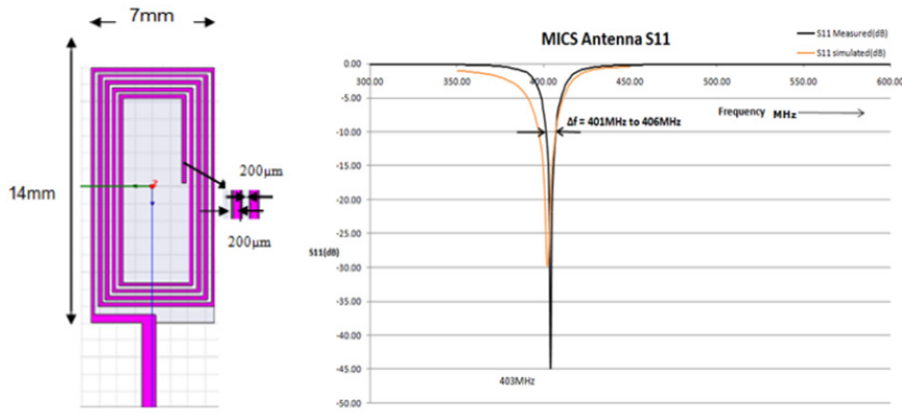
Figure 10(b) shows that the proposed antenna has an ultra wide bandwidth of 260 MHz (from 370MHz to 630 MHz) for $VSWR < 2$ and an omnidirectional radiation pattern at azimuth plane (as shown in Figure 10(c)). Using identical antenna pairs in the equivalent body phantom fluid, antenna efficiency is measured to 43.7% (3.6 dB).

5.2 Receiving antennas

The receiving antennas are operating outside of human body, which is no longer limited by its size. Therefore, the design of receiving antennas is less challenge than the design of transmitting antennas. In this subsection, several types of receiving antenna are selected as examples.

Narrow bandwidth antenna for receiver

A narrow bandwidth receiving antenna is designed using microstrip loop structure (Shirvante et al. 2010). The antenna is patterned using a milling machine on a Duroid 5880 substrate with a relative permittivity ϵ_r of 2.2 and a thickness of 500 μm as shown in Figure 11(a). The overall length of the wire is approximately a quarter wavelengths: $\lambda_{air} / 4 = 187\text{mm}$ at 402MHz for air medium.



(a)

(b)

(c)

Fig. 11. Rectangular microstrip loop antenna (Shirvante et al. 2010): (a) the geometric structure; (b) simulated and measured return losses; (c) azimuth pattern at 403MHz.

Figure 11(b) shows the simulated and measured return losses of the proposed antenna. The return loss shows a deep null of -30dB at 403MHz . The directional radiation pattern as shown in Figure 11(c) provides the possibility to aim the receiver to human body area, where the transmitter sends signals from. Therefore, for narrow bandwidth applications, such as the ASK or FSK modulation, the line loop antenna is a good choice.

Miniaturized microstrip planar antenna

To accommodate the antenna in a small communication unit, a meander line style structure is used (Babar et al., 2009). The antenna's radiating part is shorted with the ground plane, to further decrease the size of the antenna structure. The reduction of the size of the antenna by shortening also reduces the gain of the antenna, as decreasing the size of the antenna more than its wavelength affects the efficiency of the antenna.

The antenna was fabricated on a double sided copper FR4 - printed circuit board, with 1.6mm thickness as shown in Figure 12(a). The excitation is given through an SMA connector from the opposite direction of the PCB to the antenna structure. The total size of the antenna structure is $20\text{mm} \times 37\text{mm}$. There is no ground plane present on the opposite side of the PCB, where the antenna structure is present, which helps in getting an omnidirectional radiation pattern.

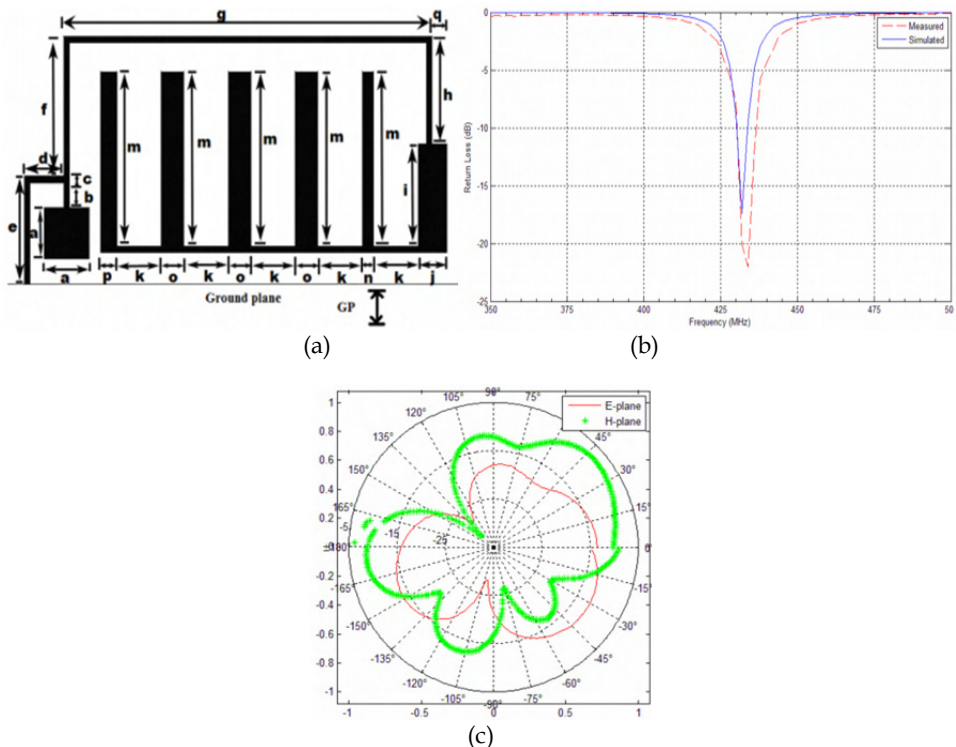


Fig. 12. Microstrip planar antenna (Babar et al. 2009): (a) the geometric structure; (b) simulated and measured return losses; (c) radiation patterns at 433MHz .

Figure 12(b) presents that the operating frequency of the antenna is 433 MHz with the bandwidth of 4MHz. Figure 12(c) shows the radiation pattern of the antenna's E and H-plane. The achieved max gain from the antenna was around -6.1 dBi.

Receiver antenna with buffer layer

The dual pentagon loop antenna having circularly polarization is proposed (Park, S. et al., 2008). The configuration of the proposed dual pentagon loop antenna is shown in Figure 13(a). The proposed antenna and the feeding structure were etched on the front and the back of a substrate (Figure 13(b)). And a-a' are b-b' are shorted as follows. The proposed antenna was designed a dual loop type to enhanced H-field since the current direction of each of loops is different. And there is a gap on each of loops to make a CP wave (Morishita & Hirasawa 1994; Sumi et al., 2004 as cited in Park, S. et al., 2008). The strip widths of the primary loop and of the CPW are 0.80 mm; the used substrate is R/flex 3850; $L_1 = 12.93$ mm, $L_2 = 10.97$ mm, $L_3 = 10.21$ mm, $G = 0.49$ mm, $S_1 = 26.01$ mm, $S_2 = 1.65$ mm, $W_1 = 5.80$ mm, $W_2 = 1.70$ mm. The CPW feeding line on the back of substrate is used to efficiently excite balanced signal power which makes to have a broadband.

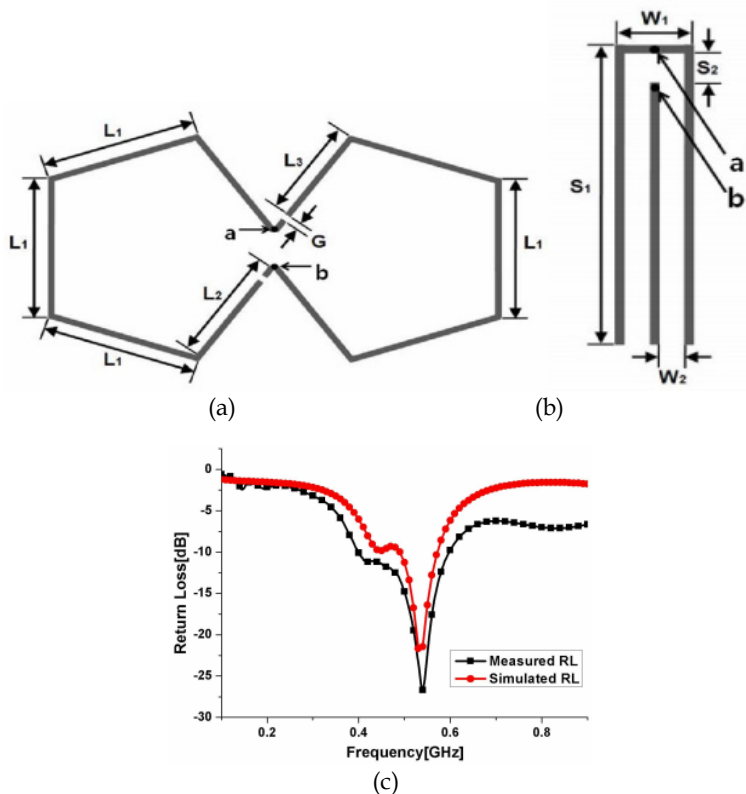


Fig. 13. Receiver antenna with buffer layer (Park, S. et al., 2008): (a) the pentagon dual loop antenna; (b) feeding structure; (c) simulated and measured return losses.

Figure 13(c) presents that the bandwidth of the receiver antenna is from 400 MHz to 600 MHz for $VSWR \leq 2$. As a wave in air meets a medium of which relative permittivity is very high over air, much reflection is inevitably generated. So we designed the buffer layer having ϵ_r between air and human body for reducing the reflection, artificially. The buffer layer which is added a little bit loss is attached on the back of the proposed antenna for reducing a size of antenna and back lobe power.

6. Conclusions

Because of the requirement of medical test for GI tract, WCE came to the world. It solves many restrictions on exploring GI tract. With the development from 2001, WCE has become a promising device with suitable requirement. It has image sensor and lighting, control unit, wireless communication unit, power source, and mechanical actuator. The system can be operated outside the human body, the size of the capsule endoscope system is smaller, and the interconnection between devices was optimized, power consumption also reduced with technology optimized. Some companies and individual are still studying on new functions and optimization.

For wireless capsule endoscopy antenna, several basic standards and situation of operation in human body were discussed. The signal transmission efficiency of the antenna will directly decide the quality of the received real-images and rate of power consumption. Because of the lossy material absorbs a number of waves and decreasing the power of receiving signal, human body presenting strong negative effects on the microwave propagation. Wireless capsule endoscopy transmitting antenna is for sending out the detected signal inside human body and receiving antenna receive the signal outside human body. Several transmitting antennas are introduced in this article. The two fundamental types of transmitting antenna are the spiral antennas and conformal antennas both feature as the small physical size, relatively large bandwidth, omni-directional pattern and polarization diversity. The receiving antennas operating outside of human body are also discussed, such as the narrow bandwidth antenna for receiver, microstrip meandered planar antenna and the receiver antenna with buffer layer. All of them operate well outside the human body.

7. Acknowledgement

This work is supported by the Natural Science Foundation of Jiangsu province (No. BK2010251 and BK2011352), Suzhou Science and Technology Bureau (No. SYG201011), and XJTLU Research Develop Fund (No. 10-03-16.).

8. References

- American Cancer Society, (2010). Key statistics about cancers, *Official website of American Cancer Society*, access at Oct. 1st, 2010. <<http://www.cancer.org/Cancer/index>>
- Babar, A. et al., (2009). Miniaturized 433 MHz antenna for card size wireless systems, *Proceeding of Antennas and Propagation Society International Symposium (APSURSI), 2009 IEEE*, Charleston, June, 2009.

- Chi, B. et al., (2007). Low-power transceiver analog front-end circuits for bidirectional high data rate wireless telemetry in medical endoscopy applications, *IEEE Trans. Biomed. Eng.*, Vol. 54, No. 7, 2007, pp. 1291-1299.
- Chirwa, L.C. et al., (2003). Radiation from ingested wireless devices in biomedical telemetry, *Electronic Letters*, Vol.39, No.2, 2003, pp.178-179.
- Eliakim, R. et al., (2006). Evaluation of the PillCam Colon capsule in the detection of colonic pathology: results of the first multicenter, prospective, comparative study. *Endoscopy* 2006, Vol.38, No.10, 2006, pp. 963-970.
- Fuyono, I., (2005). Olympus finds market rival hard to swallow, *Nature*, Vol. 438, 2005, p.913.
- Gavriel, D. M., (2000). The development of the swallowable video capsule (M2A), *Gastrointestinal Endoscopy*, Vol. 52, No. 6, 2000, pp. 817-819.
- Given Imaging, (2010). Overview of product, *Official website of Given Imaging*, access at Sep. 30th, 2011.
<<http://www.givenimaging.com/en-int/HealthCareProfessionals/Pages/pageHCP.aspx>>
- Haga, N. et al., (2009). Characteristics of cavity slot antenna for body-area networks, *IEEE Trans. Antennas Propag.*, Vol. 57, No. 4, 2009, pp. 837-843.
- Huang, Y. & Boyle, K., (2008). Radio Wave Propagation Characteristic in Media, *Antennas from Theory to Practice*, pp.93-95.
- Iddan, G. G. et al., (2000). Wireless capsule endoscopy, *Nature*, Vol. 405, 2000, pp. 417-418.
- IntroMedic, (2010). MicroCam Info, *Official website of IntroMedic*, access at Sep. 30th, 2011.
<<http://www.intromedic.com/en/product/productInfo.asp>>
- Izdebski, P. et al., (2009). Ingestible Capsule Antenna for Bio-Telemetry, *Proceeding of IEEE International Workshop on Antenna Technology (iWAT) 2009*, Santa Monica, March, 2009.
- Johnson, C. C. & Guy, A. W., (1972). Nonionizing electromagnetic wave effects in biological materials and systems, *Proceeding of IEEE*, Vol. 60, No. 6, 1972, pp.692-720.
- Kim, J. & Rahmat-Samii, Y., (2004). Implanted antennas inside a human body: simulations, designs and characterizations, *IEEE transaction of Microwave theory and techniques*, August, Vol. 52. No. 8, 2004, pp. 1934-1943.
- Kraus, J. D. & Fleisch, D. A., (1999). *Electromagnetics with Application*, 5th edition, McGraw-Hill, 1999.
- Kwak, S. I. et al., (2005). Ultra-wide band spiral shaped small antenna for the biomedical telemetry, *2005 Asia-Pacific Conference Proceedings (APMC)*, Suzhou, December, 2005.
- Lee, S. H. et al., (2007). A dual spiral antenna for wideband capsule endoscope system, *2007 Asia-Pacific Conference Proceedings (APMC)*, Bangkok, December, 2007.
- Lee, S. H. et al., (2008). A conical spiral antenna for wideband capsule endoscope system, *Proceeding of Antennas and Propagation Society International Symposium (AP-S) 2008*, San Diego, June, 2008.
- Lee, S. H. et al., (2010). Fat arm spiral antenna for wideband capsule endoscope systems, *Proceeding of Radio and Wireless Symposium (RWS) 2010*, New Orleans, LA, January, 2010.
- Lenaertes, B. & Puers, R., (2006). An omnidirectional transcutaneous power link for capsule endoscopy, in *Proceedings of International Workshop on Wearable and Implantable Body Sensor Networks*, 2006, pp.46-49.

- Mishkin, D. S. et al., (2006). ASGE Technology Status Report, Wireless Capsule Endoscopy, *Gastrointestinal Endoscopy*, Vol. 63, No. 4, 2006, pp. 539-545.
- Morishita, H. & Hirasawa, K., (1994). Wideband circularly-polarized loop antenna, *Proceeding of Antennas and Propagation Society International Symposium (AP-S) 1994*, Seattle, 1994.
- Norris, M. et al., (2007). Sub miniature antenna design for wireless implants, *Proceedings of the IET Seminar on Antennas and Propagation for Body-Centric Wireless Communication*, London, 2007.
- Olympas, (2010). EndoCapsule - Taking capsule endoscopy to next level, *Official website of Olympus*, access at Sep. 30th, 2011.
< http://www.olympus-europa.com/endoscopy/2001_5491.htm>
- Park, S. et al., (2008). A New Receiver Antenna with Buffer Layer for Wireless Capsule Endoscopy in human body, *Proceeding of Antennas and Propagation Society International Symposium (AP-S) 2008*, San Diego, June, 2008.
- Rasouli, M. et al., (2010). Wireless Capsule Endoscopes for Enhanced Diagnostic Inspection of Gastrointestinal Tract, *Proceeding of Robotics Automation and Mechatronics (RAM) 2010*, Singapore, June, 2010.
- Ravens, A. F. & Swain, P., (2002). The wireless capsule: new light in the darkness, *Digestive Diseases*, Vol. 20, No. 2, 2002, pp.127-133.
- RF System Lab, (2010), The next generation of capsule endoscopy - Sayaka, *Official website of RF System Lab*, access at Sep. 30th, 2011,
< <http://www.rfamerica.com/sayaka/index.html>>
- Schoofs, N. et al., (2006). PillCam colon capsule endoscopy compared with colonoscopy for colorectal tumor diagnosis: a prospective pilot study. *Endoscopy 2006*, Vol. 38, No.10, 2006, pp. 971-977.
- Shirvante, V. et al., (2010). Compact spiral antennas for MICS band wireless endoscope toward pediatric applications, *Proceeding of Antennas and Propagation Society International Symposium (APSURSI), 2010 IEEE*, Toronto, July, 2010.
- Sumi, M. et al., (2004). Two rectangular loops fed in series for broad-band circular polarization and impedance matching, *IEEE Transaction on Antennas and Propagation*, Vol. 52, No. 2, pp. 551-554, 2004.
- Yu, X. et al., (2006). Microstrip antennas for the wireless capsule endoscope system, *Patent CN 1851982A*, October. 2006.
- Yun, S. et al., (2010). Outer-Wall Loop Antenna for Ultrawideband Capsule Endoscope System, *IEEE Antennas and Wireless Propagation Letters*, Vol. 9, pp.1135-1138, 2010.
- Zhou, Y. et al., (2009). A wideband OOK receiver for wireless capsule endoscope, *European Microwave Conference 2009*, Rome, October, 2009.

Travelling Planar Wave Antenna for Wireless Communications

Onofrio Losito and Vincenzo Dimiccoli
Itel Telecomunicazioni srl, Ruvo di Puglia (BA)
Italy

1. Introduction

Microstrip antennas are one of the most widely used types of antennas in the microwave frequency range, and they are often used in the millimeter-wave frequency range. Actually as the demand for high data rates grows and microwave frequency bands become congested, the millimeter-wave spectrum is becoming increasingly attractive for emerging wireless applications. The abundance of bandwidth and large propagation losses at millimeter-wave frequencies makes these bands best-suited for short-range or localized systems that provide broad bandwidth. Automotive radar systems including cruise control, collision avoidance and radiolocation with operation up to 10 GHz have a large market potential in the near future of millimetre wave applications.

One advantage of the microstrip antenna is easy matching, fabrication simplicity and low profile, in the sense that the substrate is fairly thin. If the substrate is thin enough, the antenna actually becomes conformal, meaning that the substrate can be bent to conform to a curved surface. Disadvantages of the microstrip antenna include the fact that it is usually narrowband, with bandwidths of a few percent being typical. Also, the radiation efficiency of the microstrip antenna tends to be lower than some other types of antennas, with efficiencies between 70% and 90% being typical. A microstrip antenna operating in a travelling wave configuration could provide the bandwidth and the efficiencies needed.

Travelling-wave antennas are a class of antennas that use a travelling wave on a guiding structure as the main radiating mechanism. It is well known that antennas with open-ended wires where the current must go to zero (dipoles, monopoles, etc.) can be characterized as standing wave antennas or resonant antennas. The current on these antennas can be written as a sum of waves traveling in opposite directions (waves which travel toward the end of the wire and are reflected in the opposite direction). For example, the current on a dipole of length l is given by:

$$\begin{aligned}
 I(z) &= I_o \sin \left[k \left(\frac{l}{2} - z' \right) \right] \\
 &= \frac{I_o}{2j} \left[e^{jk \left(\frac{l}{2} - z' \right)} - e^{-jk \left(\frac{l}{2} - z' \right)} \right]
 \end{aligned} \tag{1}$$

$$= \frac{I_o}{2j} \left[\underbrace{e^{j\frac{kl}{2}} e^{-jkz'}}_{+z \text{ directed wave}} - \underbrace{e^{-j\frac{kl}{2}} e^{jkz'}}_{-z \text{ directed wave}} \right]$$

Traveling wave antennas are characterized by matched terminations (not open circuits) so that the current is defined in terms of waves traveling in only one direction (a complex exponential as opposed to a sine or cosine). A traveling wave antenna can be formed by a single wire transmission line (single wire over ground) which is terminated with a matched load (no reflection). Typically, the length of the transmission line is several wavelengths.

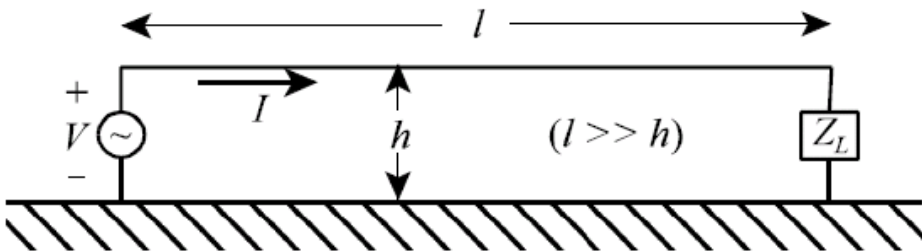


Fig. 1. Beverage or wave antenna.

The antenna shown in Fig. 1 above is commonly called a Beverage or wave antenna. This antenna can be analyzed as a rectangular loop, according to image theory. However, the effects of an imperfect ground may be significant and can be included using the reflection coefficient approach. The contribution to the far fields due to the vertical conductors is typically neglected since it is small if $l \gg h$. Note that the antenna does not radiate efficiently if the height h is small relative to wavelength. In an alternative technique of analyzing this antenna, the far field produced by a long isolated wire of length l can be determined and the overall far field found using the 2 element array factor. Traveling wave antennas are commonly formed using wire segments with different geometries. Therefore, the antenna far field can be obtained by superposition using the far fields of the individual segments. Thus, the radiation characteristics of a long straight segment of wire carrying a traveling wave type of current are necessary to analyze the typical traveling wave antenna. Traveling-wave antennas are distinguished from other antennas by the presence of a traveling wave along the structure and by the propagation of power in a single direction. Linear wire antennas are the dominant type of traveling-wave antennas.

There are in general two types of traveling-wave antennas [1-2]. The first one is the surface-wave antenna, which is a slow-wave structure, where the phase velocity of the wave is smaller than the velocity of light in free space and the radiation occurs from discontinuities in the structure (typically the feed and the termination regions). The propagation wavenumber of the traveling wave is therefore a real number (ignoring conductors or other losses). Because the wave radiates only at the discontinuities, the radiation pattern physically arises from two equivalent sources, one at the beginning and one at the end of the

structure. This makes it difficult to obtain highly-directive singlebeam radiation patterns. However, moderately directly patterns having a main beam near endfire can be achieved, although with a significant sidelobe level. For these antennas there is an optimum length depending on the desired location of the main beam. Examples include wires in free space or over a ground plane, helices, dielectric slabs or rods, corrugated conductors, "beverage" antenna, or the V antenna. An independent control of the beam angle and the beam width is not possible.

The second type of the travelling wave antennas are a fast-wave structure as leaky-wave antenna (LWA) where the phase velocity of the wave is greater than the velocity of light in free space. The structure radiates all its power with the fields decaying in the direction of wave travel.

A popular and practical traveling-wave antenna is the Yagi-Uda antenna. It uses an arrangement of parasitic elements around the feed element to act as reflectors and directors to produce an endfire beam. The elements are linear dipoles with a folded dipole used as the feed. The mutual coupling between the standing-wave current elements in the antenna is used to produce a traveling-wave unidirectional pattern. Recently has been developed a new simple analytical and technical design of meanderline antenna, taped leaky wave antenna (LWA) and taped composite right/left-handed transmission-line (CRLHTL) LWA. The meanderline antenna is a traveling-wave structure, which enables reduction of the antenna length. It has a periodical array structure of alternative square patterns. With this pattern, the extended wire can be made much longer than the initial antenna (dipole) length, so that the selfresonance can be attained. The resonance frequency is then lower and radiation resistance is higher than that of a dipole of the same length. This in turn implies that the antenna is effectively made small.

2. Leaky wave antennas

In detail this type of wave radiates continuously along its length, and hence the propagation wavenumber kz is complex, consisting of both a phase and an attenuation constant. Highly-directive beams at an arbitrary specified angle can be achieved with this type of antenna, with a low sidelobe level. The phase constant β of the wave controls the beam angle (and this can be varied changing the frequency), while the attenuation constant α controls the beamwidth. The aperture distribution can also be easily tapered to control the sidelobe level or beam shape.

All kinds of open planar transmission lines are predisposed to excite leaky waves. There are two kinds of leaky waves. Surface leaky waves radiate power into the substrate. These waves are in most cases undesirable as they increase losses, cause distortion of the transmitted signal and cross-talk to other parts of the circuit. Space leaky waves radiate power into a space and mostly also into the substrate. These waves can be utilized in leaky wave antennas. Leaky-wave antennas can be divided into two important categories, uniform and periodic, depending on the type of guiding structure. A uniform structure has a cross section that is uniform (constant) along the length of the structure, usually in the form of a waveguide that has been partially opened to allow radiation to occur. The guided wave on the uniform structure is a fast wave, and thus radiates as it propagates.

As said previously leaky-wave antennas form part of the general class of travelling-wave antennas which are a class of antennas that use a travelling wave on a guiding structure as the main radiating mechanism [3], as defined by standard IEEE 145-1993: "An antenna that couples power in small increments per unit length, either continuously or discretely, from a travelling wave structure to free space".

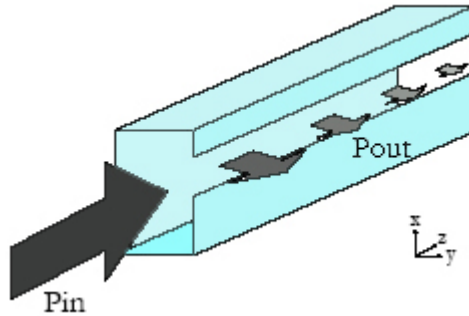


Fig. 2. Rectangular metal waveguide with a slit, aperture of the leaky wave antenna.

Leaky-wave antennas are a fast-wave travelling-wave antennas in which the guided wave is a fast wave, meaning a wave that propagates with a phase velocity that is more than the speed of light in free space.

The slow wave travelling antenna does not fundamentally radiate by its nature, and radiation occurs only at discontinuities (typically the feed and the termination regions). The propagation wavenumber of the travelling wave is therefore a real number (ignoring conductors or other losses). Because the wave radiates only at the discontinuities, the radiation pattern physically arises from two equivalent sources, one at the beginning and one at the end of the structure. This makes it difficult to obtain highly-directive singlebeam radiation patterns. However, moderately directive patterns having a main beam near endfire can be achieved, although with a significant sidelobe level. For these antennas there is an optimum length depending on the desired location of the main beam. An independent control of the beam angle and the beam width is not possible. By contrast, the wave on a leaky-wave antenna (LWA) may be a fast wave, with a phase velocity greater than the speed of light. Leakage is caused by asymmetry, introduced in radiating structure transversal section (e.g.: aperture offset, waveguide shape, etc...), feeding modes or a combination of them. In this type of antennas, the power flux leaking from waveguide to free space (P_{out} in Fig. 2 and Fig. 3), introduces a loss inside structure, determining a complex propagation wavenumber k_z [4-5]:

$$(k_z = \beta - j\alpha) \quad (2)$$

Where α is the leakage constant and β is the propagation constant. The phase constant β of the wave controls the beam angle (and this can be varied changing the frequency), while the attenuation constant α controls the beamwidth. Highly-directive beams at an arbitrary specified angle can be achieved with this type of antenna, with a low sidelobe level.

Moreover the aperture distribution can also be easily tapered to control the sidelobe level or beam shape. Leaky-wave antennas can be divided into two important categories, uniform and periodic, depending on the type of guiding structure.

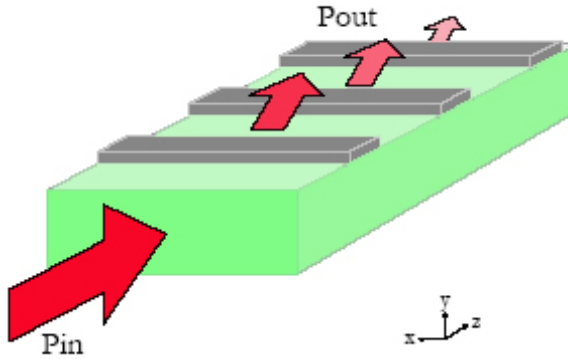


Fig. 3. Example of periodic leaky wave antenna, using a dielectric substrate upon which are placed rods of other material, even metal, in a periodic layout.

A uniform structure has a cross section that is uniform (constant) along the length of the structure, usually in the form of a waveguide that has been partially opened to allow radiation to occur [6]. The guided wave on the uniform structure is a fast wave, and thus radiates as it propagates. A periodic leaky-wave antenna structure is one that consists of a uniform structure that supports a slow (non radiating) wave that has been periodically modulated in some fashion. Since a slow wave radiates at discontinuities, the periodic modulations (discontinuities) cause the wave to radiate continuously along the length of the structure. From a more sophisticated point of view, the periodic modulation creates a guided wave that consists of an infinite number of space harmonics (Floquet modes) [7]. Although the main ($n = 0$) space harmonic is a slow wave, one of the space harmonics (usually the $n = -1$) is designed to be a fast wave, and hence a radiating wave.

3. LWA in waveguide

A typical example of a uniform leaky-wave antenna is a rectangular waveguide with a longitudinal slot. This simple structure illustrates the basic properties common to all uniform leaky-wave antennas. The fundamental TE_{10} waveguide mode is a fast wave, with

$\beta = \sqrt{k_0^2 - (\frac{\pi}{a})^2}$ lower than k_0 . As mentioned, the radiation causes the wavenumber k_z of the propagating mode within the open waveguide structure to become complex. By means of an application of the stationary-phase principle, it can be found in fact that [5]:

$$\sin \theta_m \cong \frac{\beta}{k_0} = \frac{c}{v_{ph}} \quad (3)$$

where ϑ_m is the angle of maximum radiation taken from broadside. As is typical for a uniform LWA, the beam cannot be scanned too close to broadside ($\vartheta_m = 0$), since this corresponds to the cutoff frequency of the waveguide. In addition, the beam cannot be scanned too close to endfire ($\vartheta_m = 90$) since this requires operation at frequencies significantly above cutoff, where higher-order modes are in a bound condition or can propagate, at least for an air-filled waveguide. Scanning is limited to the forward quadrant only ($0 < \vartheta_m < \frac{\pi}{2}$) for a wave travelling in the positive z direction.

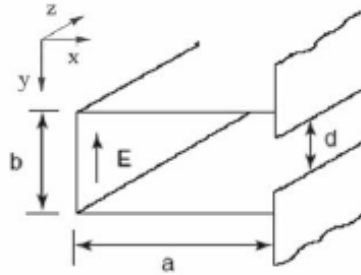


Fig. 4. Slotted guide (patented by W. W. Hansen in 1940).

This one-dimensional (1D) leaky-wave aperture distribution (see Fig. 4), results in a “fan beam” having a narrow beam in the x - z plane (H plane), and a broad beam in the cross-plane. Unlike the slow-wave structure, a very narrow beam can be created at any angle by choosing a sufficiently small value of α . From diffraction theory, a simple formula for the beam width, measured between half power points (3dB), is:

$$\Delta\vartheta \cong \frac{\text{const}}{\frac{L}{\lambda_0} \cos \vartheta_m} \quad (4)$$

“const” is a parameter which is influenced by the type of aperture and illumination; for example, if at the aperture there’s a constant field, $\text{const} = 0.88$ and, if the structure is uniform, $\text{const} = 0.91$. As a rule of thumb, supposing:

$$\Delta\vartheta \cong \frac{1}{\frac{L}{\lambda_0} \cos \vartheta_m} \quad (5)$$

a good approximation of beamwidth is yielded, where L is the length of the leaky-wave antenna, and $\Delta\vartheta$ is expressed in radians. For 90% of the power radiated it can be assumed:

$$\frac{L}{\lambda_0} \cong \frac{0.18}{\frac{\alpha}{k_0}} \Rightarrow$$

$$\Delta\vartheta \cong \frac{\alpha}{k_0}$$

If the antenna has a constant attenuation throughout its length $\alpha_z(z) = \alpha_z$ results:

$$P(z) = P(0)e^{-2\alpha_z z}$$

Therefore, being L the length of antenna, if a perfectly matched load is connected at the end of it, it's possible to express antenna efficiency as:

$$\eta_{rad} = \frac{P(0) - P(L)}{P(0)} = 1 - \frac{P(L)}{P(0)} = 1 - e^{-2\alpha_z L} \quad (6)$$

Rearranging:

$$L = -\frac{\ln(1 - \eta_{rad})}{2\alpha_z} \quad (7)$$

For most application, to gain a 90% efficiency, means that the antenna length is within $10\lambda_0 \div 100\lambda_0$ interval.

Fixing the antenna efficiency, using (7), makes possible to express attenuation constant in terms of antenna length, and vice versa. Using antenna efficiencies grater than 90%-95% is not advisable; in fact, supposing constant antenna cross section and, as a consequence, fixed leakage constant, the necessary length L grows exactly as $\alpha_z L$, which increases asymptotically, as shown in Fig 5. If we want a 100% efficiency ($\eta_{rad} = 1$) from (6):

$$P(L) = 0 \Rightarrow e^{-2\alpha_z L} = 0 \Rightarrow L = \infty$$

we note that is necessary an infinite antenna length.

Substuting (7) in (5), being $\lambda_0 = 2\pi / k_0$:

$$\Delta\vartheta \approx \left(\frac{-4\pi}{\ln(1 - \eta_{rad}) \cos \vartheta_m} \right) \frac{\alpha_z}{k_0}$$

Because $\cos \vartheta_m = \sqrt{1 - \sin^2 \vartheta_m}$, considering (7)

$$\Delta\vartheta \approx \left(\frac{-4\pi}{\ln(1 - \eta_{rad}) \sqrt{1 - \left(\frac{\beta_z}{k_0}\right)^2}} \right) \frac{\alpha_z}{k_0} \quad (8)$$

Since $k_0^2 = k_c^2 + k_z^2$ and having supposed the attenuation constant much smaller than the phase constant, $k_z \approx \beta_z$, getting:

$$\Delta\theta \approx \left(\frac{-4\pi}{\ln(1-\eta_{rad})} \frac{k_c}{k_0} \right) \frac{\alpha_z}{k_0} \quad (9)$$

where k_c is the transverse propagation constant. Alternatively, considering (7):

$$\Delta\theta \approx \frac{2\pi}{L \cdot k_c}$$

Using waveguide theory notation, supposing λ_c the cut-off wavelength:

$$\Delta\theta \approx \frac{\lambda_c}{L} \quad (10)$$

(3) and (10), provided the approximations used to be valid, are a valid tool for describing the main parameters of radiated beam.

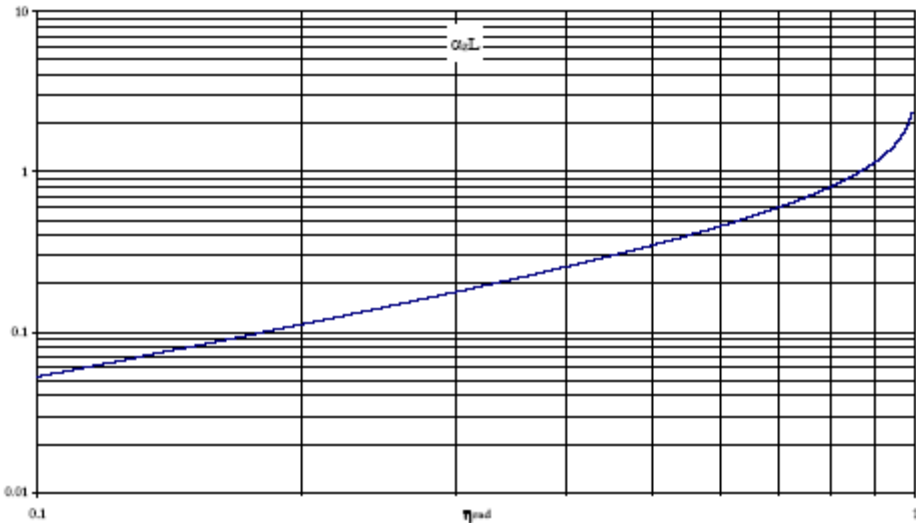


Fig. 5. Variation of $\alpha_z L$ versus antenna efficiency.

Radiation properties of leaky wave antennas are well described by dispersion diagrams. In fact since leakage occurs over the length of the slit in the waveguiding structure, the whole length constitutes the antenna's effective aperture unless the leakage rate is so great that the power has effectively leaked away before reaching the end of the slit. A large attenuation constant implies a short effective aperture, so that the radiated beam has a large beamwidth. Conversely, a low value of α results in a long effective aperture and a narrow beam, provided the physical aperture is sufficiently long.

Moreover since power is radiated continuously along the length, the aperture field of a leakywave antenna with strictly uniform geometry has an exponential decay (usually slow), so that the sidelobe behaviour is poor. The presence of the sidelobes is essentially due to the fact that the structure is finite along z .

When we change the cross-sectional geometry of the guiding structure to modify the value of α at some point z , however, it is likely that the value of β at that point is also modified slightly. However, since β must not be changed, the geometry must be further altered to restore the value of β , thereby changing α somewhat as well.

In practice, this difficulty may require a two-step process. The practice is then to vary the value of α slowly along the length in a specified way while maintaining β constant (that is the angle of maximum radiation), so as to adjust the amplitude of the aperture distribution to yield the desired sidelobe performance.

Radiation modes

Let us consider a generic plane wave, whose propagation vector belongs to plane $(y-z)$, directed towards a dielectric film grounded on a perfect electric conductor (PEC) parallel to plane $(x-z)$, as shown in Fig. 6 [8-9].

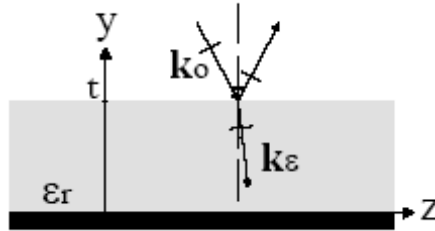


Fig. 6. Incident wave on a grounded dielectric film, whose thickness is t .

If the incident wave polarization is linear and parallel to the x axis, since both reflection and refraction occur:

$$\begin{cases} E_{x_0} = Ae^{-jk_{y_0}(y-t)} + Ce^{jk_{y_0}(y-t)} & y \geq t \\ E_{x_\epsilon} = B \cos(k_{y_\epsilon} y) + D \sin(k_{y_\epsilon} y) & t \geq y \geq 0 \end{cases}$$

Being the tangent components of electric field null on a PEC surface, $B = 0$:

$$\begin{cases} E_{x_0} = Ae^{-jk_{y_0}(y-t)} + Ce^{jk_{y_0}(y-t)} & y \geq t \\ E_{x_\epsilon} = D \sin(k_{y_\epsilon} y) & t \geq y \geq 0 \end{cases} \quad (11)$$

One constant can be expressed by the remaining two, as soon as continuity of tangent components of electric field is considered in $y = t$. First equation of (11) contains an exponential term which, diverging for $y \rightarrow \infty$, violates the radiation condition at infinite distance. Therefore, $C \neq 0$ only near plane $y = t$, at the incidence point. k_z can assume any

value from 0 to k_0 (i.e.: radiating modes); above it, only discrete values of kz exist, identifying the associated guided modes. Since separability condition must be satisfied, in air:

$$k_0^2 = \omega^2 \mu_0 \varepsilon_0 = k_{y_0}^2 + k_z^2 \text{ where } k_0 \in \Re \quad (12)$$

For every k_z , it's now possible calculate k_{y_0} . In fact, considering only positive solutions:

$$k_{y_0} = \sqrt{k_0^2 - k_z^2}$$

Obtaining:

Mode	Wave numbers	
Guided	$k_z > k_0$	$k_{y_0} \in \Im$
Radiating	$k_0 > k_z > 0$	$k_0 > k_{y_0} > 0$
Evanescent	$0 > k_z > -j\infty$	$\infty > k_{y_0} > k_0$

Table 1. Wave modes identified by k_z .

Thus, a spectral representation of electromagnetic field near the air-dielectric interface, must contain all values of k_{y_0} , from 0 to ∞ : the associated integral is complex and slowly convergent. Alternatively, a description, which uses leaky waves and guided modes, both discrete, can well approximate such field.

It's been observed that it's often enough a single leaky wave to obtain a good far field description.

Letting $k_y = k_{y_0}$, from (2) and (12), in general:

$$\begin{cases} k_0 = \beta_y^2 + \beta_z^2 - \alpha_y^2 - \alpha_z^2 \\ 0 = \beta_y \alpha_y + \beta_z \alpha_z \end{cases}$$

alternatively

$$\begin{cases} k_0^2 = |\bar{\beta}|^2 \cdot |\bar{\alpha}|^2 \\ 0 = \bar{\beta} \cdot \bar{\alpha} \end{cases} \quad (13)$$

Having defined the attenuation and the phase vectors, respectively, as:

$$\alpha = \alpha_y \bar{y}_0 + \alpha_z z_0$$

$$\beta = \beta_y \bar{y}_0 + \beta_z z_0$$

Being $k_0 \in \Re$, from (9) $|\bar{\beta}| \neq 0$, and $|\bar{\beta}| > |\bar{\alpha}|$

Considering waves propagating in the positive direction of z axis, $\beta_z > 0$ and supposing no losses in z direction, $\alpha_z = 0$, from (13):

$$0 = \beta \cdot \alpha$$

Leaving out $\beta_y, \alpha_y = 0$, two situations can occur: $\alpha_y = 0$ and $\beta_y = 0$. If $\alpha_y = 0$, equations describe a uniform plane wave passing the air-dielectric interface. On the other hand, if $\beta_y = 0$, two types of superficial waves exist, depending on $\alpha_y = 0$ sign:

$\alpha_y > 0$		Confined superficial wave
$\alpha_y < 0$		Improper superficial wave

Fig. 7. Superficial waves at air-dielectric interface when $\beta_y = 0$.

Because confined superficial waves amplitude decreases exponentially as distance from interface increases, when y is greater than 10 times radiation wavelength, electromagnetic field practically ceases to exist. Improper superficial wave, whose amplitude increases exponentially as distance from interface increases, are not physically possible because they violate the infinite radiation condition.

Removing the hypothesis $\alpha_z = 0$, both $\alpha_y \neq 0$, and $\beta_y \neq 0$.

	Losses in Dielectric
	Leaky Wave

Fig. 8. General mutual β and α configurations depicting condition $0 = \beta \cdot \alpha$.

When losses in dielectric occur, β must point towards the inner part of dielectric to compensate such losses (see Fig.8). In the other configuration, when β points upwards, even though a non-physical solution is described, the associated wave is useful to describe electromagnetic field near air-dielectric interface.

4. LWA in microstrip

Microstrip antenna technology has been the most rapidly developing topic in antennas during the last twenty years [10]. Microstrip is an open structure that consists of a very thin metallic strip or patch of a width, w, separated from a ground plate by a dielectric sheet

called substrate (Fig. 9). The thickness of the conductor, t , is much less than a wavelength, and may be of various shapes. The height of the substrate, h , is usually very thin compared to the wavelength ($.0003 \lambda \leq h \leq 0.05 \lambda$) [11]. The substrate is designed to have a known relative permittivity, ϵ_r , that is homogeneous within specified temperature limits.

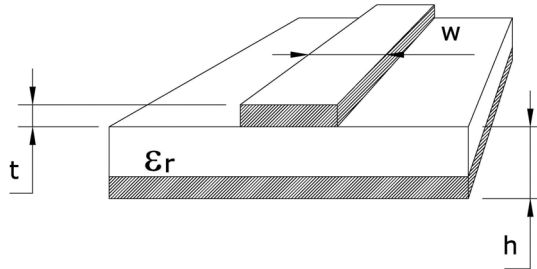


Fig. 9. Geometry of a microstrip transmission line.

The antenna can be excited directly by a microstrip line, by a coaxial cable, or a combination of the two. The antenna can also be fed from a microstrip line without direct contact through electromagnetic coupling. Feeding by electromagnetic coupling through an aperture in the ground plane tends to improve bandwidth. To maximize efficiency, the impedance of the feed must be matched to the input impedance of the antenna. There are a variety of stubs, shunts, and other devices used for matching. The major disadvantages of microstrip are lower gain, very narrow bandwidth, low efficiency and low power handling ability. In addition, antennas made with microstrip typically have poor polarization purity and poor scan performance [12].

Operating above the cutoff frequency, the field lines of microstrip extend throughout the substrate as well as into the free space region above the substrate, as seen in Fig. 10. The phase velocity of the field in the free space surrounding the structure is the speed of light, c , and the phase velocity of the field in the substrate is given by Equation (14)

$$v_p = \frac{c}{\sqrt{\epsilon_r}} \quad (14)$$

This difference in phase velocity at the interface between the substrate and free space makes the TEM mode impossible. Instead, the fundamental mode for microstrip is a quasi-TEM mode, in which both the electric and magnetic fields have a component in the direction of propagation. Likewise, a higher order mode in microstrip is not purely TE or TM, but a hybrid combination of the two. The n th higher order mode is termed the TE_n mode. The fundamental mode of microstrip, as seen in Fig. 10, does not radiate since the fields produced do not decouple from the structure. If the fundamental mode is not allowed to propagate, the next higher order mode will dominate. Fig. 11 shows the fields due to the first higher order mode, TE_{10} . A phase reversal, or null, appears along the centerline, allowing the fields to decouple and radiate.

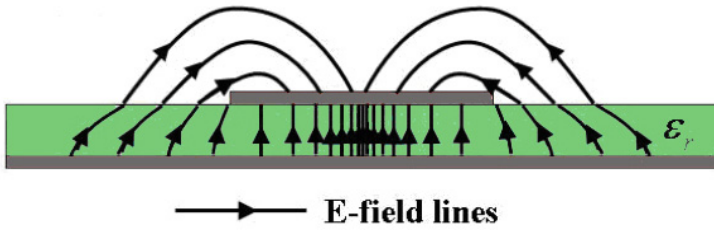


Fig. 10. Field pattern associated with the fundamental mode of microstrip.

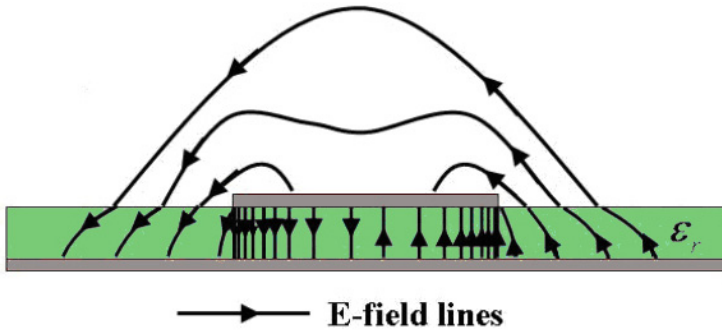


Fig. 11. Field pattern associated with the first higher order mode of microstrip.

Recently, there has been significant interest in the microstrip leaky-wave antenna which utilizes a higher order radiative microstrip mode. Since Menzel in 1979, published the first account of a travelling wave microstrip antenna that used a higher order mode to produce leaky waves [13], many microstrip leaky-wave antenna designs incorporating various modifications have been investigated. The design of Menzel antenna [13], can be seen in Fig. 12. Menzel's antenna uses seven slots cut from the conductor along the centerline to suppress the fundamental mode allowing leaky wave radiation via the first higher order mode. Menzel's antenna has been analyzed by a host of researchers over the past 25 years [14] and its performance is known and reproducible. Instead of transverse slots, we can use a metal wall down the centerline of the antenna to block the fundamental mode. Symmetry along this metal wall invites the application of image theory. One entire side of the antenna

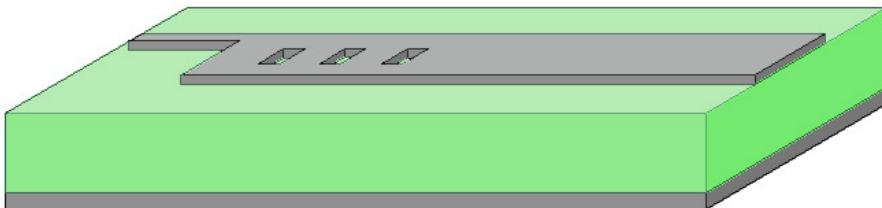


Fig. 12. Menzel's original antenna [13].

is now an image of the other side, making it redundant and unneeded. This property allows to design the resulting antenna half of the width of Menzel's antenna, as shown in Fig. 13.

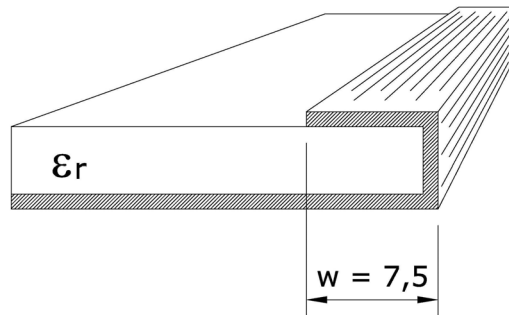


Fig. 13. Half Width Leaky Wave Antenna.

As mentioned the microstrip structures do not radiate for the fundamental mode, therefore, a higher order mode must be excited to produce leaky waves. This method of producing radiation by exciting higher order modes in a transmission line has been documented since the 1950's [6]. By the 1970's, rectangular waveguides, circular waveguides, and coaxial cables were in use as leaky traveling wave antennas. However, until Menzel, the jump to microstrip had not been made. By looking at a cross section of microstrip excited in the fundamental mode, the E field is strongest in the center and tapers off to zero at the sides, as depicted in Fig 10. If the electric field down the centerline is suppressed, the fundamental mode will be prohibited, forcing the energy to propagate at the next higher mode, TE_{10} . As seen in Fig. 11, TE_{10} mode causes E to be strongest at the edges. Menzel attempted to force the TE_{10} mode using several means. Feeding two equal magnitude waves 180° out of phase with a "T" or "Y" feed produced TE_{10} as desired, but did not fully eliminate the fundamental mode. Easier to produce and providing an even better response was given using transverse slots down the centerline (Fig. 12). The multiple feeds were not necessary to produce the TE_{10} mode when the fundamental mode was suppressed. Menzel demonstrated that the beam angle can be predictively steered by input frequency if the electrical length of the antenna is at least 3λ . If the length is less than 3λ , too little of the incident wave is being radiated and a resonance standing wave pattern is forcing the beam toward broadside. Qualitative analysis shows that the beamwidth of Menzel's antenna is not frequency dependent, however, it is inversely related to length. The 3 dB beamwidth approaches 10° for electrical length of over 6λ and approaches nearly 90° for fractions of a wavelength. Menzel's gain varied from 7 dB for $l = 0.2\lambda$ to 14 dB for $l = 4\lambda$. 7 dB is comparable to a similar sized resonant antenna. An antenna longer than $l = 4\lambda$ would have an even higher gain as the radiation aperture increases. Lee notes that Menzel assumed that his antenna should radiate simply because the phase constant due to his operating frequency was less than k_0 [15]. If Menzel had considered the complex propagation constant, he would have realized that his antenna was operating in a leaky regime. The length would need to be roughly 220 mm, or more than twice as long as his design, to

radiate at 90% efficiency. Radiation patterns in Menzel's paper clearly show the presence of a large backlobe due to the reflected traveling wave.

Now this class of printed antennas that is particularly well suited for operation at mm-wave frequencies, alleviate some of the problems associated with resonant antennas since they provide higher gain, broader bandwidth performance, and frequency scanning capabilities. These microwave and millimeter leaky wave antennas, have the same properties of the waveguide leaky wave antennas described previously. In addition, when opening a waveguide to free space, a discrete spectrum is not enough to express an arbitrary solution [16].

In fact, when considering a closed region, all characteristic solutions, individuated by the associated eigenvalues, constitute a complete and orthogonal set of modes, whose linear combination can express any field satisfying boundary conditions. As soon as the region is not perfectly bounded, an arbitrary field solution cannot be expressed only using discrete eigenmodes but, generally, a continuous spectrum of modes, which don't necessarily have finite energy (e.g.: plane waves), must be considered, too.

Fortunately, for leaky wave antennas, an approximation that uses particular waves, called leaky, can be used instead of the continuous spectrum. Moreover, leaky waves are well described by dispersion constants (i.e.: leakage and phase constants) that strongly affect the radiated beam width and elevation.

5. Dispersion curves, spectral-gap

Dispersion curves, describing how attenuation and phase vectors, solutions of dispersion equation (12), evolve, are a valid tool to study leaky waves.

As discussed previously, the radiation mechanism of higher order modes on microstrip LWA is attributed to a traveling wave instead of the standing wave as in patch antennas and above cutoff frequency, where the phase constant equals the attenuation constant ($\alpha_c = \beta_c$), it is possible to observe three different range of propagation: bound wave, surface wave and leaky wave [15]. At low frequency, below the cutoff frequency, we have the reactive region due to evanescent property of LWA.

From (3) we can observe that the leaky mode leaks away in the form of space wave when $\beta < K_0$, therefore we can define the radiation leaky region from the cutoff frequency to the frequency at which the phase constant equals the free-space wavenumber ($\beta = K_0$). For ($\beta > K_s$) we have the bound mode region and for $K_0 < \beta < K_s$, exists a narrow frequency range ($K_s < \frac{1}{\sqrt{\epsilon_r}}$), in which we can have surface-wave leakage, where K_s is the surface wavenumber.

Moreover the transition region between surface wave leakage and space wave leakage including a small range in frequency for which the solution is non-physical, and it therefore cannot be seen. For this reason, the transition region is called a spectral gap. Such a spectral gap occurs commonly (but not always) at such transitions in printed circuits, but it also occurs in almost all situations for which there is a change from a bound mode to a leaky

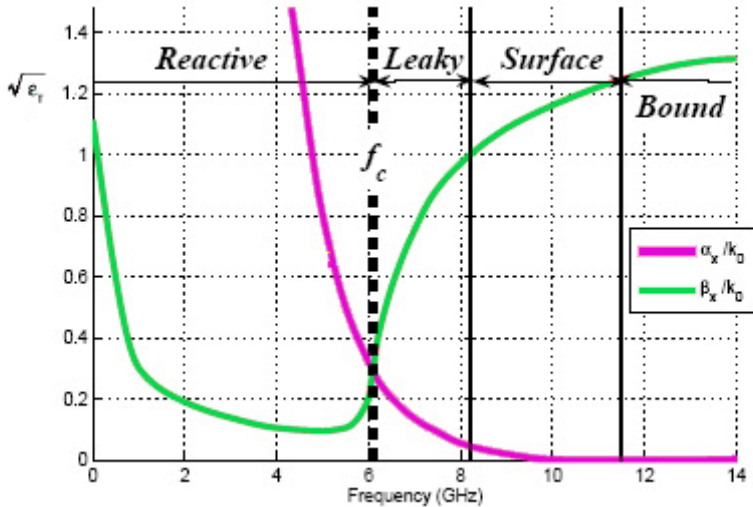


Fig. 14. The typical normalized attenuation constant, α / k_0 , and phase constant β / k_0 , in the direction of propagation of the first higher order mode, TE_{10} . There are four frequency regions associated with propagation regimes: Reactive, Leaky, Surface, and Bound.

mode, or vice versa. For example, a spectral gap will appear when the beam approaches endfire in all leaky-wave antennas whose cross section is partly loaded with dielectric material. It is necessary to employ a greatly enlarged scale, on which the dispersion plot is sketched qualitatively. The transition region itself is divided into two distinct frequency ranges, one from point A to point B and the second from point B to point C. Before point B, a leaky wave occurs. As soon as frequency reaches f_1 (point B), an improper superficial wave is solution of dispersion equation [9]. Because, both α_z and β_z cannot increase with frequency between f_1 and f_2 , their trend will change until point C, from which a confined superficial wave is an acceptable solution for increasing values of frequency.

To depict normalized constants behaviour around the spectral-gap, it's necessary a very precise numerical method since, leaving out particular structures, its width (Δf) is very small compared to working frequencies.

The dispersion characteristics for microstrip has been investigated by a number of authors using different full wave methods and evaluating different regimes of the dispersion characteristic.

The spectral domain analysis has proven to be one of the most efficient and fruitful techniques to study the dispersion characteristics of printed circuit lines [17]. As is explained in literature, the Galerkin method in conjunction with Parseval theorem can be used to pose the dispersion relation of an infinite printed circuit line as the zeros of the following equation:

$$F(k_z) = \int_{C_x} \tilde{G}_{zz}(k_x, k_z) \tilde{T}^2(k_x) dk_x = 0 \quad (15)$$

where $\tilde{T}(k_x)$, is the Fourier transform of the basis function $T(k_x)$ used to expand the longitudinal current density on the strip conductor as

$$J_{sz}(x, z) = T(x)e^{-jk_z z}$$

The term $\tilde{G}_{zz}(k_x, k_z, \omega)$ is the zz component of the spectral dyadic Green's function, and C_x is an appropriate integration path in the complex k_x plane to allow for an inverse Fourier transform non uniformly convergent function. The spectral dyadic Green's function has the following singularities in the complex k_x plane: branch point, a finite set of poles on the proper sheet and a infinite set of poles on the improper sheet. For a fixed frequency, the function $F(k_z)$ is not uniquely defined because of the many possible different C_x integration paths that can be used to carry out the integral (15) [18].

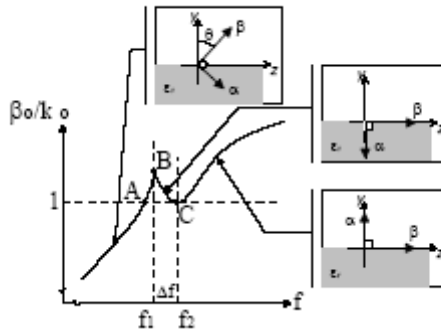


Fig. 15. Transition region between leaky wave and confined superficial wave showing the spectral-gap occurring f_1 and f_2 .

The different C_x paths come from the different singularities of the spectral dyadic Green's function, that can be detoured around. For complex leaky mode solution, an integration path detouring around only the proper poles of the spectral dyadic Green's function is associated with an surface-wave leaky mode solution. If the path also detours around the branch points, passing through the branch cuts and, therefore, lying partly on the lower Riemann sheet, the path will be associated with an space-wave leaky mode solution (see Fig. 16). This procedure, is not trivial. We shown in the next chapters how it is possible to extract the propagation constant of a microstrip LWA more simply using an FDTD code with UPML boundary condition, who directly solves the $_elds$ in the time domain using Maxwell's equations and with which the analysis is easy modifying the geometries of the LWAs. The results are in a good agreement with transverse resonance approximation (a full wave method) derived by Kuestner [19].

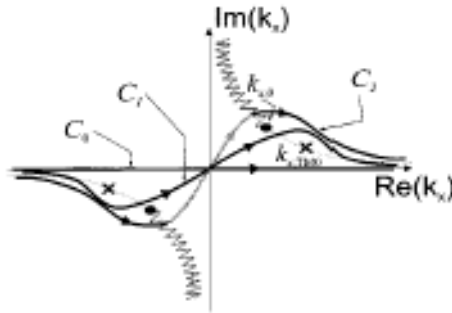


Fig. 16. Possible integration paths C . The three different are denoted as C_0 for the real-axis path (bound mode solution), C_1 for the path that detours around only the spectral dyadic Green's function poles (surface-wave leaky modes solution), and C_2 for the path that also passes around the branch points (space-wave leaky modes solution).

6. Tapered leaky wave antennas

Nowdays, some applications especially with regard to communication applications like the indoor wireless LAN(WLAN) actually are increasing the use of millimeterwave antennas like leaky-wave antennas (LWA), suited for more purpose. In detail, the transmitting/receiving antennas with relatively broadbeam and broadband can be obtained from the curved and tapered leakywave structures. In fact, the microstrips (LWA), are very popular and widely used in applications thanks to their advantages of low-profile, easy matching, narrow beamwidth, fabrication simplicity, and frequency/electrical scanning capability. Is well know that the radiation mechanism of the higher order mode on microstrip LWAs is attributed to a traveling wave instead of the standing wave as in patch antennas. Moreover the symmetry of the structure along this physical grounding structure, thanks to the image theory, allows to design only half of an antenna with the same property of one in its entirety, and reducing up to 60% the antenna's dimensions. Using this tapered antenna we can obtained a quasi linear variations of the phase normalized constant and than a quasi linear variations of the its radiation angle. Moreover the profile of the longitudinal edges of the LWA, was designed, by means of the reciprocal slope of the cutoff curve, symmetrically to the centerline of the antenna, allows a liner started of leaky region.

Nevertheless the variation of the cross section of the antenna, allowing a non-parallel emitted rays, such as happens in a non-tapered LWA. In fact, using the alternative geometrical optics approach proposed in the tapering of the LWA, for a fixed frequency, involves the variation of the phase constant β and the attenuation constant a , obtained as a cut plane of 3D dispersion surface plot varying width and frequency. We can be determined a corresponding beam radiation interval with respect to endfire direction. As mentioned previously, for a tapered antenna with a curve profile (square root law profile) the radiation angle in the leaky regions, vary quasi linearly whit the longitudinal dimension, so it is possible to calculate the radiation angle of the antenna as a average of the phase constant using the simple formula.

Alternatively using the geometrical optics approach it is easy to determine the closed formula to predict the angle of main beam of a tapered LWA.

7. Design of tapered LWA

The radiation mechanism of the higher order mode on microstrip LWAs is attributed to a traveling wave instead of the standing wave as in patch antennas [13,20].

We can explain the character of microstrip LWAs through the complex propagation constant $k = \beta - j\alpha$, where β is the phase constant of the first higher mode, and α is the leakage constant. Above the cutoff frequency, where the phase constant equals the attenuation constant ($\alpha_c = \beta_c$), it is possible to observe three different propagation regions: bound wave, surface wave and leaky wave.

The main-beam radiation angle of LWA can be approximated by:

$$\theta = \cos^{-1}\left(\frac{\beta}{K_0}\right) \quad (16)$$

where θ is the angle measured from the endfire direction and K_0 is the free space wavenumber. According to (16) we can observe that the leaky mode leaks away in the form of space wave when $\beta < K_0$, therefore we can define the radiation leaky region, from the cutoff frequency to the frequency at which the phase constant equals the free-space wavenumber ($\beta = K_0$). An example of tapered LWA was proposed in [21-22], using an appropriate curve design to taper LWA.

In fact through the dispersion characteristic equation, evaluated with FDTD code, we can obtain the radiation region of the leaky waves indicated in the more useful way for the design of our antenna:

$$\frac{c}{2w_{eff}\sqrt{\epsilon_r}} = f_c < f < \frac{f_c\sqrt{\epsilon_r}}{\sqrt{\epsilon_r - 1}} \quad (17)$$

From equation (17) we can observe that the cutoff frequency increases when the width of the antenna decrease, shift toward high frequencies, the beginning of the radiation region as shown in Fig. 17. 1.

Therefore it is possible to design a multisection microstrip antenna [as Type I antenna in Fig. 17.a], in which each section able to radiate at a desired frequency range, can be superimposed, obtaining an antenna with the bandwidth more than a uniform microstrip antenna. In this way every infinitesimal section of the multisection LWA obtained overlapping different section should be into bound region, radiation region or reactive region, permitting the power, to uniformly radiated at different frequencies.

Using the same start width and substrate of Menzel travelling microstrip antenna (TMA) [13], and total length of 120 mm., we have started the iterative procedure mentioned in [23] to obtain the number, the width and the length of each microstrip section. From Menzel

TMA width, we have calculated the f_{START} (onset cutoff frequency) of the curve tapered LWA, then, choosing the survival power ratio ($\tau = e^{-2\alpha_i L_i}$) opportunely, at the end of the first section, we have obtained the length of this section. The cutoff frequency of subsequent section (f_i), was determined by FDTD code, while the length of this section was determined, repeating the process described previously. This iterative procedure was repeated, until the upper cutoff frequency of the last microstrip section.

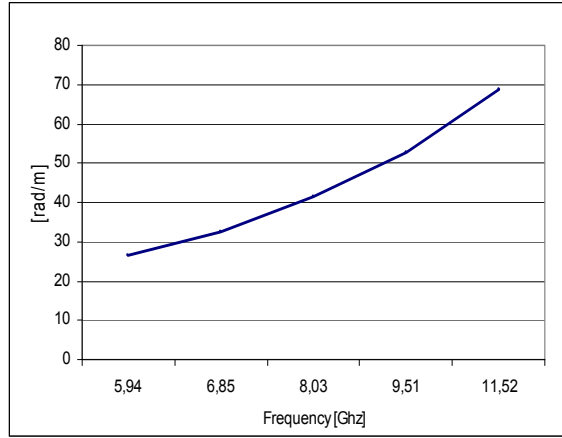


Fig. 17.1 Cutoff frequency of multisection microstrip LWA.

The presence of ripples in return loss curve and the presence of spurious sidelobes shows the impedance mismatch and discontinuity effect of this multisection LWA that reduce the bandwidth. A simple way to reducing these effects is to design a tapered antenna in which the begin and the end respectively of the first and the last sections are linearly connected together (as the Type II antenna in Fig. 17.a).

Alternatively the ours idea was to design a LWA using a physical grounding structure along the length of the antenna, with the same contour of the cutoff phase constant or attenuation constant curve ($\alpha_c = \beta_c$), obtained varying the frequency (the cutoff frequency f_c is the frequency at which $\alpha_c = \beta_c$), for different width and length of each microstrip section as shown in Fig. 17.1, employing the following simple equation (18):

$$\beta_c = c_1 f^2 + c_2 f + c_3 \quad (18)$$

obtained from linear polynomials interpolation, where $c_1 = 0.0016$, $c_2 = 0.03$, $c_3 = -15.56$.

The antenna layout (as the Type III antenna in Fig. 17.a), was optimized through an 3D electromagnetic simulator, and the return loss and the radiation pattern was compared with Type I antenna and Type II antenna.

8. Simulation results

An asymmetrical planar 50Ω feeding line was used to excite the first higher-order mode while a metal wall down the centerline connecting the conductor strip and the ground plane was used to suppress the dominant mode for Type I - III. The chosen substrate had a dielectric constant of 2.32 and a thickness of 0.787 mm, while the total length of the leaky wave antenna was chosen to be 120 mm.

The leaky multisection tapered antenna Type I was open-circuited, with a 15 mm start width, and 8.9 mm of final width obtained according to [23]. For LWA layout Type I, we used four microstrip steps, for layout Type II we tapered the steps linearly, while the curve contour of the LWA layout Type III, was designed through equation (18).

Fig. 17.b shows the simulated return loss of three layouts. We can see that the return loss (S11) of Type I is below -5 dB from 6 to 10.3 GHz, but only three short-range frequencies are below -10 dB. S11 of Type II is below -5 dB from 6.1 to 9.1 GHz, and below -10 dB from 6.8 to 8.6 GHz. At last, S11 of Type III is below -5 dB from 6.8 to 11.8 GHz, and below -10 dB from 8.0 to 11.2 GHz. In Fig.17.c are shows the mainlobe direction at 9.5 GHz for the different Type I to Type III. We can see a reduction of sidelobe and only few degrees of mainlobe variation between Type I to Type III. Moreover, in Fig. 18 is shown the variation of mainlobe of antenna Type III, for different frequency, while in Fig. 19 is shown the trend of gain versus frequency of the same antenna. It is clear that, the peak power gain is more than 12 dBi, which is almost 3 dBi higher than uniform LWAs.

Finally the simulated VSWR is less than 2 between 8.01 and 11.17 GHz (33%), yielding an interesting relative bandwidth of 1.39:1, as shown in Fig. 20, compared with uniform microstrip LWAs (20% for VSWR < 2) as mentioned in [24].

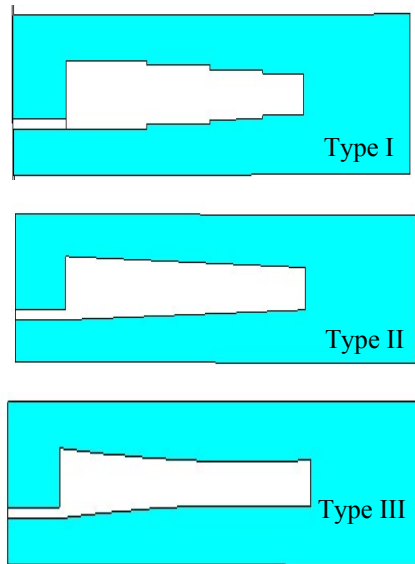


Fig. 17a. Layout of leaky wave antennas Type I-III. A physical grounding structure was used to connecting the conductor strip and the ground plane.

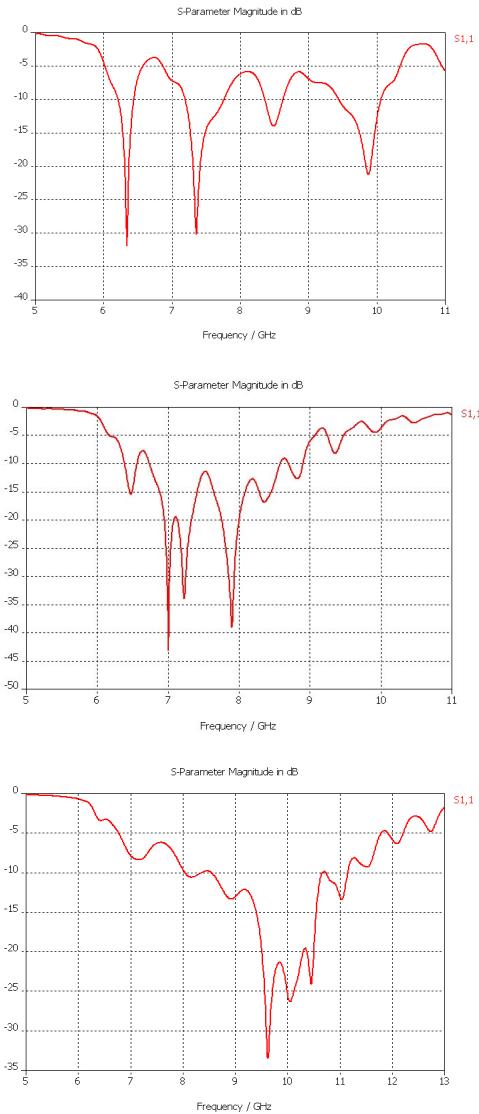


Fig. 17b. Simulated Return loss of Type I-III LWA.

These results indicate a high performance of Type III LWA: high efficiency excitation of the leaky mode, increases of the bandwidth, improves the return loss and reduction of 19% of metallic surface with respect to uniform LWA. Moreover, these results are in a good agreement whit the experimental results of return loss and radiation pattern of a prototype made using a RT/Duroid 5880 substrate with thickness of 0.787 mm and relative dielectric constant of 2.32, as shown in Fig.21 and Fig.22.

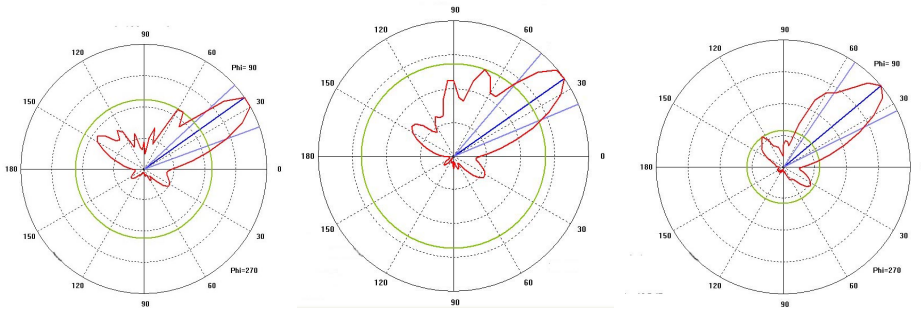


Fig. 17c. Radiation patterns of Electric field (H plane) of Type I-III, LWA at 9.5 GHz.

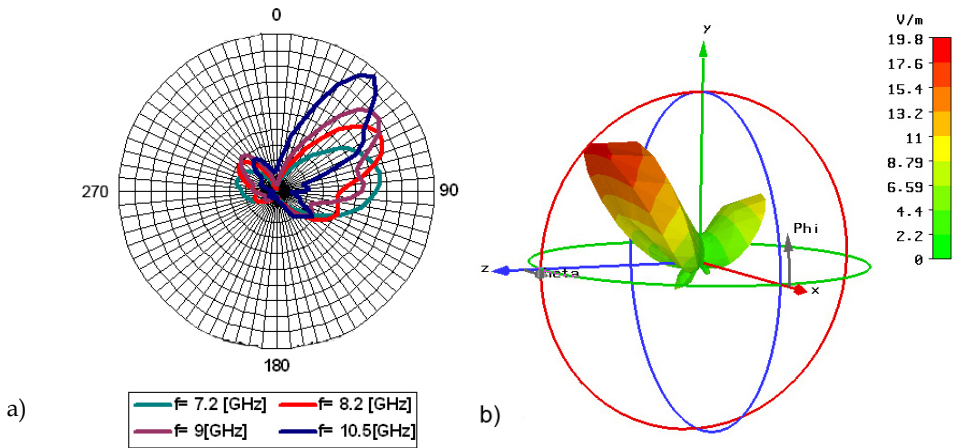


Fig. 18. a) Simulated radiation patterns of E field of LWA Type III for different frequency. b) 3-D radiation pattern of Electric field.

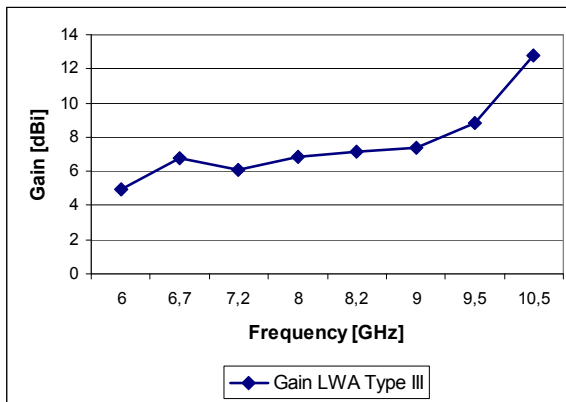


Fig. 19. The gain versus frequency of the LWA Type III.

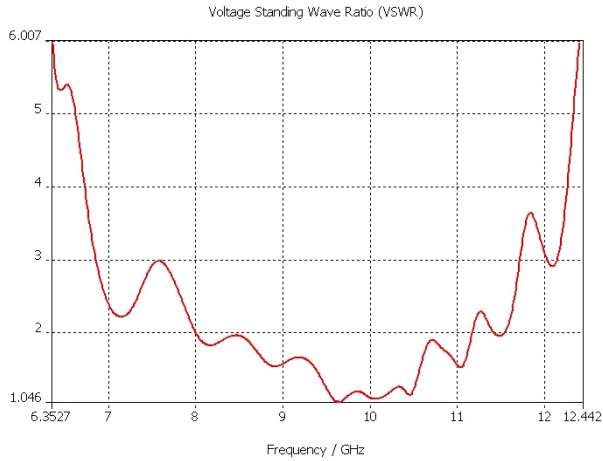


Fig. 20. Simulated VSWR of LWA Type III.

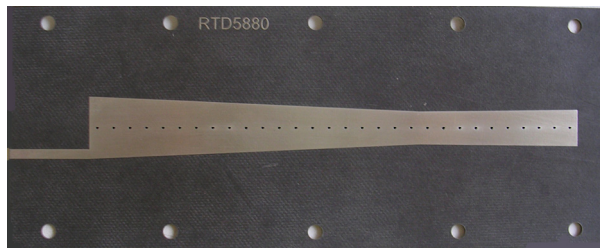
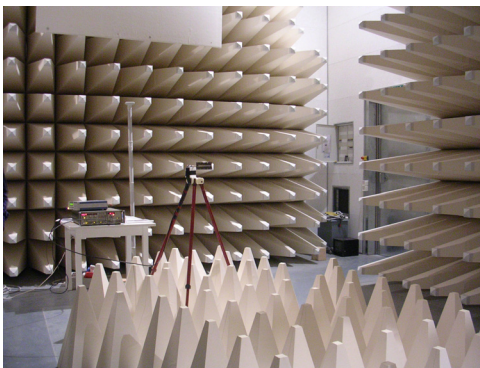
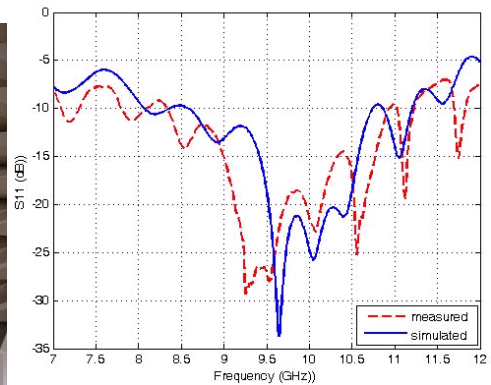


Fig. 21. A prototype of tapered LWA Type III with holes made in the centerline of the antenna.



a)



b)

Fig. 22. a) Measurement set-up of LWA Type III. b) Experimental and simulated return loss of LWA Type III.



Fig. 23. A prototype of half tapered LWA.

Moreover the use of a physical grounding structure along the length of the antenna, as suggested in [21-22], allows the suppression of the dominant mode (the bound mode), the adoption of a simple feeding, and due to the image theory, it is also possible to design only half LWA (see Fig. 23) with the same property of one entire, as shown in Fig. 24 and in Fig. 25, reducing up to 60% the antenna's dimensions compared to uniform LWAs [24].

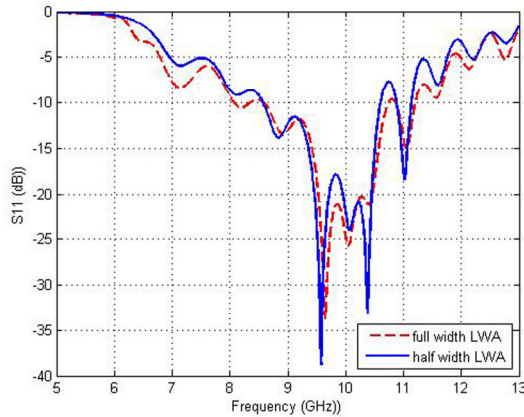


Fig. 24. Measured return loss of full and half Leaky Wave Antennas

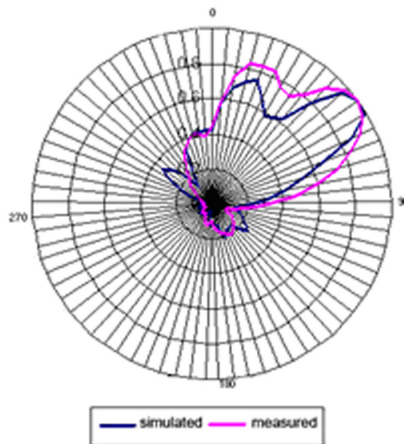


Fig. 25. The measured and simulated radiation patterns of E field of half tapered LWA at 8 GHz.

9. Focusing-diverging property

As described in [21-22], the profile of the longitudinal edges of the LWA, was designed, by means of the reciprocal slope of the cutoff curve, symmetrically to the centerline of the antenna, allows a linear started of leaky region. Using this tapered antenna we can obtained a quasi linear variations of the phase normalized constant and than a quasi linear variations of the its radiation angle as we can see in Fig. 26 and in Fig. 27. Nevertheless the variation of the cross section of the antenna, allowing a non-parallel emitted rays, such as happens in a non-tapered LWA (see Fig. 28). In fact, as was described in the alternative geometrical optics approach proposed in [24] the tapering of the LWA, for a fixed frequency, involves the variation of the phase constant β and the attenuation constant α , as shown in Fig. 29, obtained as a cut plane of 3D dispersion surface plot varying width and frequency (see Fig. 30).

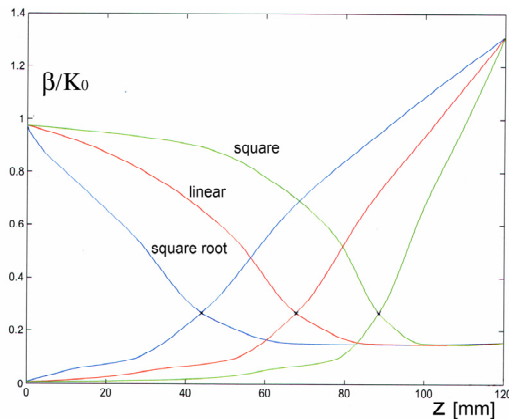


Fig. 26. The variation of the main beam radiation angle versus length of the antenna, at $f = 8$ GHz, for linear, square and square root profile of the LWA.

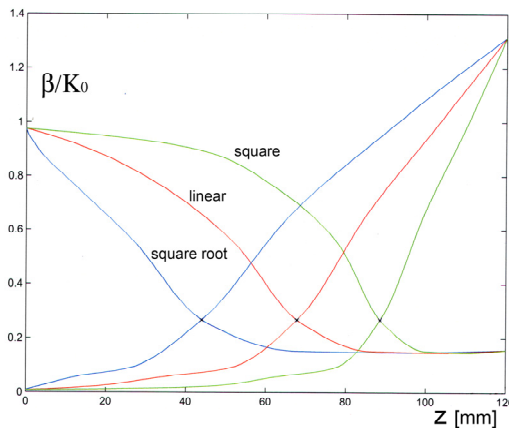


Fig. 27. The variation of the phase constant versus length of the antenna, at $f = 8$ GHz, for linear, square and square root profile of the LWA.

From (16) can be determined in the leaky regions of the antenna, a corresponding beam radiation interval $[\vartheta_{\min}, \vartheta_{\max}]$, with respect to endfire direction.

As mentioned previously, for a tapered antenna with a curve profile (square root law profile) the radiation angle in the leaky regions, vary quasi linearly whit the longitudinal dimension, so it is possible to calculate the radiation angle of the antenna as a average of the phase constant using the simple equation (19).

$$\vartheta_m = \text{sen}^{-1} \left(-\frac{1}{K_0 L} \int_0^L \beta(z) dz \right) \tag{19}$$

Alternatively using the geometrical optics it is easy to determine the closed formula to predict the angle of main beam of a tapered LWA. Through simple mathematical passages, the main beam angle ϑ_m can be obtained by the equation (20).

$$\vartheta_m = \text{sen}^{-1} \left(\frac{A \text{sen} \vartheta_{\min}}{\frac{1}{2} \sqrt{(2A \text{sen} \vartheta_{\min})^2 + (L + 2C \cos \vartheta_{\max})^2}} \right) \tag{20}$$

Where A and C are respectively the distance between real focus F and the beginning and the end of the length of the antenna L . Therefore, if we know the begin width and the end width of the antenna, from the curves of normalized phase and attenuation constant at fixed frequency, we can determine the beam radiation range from (16), and the main beam angle through (20).

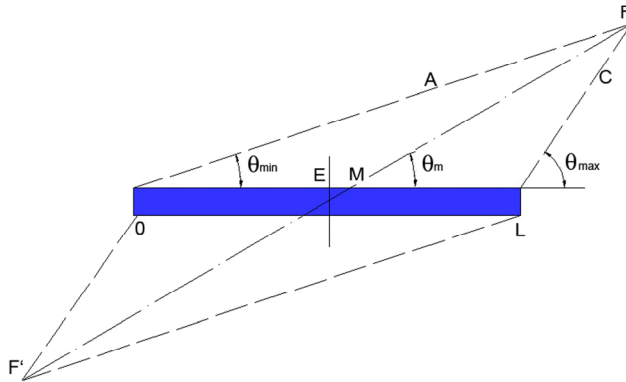


Fig. 28. The ray optical model for a tapered Leaky Wave antenna.

Furthermore this focusing phenomena of a tapered LWA can determine a wide-beam pattern in a beam radiation range which is evident when the antenna length is increased ($L \cong 50\lambda_0$) [25].

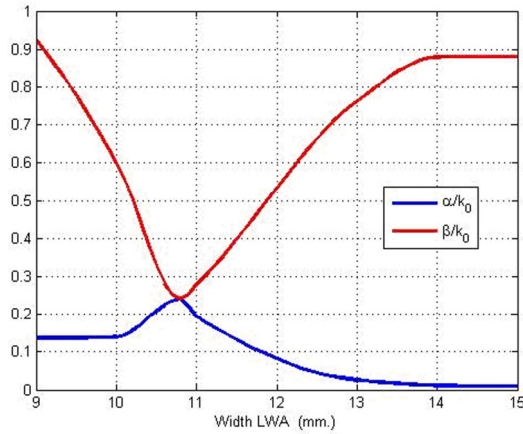


Fig. 29. The curve of normalized phase and attenuation constants versus the width, at 8 GHz for the LWA with the angular range $[28^\circ, 76^\circ]$. The leaky region start from 10.8 mm. (cutoff frequency).

To obtain a broad beam pattern without the use of a longer LWA, we can bend a tapered LWA (see Fig. 31), leading the electromagnetic waves to diverge. This, increases the beam of the radiation pattern and reduce furthermore the back lobes as we can see compared the curves of Fig. 32. Finally in Fig. 33 is shown the measured return loss of half bend LWA Type III.

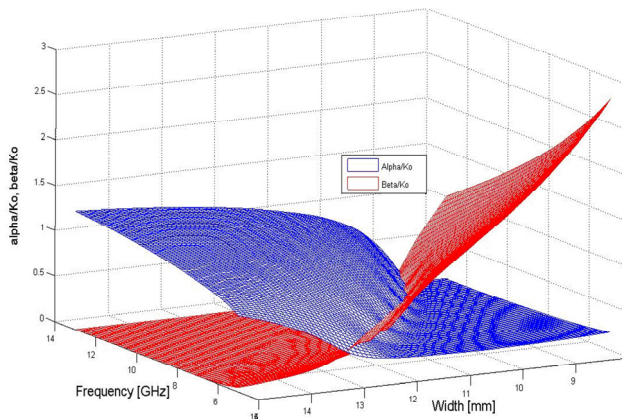
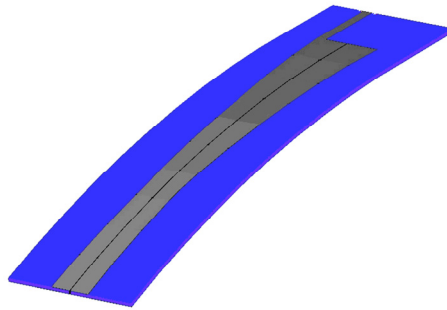
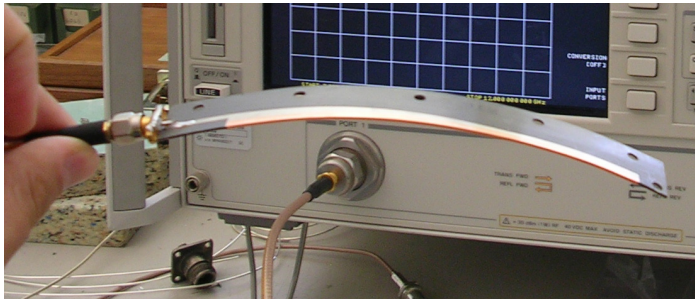


Fig. 30. The 3D normalized phase constant and attenuation constant of tapered LWA versus frequency and width.

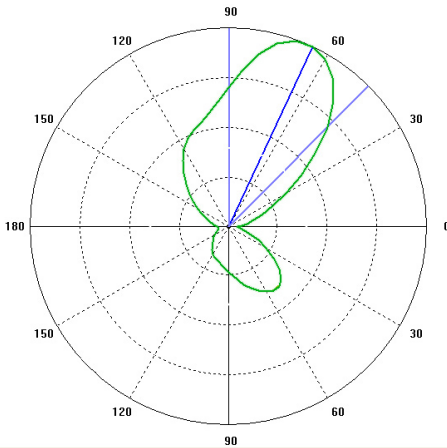


a)

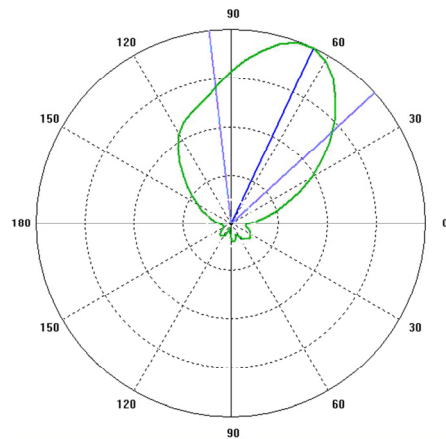


b)

Fig. 31. a) Layout of a bend tapered LWA. b) A prototype of bend half LWA Type III made using Roger 5880 RT/Duroid.



a)



b)

Fig. 32. a) The radiation patterns of E field of tapered LWA at $f = 8$ GHz. b) The radiation patterns of E field of bend tapered LWA at $f = 8$ GHz.

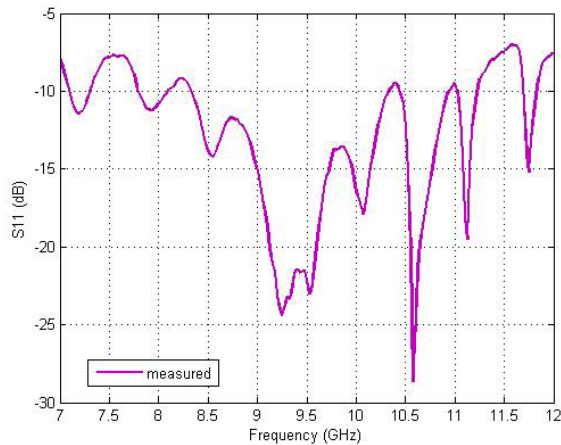


Fig. 33. Experimental return loss of half bend LWA Type III.

10. Tapered composite right/left-handed transmission-line (CRLH-TL) leaky-wave antennas (LWAs)

Recently composite right/left-handed (CRLH) leaky-wave antennas (LWAs) have been shown as one of the applications of the CRLH transmission line (TL) metamaterials thanks to their advantages of fabrication simplicity and frequency/electrically scanning capability without any complex feeding network. Nevertheless the fixed geometrical size of a unit cell of the CRLH-TL Leaky-wave antennas, prevents the possibility to improve the antenna bandwidth “tapering” the geometrical size of unit cell.

It is well known as a composite right/left-handed transmission-line (CRLH-TL) metamaterials, used for the leaky-wave antennas (LWAs) allow to obtain a superior frequency scanning ability than its conventional counterpart [26-27]. The leaky-wave antennas possess the advantages of low-profile, easy matching, fabrication simplicity, and frequency/electrically scanning capability without any complex feeding network.

However, the conventional leaky-wave antennas suffer from major limitations in their scanning capabilities. In fact the radiation pattern is restricted to strictly positive θ for uniform configurations, or to a discontinuous range of negative or positive θ excluding broadside direction, for periodic configurations. The CRLH LWAs have essentially suppressed these limitations, being able to scans the entire space from $\theta = -90^\circ$ to $\theta = +90^\circ$ and thereby paved the way for novel perspectives for leaky-wave antennas.

Although actually the designs of CRLH-TL for LWAs available in the literature, are developed as a different number of unit cell with a fixed geometrical size for all the unit cells of the entire antenna.

These design prevents the possibility to improve the antenna bandwidth “tapering” the geometrical size of unit cell. In order to obtain an improvement of the antenna bandwidth a novel design of CRLH LWAs was used in our work. The simulation results of the the CRLH

unit-cell with different size, obtained by a commercial 3D EM simulator has shown the good performance of this antenna compared with the performance of the uniform CRLH TL LWA antenna.

The good performance of this composite right/left-handed LWA are also demonstrated by measured results, which shown a good agreement with simulation results paving the way for the future applications of the antenna.

11. Antenna design

It has been shown that the leakage rate of the CRLH-TL LWA can be altered by using different sizes of the unit-cell [27] as shown in Fig.34.



Fig. 34. Different size and number of fingers of CRLH-TL unit cell.

In detail the radiation resistance, of the unit-cell having four fingers and the unit-cell of six fingers, both the unit-cells designed to have the phase origin at the same frequency, shows two different bandwidth as mentioned in [27-28].

The radiation resistance of the four finger antenna is always higher than that of the six-finger one, which implies faster decay of power (more leakage) along the structure for the former.

Moreover it should be noted that increasing the number of fingers the size of the unit-cell has to be reduced in order to have the same centre-frequency for the antenna antennas, otherwise, the centre-frequency for the antenna with unit-cell which have the larger number of fingers will shift down to a lower frequency [29-30].

As shown in [21-22] for a simple microstrip leaky wave antenna the radiation bandwidth is governed by the line width once the substrate is fixed. The bandwidth can be improved by adopting a tapered line structure (Fig.35), where, the radiation of different frequency regions leaks from different parts of the antenna.

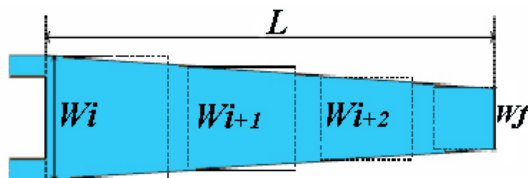


Fig. 35. A taper layout of LWA with a different frequency regions leaks from different parts of the antenna.

In fact from the propagation characteristics of the leaky wave antenna, we know that the leakage radiate phenomena, can only be noted above the cutoff frequency of higher order mode, and below the frequency such that, the phase constant is equal at the free space wave number. Decreasing the width of the antenna for a microstrip leaky-wave antenna the cutoff frequency increases shift toward high frequency. This behaviour allows to design a multisection microstrip LWA according [21-22] superimposing different section, in which each section can radiate in a different and subsequence frequency range, obtaining a broadband antennas. In this way each section should be into bound region, radiation region or reactive region, permitting the power, to uniformly radiated at different frequencies.

Following these idea in the our developed procedure we have applied a process to get the dimension of the physical parameters of the unit cell shows in Fig. 36 whit different optimized number of fingers (see Fig. 37). Naturally we have calculated the extraction parameters of every cell of CRLH implementation: LR, CR, LL, and CL using the equation mentioned in [26].

In the case of CRLH transmission line based LWA the amount of radiation by the unit cell can be related to the beam shape required and thus can be used to determine the total size of the structure as mentioned in [31].

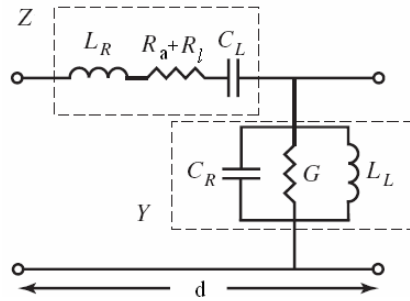


Fig. 36. Unit cell equivalent circuit with radiation resistance.

Given the unit cell equivalent circuit in Fig. 36, we have the per unit length series impedance and per unit length shunt admittance as follows [31]:

$$Z' = R_a' + R_l' + j\omega L_R' - \frac{j}{\omega C_L'} \quad (21)$$

$$Y' = G' + R_l' + j\omega C_R' - \frac{j}{\omega L_L'} \quad (22)$$

Where R_a' represents radiation resistance per unit length, R_l' represents the per unit length resistance associated with transmission loss, L_R' and C_R' denotes per unit length parasitic inductance and capacitance respectively, C_L' and L_L' denotes times unit cell length left-hand capacitance and inductance respectively.

Propagation constant and characteristics impedance are given by the following relations:

$$\gamma = \alpha + j\beta = \sqrt{Z'Y'} \quad (23)$$

$$Z_c = \sqrt{\frac{Z'}{Y'}} \quad (24)$$

From the above expression of propagation constant and line impedance we can find the centre frequency of the CRLH LWA [31].

These procedure was applied for subsequent frequency range of interest able to obtain a broadband antenna and a narrow-beam radiation pattern more than the uniform CRLH-TL LWA.

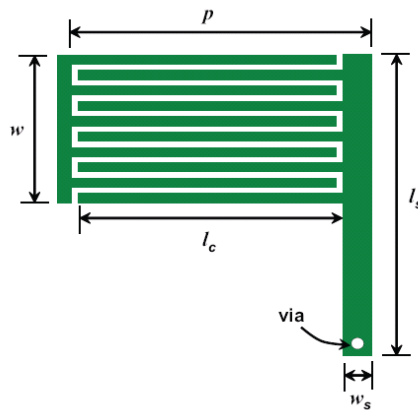


Fig. 37. Layout of a single unit cell of CRLH-TL LWA where p is the length of the unit cell period, l_c is the length of the capacitor w and w_s represent, the overall width of its finger and the width of the stub respectively.

The optimized antenna design as we can see in Fig. 38 was obtained considering the sequence of 16 cells composed respectively by 4 cells with 12 fingers, 4 cells with 10 fingers,

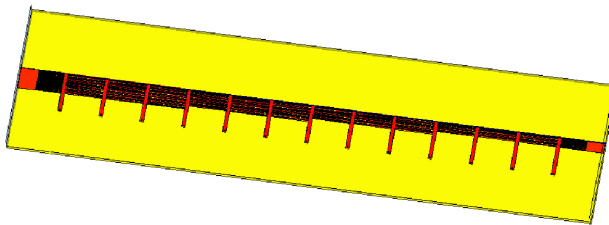


Fig. 38. Layout of the 16 unit-cell CRLH-TL LWA with cells of 10, 8 and 6 fingers.

4 cell with 8 fingers and 4 cell with 6 fingers for the entire length of the antenna of 207.55 mm and width between 2.9 mm (cells with 12 fingers) and 5.9 mm (cells with 6 fingers).

12. Simulation and experimental results

In the following Fig. 39 and Fig. 40 are showing the simulation data of the return loss obtained with a 3D EM commercial software, of the uniform 16 unit-cell CRLH-TL LWA of 10 fingers compared with the results of tapered 16 unit-cell CRLH-TL LWA. Instead in Fig. 41 and in Fig. 42, are shown the results of the radiation pattern of the uniform CRLH-TL LWA compared with 16 unit-cell of 10 fingers and tapered 16 unit-cell CRLH-TL LWA. It is evident the good performance of the tapered 16 unit-cell CRLH-TL LWA compared with uniform CRLH-TL LWA in term of broadband and narrowbeam.

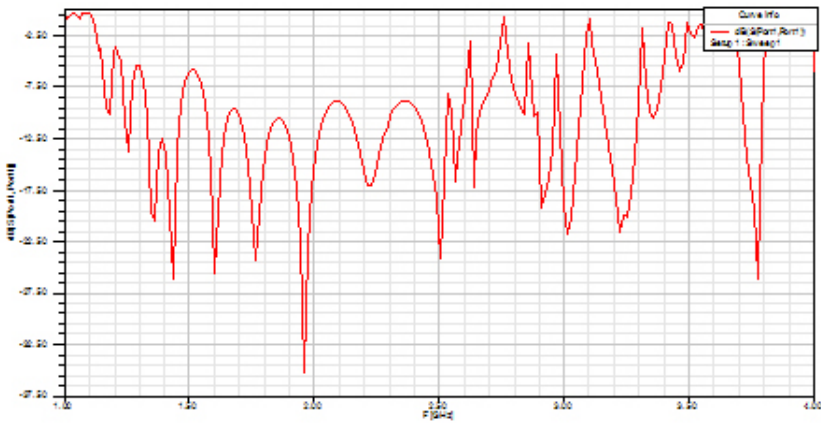


Fig. 39. Return loss (S11) of the uniform 16 unit-cell CRLH-TL LWA with 10 fingers.

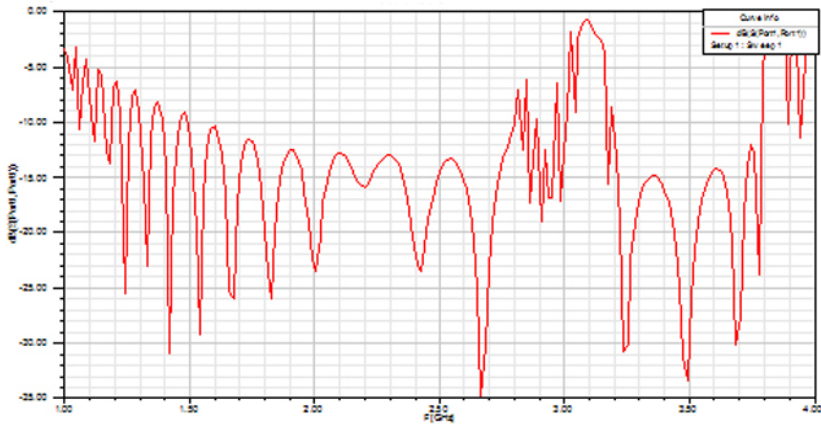


Fig. 40. Return loss (S11) of the 16 unit-cell CRLH-TL LWA with cells of 12, 10, 8 and 6 fingers.

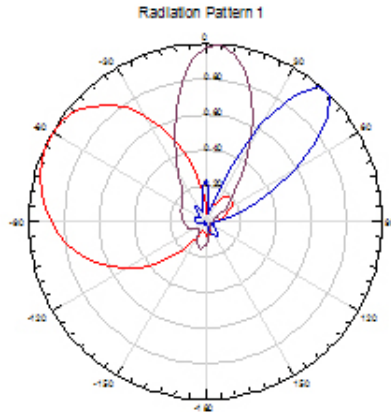


Fig. 41. Radiation pattern of the E field of the uniform CRLH-TL LWA for $f=1.12$ GHz (red line), $f=2.30$ GHz (brown line), $f= 3.20$ GHz (blu line).

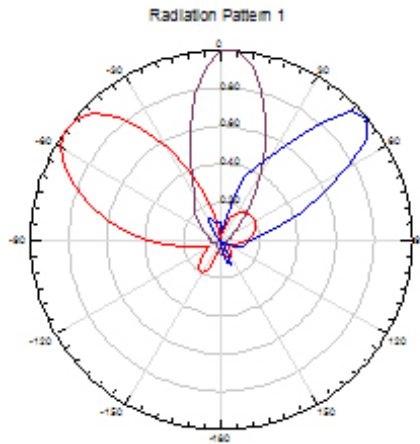


Fig. 42. Radiation pattern of the E field of the tapered CRLH-TL LWA for $f=1.12$ GHz (red line), $f=2.30$ GHz (brown line), $f= 3.20$ GHz (blu line).

The simulation results were compared with experimental results made on a prototype of CRLH-TL LWA (see Fig. 43) designed with 16 unit-cell on Rogers RT/duroid 5880 substrate with dielectric constant $\epsilon_r = 2.2$ and thickness $h = 62$ mil (loss tangent = 0.0009) showing a quite good agreement with simulated results of tapered 16 unit-cell CRLH-TL LWA as we can see in Fig. 44 and Fig. 45.



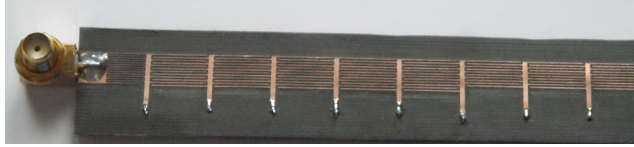


Fig. 43. A prototype and its detail of Radiation patter of tapered 16 unit-cell CRLH-TL LWA made on Rogers RT/duroid 5880.

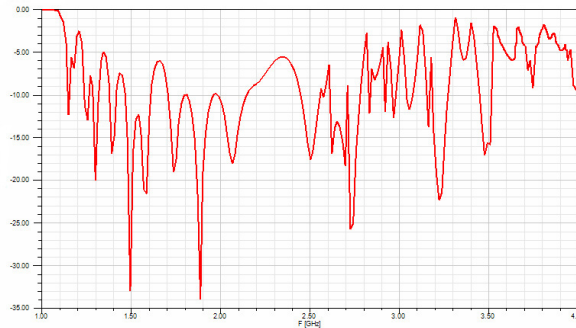


Fig. 44. Experimental return loss (S11) of the 16 unit-cell prototype CRLH-TL LWA with cells of 12, 10, 8 and 6 fingers.

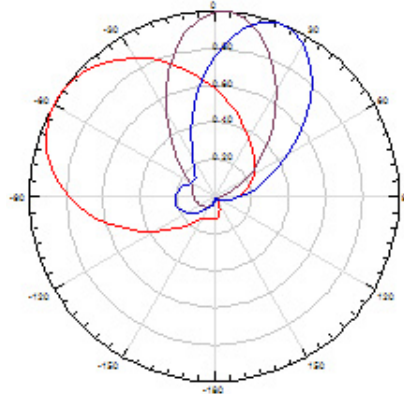


Fig. 45. Experimental E field radiation pattern of tapered prototype CRLH-TL LWA for $f=1.12$ GHz (red line), $f=2.30$ GHz (brown line), $f= 3.20$ GHz (blu line).

13. Meander antenna

Nowadays, miniaturization of electronic devices is the main request that productions have to fulfil. In this process, the reduction of the antenna size is the crucial challenge to face, being its dimensions related to the working frequency.

The rapid developing of the modern society has become to a crescent interest for the wireless communications. Nowadays, everybody wants to be connected everywhere

without the use of cumbersome devices. The current tendency goes towards portable terminals that have to be light and hand pocket. A suitable antenna for the portable terminals should be low cost, low profile, light weight and especially small size.

Printed antennas are commonly used for their simple structure and easy fabrication. As applications are space limited, it is challenging to design an antenna of small size but with simple tunable feature. For microstrip antennas, some techniques, such as making slots in their structure or using high dielectric constant substrates, can be used to reduce the antenna size. However, it results in the narrow bandwidths for its high Q factor and the low radiation efficiency.

In the following sentences is described an antenna printed on a substrate with low dielectric constant in order to get a reliable bandwidth. Moreover, the meander configuration allows us to reduce the antenna size keeping good radiation performance.

Meander dipole antennas have been already designed through numerical techniques that apply either time- or frequency-domain algorithms demanding high computational efforts and long-time processing. The common approach in the design of meander antenna is to draw a meander path with a suited length to the working frequency in commercial software and run simulations. Nevertheless, This empirical approach may lead to several consecutive trials and verifications. In order to decrease the long-time processing and avoid these cut and try methods, it would be convenient to start simulations with a commercial software having an antenna size close to its optimized dimensions. Thus, a good initial configuration can strongly affect the numerical convergence efficiency and the design process would be quicker.

This paper presents a transmission line model that provides an initial geometrical configuration of the antenna that allows us a computational improvement in the design of meander antennas. The dimensions obtained from the model have been used to run a simulation with a commercial software and an antenna resonating very close to that working frequency has been achieved. Finally, a quick optimization has been performed to definitely tune the antenna according to the ISM band.

14. TL model for meander antennas

Commercial and military mobile wireless systems demand for high compactness devices. An important component of any wireless system is its antenna. Whereas significant efforts have been devoted towards achieving low power and miniaturized electronic and RF components, issues related to design and fabrication of efficient, miniaturized, and easily integrable antennas have been overlooked. In this paper a novel approach for antenna miniaturization is presented. The meander topology is proposed as a good approach to achieve miniaturization and a transmission line model for the analysis and synthesis of meander antennas is developed.

Indeed, the miniaturization of an antenna can be accomplished through loads placed on the radiating structure. [32-33]. For example, monopoles were made shorter through center loaded (inductive) or top loaded (capacitive). Hence appropriate loading of a radiating element can drastically reduce the size, however, antenna efficiency may be reduced as well. To overcome this drawback, lumped elements of large dimensions can be created using

distributed reactive elements. For this reason, we propose a meander topology that allows us to distribute loading through short-circuited transmission lines.

In the past, meander structures were suitably introduced to reduce the resonant length of an antenna without great deterioration of its performances [34-36].

To exploit the meander topology to miniaturize printed antennas and develop a transmission line model for the analysis and the synthesis of this kind of antennas is proposed an antenna shown in Fig 46. It is a meander printed on the same side of the chassis of a circuit board on a FR4 substrate with $\epsilon_r = 3.38$ and thickness 0.787 mm. The feeding is between the meander structure and the ground plane. Even if the antenna is a printed monopole, it can be studied as an asymmetric dipole and its input reactance has been studied through a transmission line model. It is well known that the resonance condition is obtained when the input impedance is purely resistive [37-39]. The antenna has been modelled as a transmission line periodically loaded from inductive reactances X_m represented by the half meanders shown in Fig 46. We have named half-meander the shorted transmission line of length $w/2$.

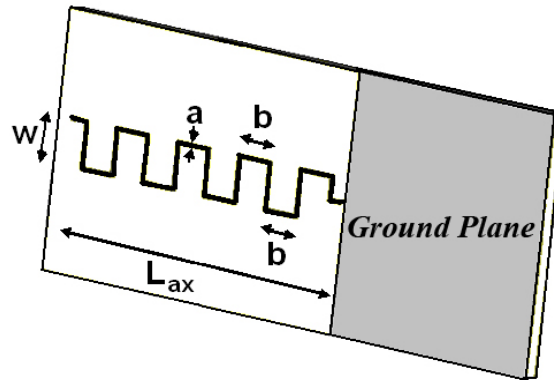


Fig. 46. Meander antenna monopole printed on the same side of the ground plane of the substrate.

The height b of each half meander and their total number $2n$ are related to the total length of the monopole L_{ax} with the following formula:

$$L_{ax} = (2n + 1)b \quad (25)$$

Each half meander was studied as a short transmission line $w/2$ long with a characteristic impedance Z_{cm} obtained as:

$$Z_{cm} = 120 \ln(b/a) \quad (26)$$

and terminating with a metallic strip having an inductance L_{sc} :

$$L_{sc} = 2 * 10^7 b [\ln(8b/a) - 1]. \quad (27)$$

The inductance L_{sc} of the strip with the length b and width a was substituted by a line L_{all} long terminating with a short circuit. The length L_{all} was properly chosen because this line had the same inductance of the strip (L_{sc}).

At the end, the inductance of each meander X_m was obtained by the formula (28):

$$X_m = Z_{cm} \tan g [2\pi \sqrt{\epsilon_{e,eff}} (w/2 + L_{all} - a/2)] \quad (28)$$

The total characteristic impedance of the transmission line with a length L_{ax} and loaded by $2n$ half meanders was:

$$Z_c = 120 [\ln(8L_{ax}/a) - 1] \quad (29)$$

In Fig. 47 the normalized length L_{ax} of the printed monopole versus the normalized thickness a according to the transmission line model for $w/b = 1$ is shown. It is pointed out from the figure that, when $b=w$, the meander length is smaller than the conventional monopole at its resonant frequency.

In Fig 48, for several values of the parameter $x=w/\lambda$, the resonant length L_{ax}/λ versus the ratio w/b is plotted.

It can be seen that, for each value of w/b , a remarkable reduction of the antenna length is obtained by increasing the values of w at the resonant frequency. Therefore, by choosing a value of w/b , the model allows to detect the correspondent resonant length of the antenna.

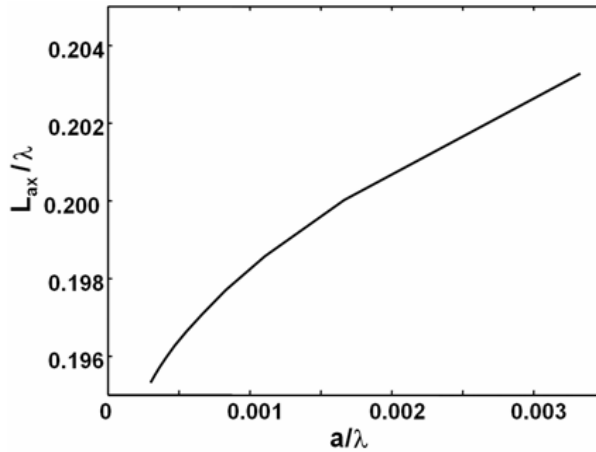


Fig. 47. The normalized length L_{ax} of the printed monopole versus the normalized thickness a according to the transmission line model for $w/b = 1$.

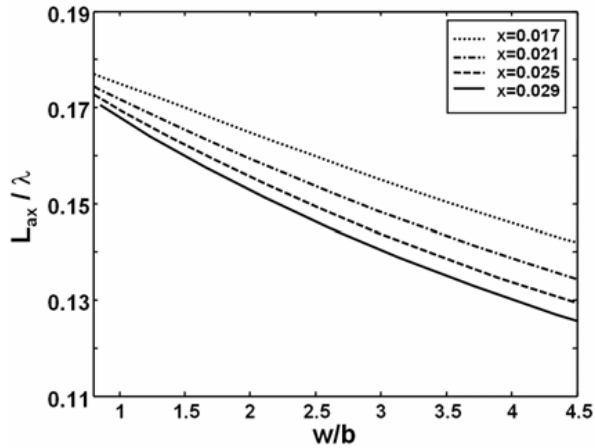


Fig. 48. Transmission line model for meander antenna printed on substrate with $\epsilon_r = 3.38$ and $s = 0.81$ mm for different $x = w/\lambda$.

15. Simulated and experimental results

To test the validity of the model, simulations were run with full wave commercial software CST Microwave Studio © at a frequency of 2.45 GHz with appropriate model as shows in Fig 49.

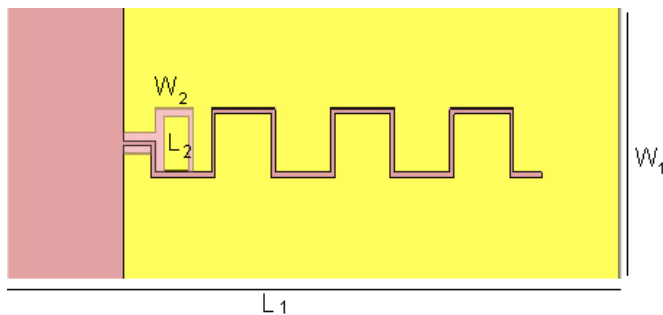


Fig. 49. Top and bottom model of meander antenna monopole designed with CST Microwave Studio ©.

The transmission line model (TLM) and the full wave simulation were in a good agreement and the difference between the resonant length obtained by TLM and the full wave

simulations was within 250 MHz as it has been summarised in Table 2. Figs 47 and Fig. 48 show, respectively, the normalized length L_{ax} versus antenna thickness a for different values of w/b and the normalized length L_{ax} versus w/b for different values of $x=w/\lambda$.

		w/b=1.4					
		model		1° run		2° run	
w/λ	n	L_{ax}/λ	fr sim.	L_{ax}/λ	fr sim.	L_{ax}/λ	fr sim.
		[GHz]		[GHz]		[GHz]	
0.029	3	0.158	2.26	0.137	2.54	0.143	2.448
0.025	4	0.177	2.13	0.154	2.40	-	-
0.021	5	0.185	2.12	0.162	2.37	0.157	2.430
0.017	6	0.180	2.23	0.159	2.49	-	-

		w/b=1					
		model		1° run		2° run	
w/λ	n	L_{ax}/λ	fr sim.	L_{ax}/λ	fr sim.	L_{ax}/λ	fr sim.
		[GHz]		[GHz]		[GHz]	
0.029	2	0.155	2.43	-	-	-	-
0.025	3	0.189	2.14	0.160	2.446	-	-
0.021	3	0.160	2.45	-	-	-	-
0.017	4	0.169	2.41	-	-	-	-

Table 2. Resonant frequencies calculated with FIT method.

The antenna sizes derived from the model allow us to obtain a design very close to the final project which can be quickly optimized by avoiding long simulations with commercial software.

Table 2 shows that the antenna sizes derived from the model allow us to get antenna sizes close to the final structure as the antenna resonates almost at 2.45 GHz. Moreover, in order to get exactly 2.45GHz, a quick optimization has to be carried out by running few simulations with a commercial software.

To validate the proposed TLM method, simulations and measurements have been performed. The antenna has been printed on a Rogers R04003C with $\epsilon_r = 3.38$ and thickness 0.81 mm. A prototype is presented in Fig 50. The geometrical sizes chosen were $a=0.5$ mm, $b=8$ mm and $w=b$ that has led to a length $L_{ax} = 56$ mm by considering 6 half meanders (Fig 50). The board total size is $L1=72$ mm and $W1=32$ mm, by considering also the chassis. The antenna is fed by a microstrip printed on the back of the chassis by terminating with a stub for achieving good matching. The microstrip is 20mm length and the stub is $L2=8$ mm and $W2= 4$ mm. The dimensions of the microstrip line has been optimized using full wave software to provide better impedance matching for the frequency antenna-resonance.

The simulated and measured return loss is shown in Fig 51. Simulation has been performed by using CST Microwave Studio © 2009 and it has shown a value of -44dB at 2.45 GHz.

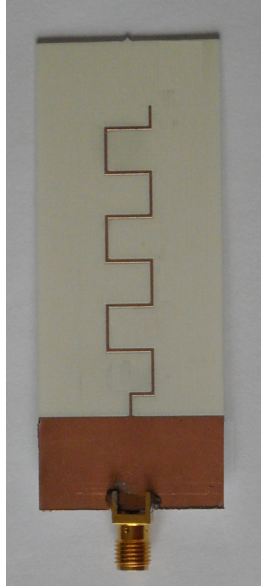


Fig. 50. Meander antenna monopole printed on the Rogers R04003C substrate.

The measurements were carried out in an anechoic chamber by connecting the antenna at a network analyser through coaxial cables.

The measured return loss in Fig 51 shows a slight shift of the antenna resonant frequency towards lower frequencies from 2.45 GHz to 2.42 GHz. Nevertheless, a good matching is still observed because the reflection coefficient assumes the value -26 dB instead of -44 dB.

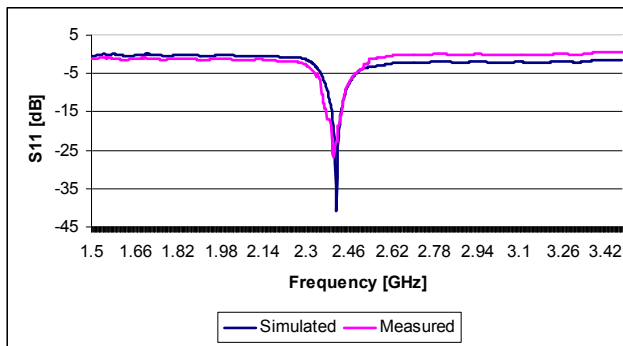
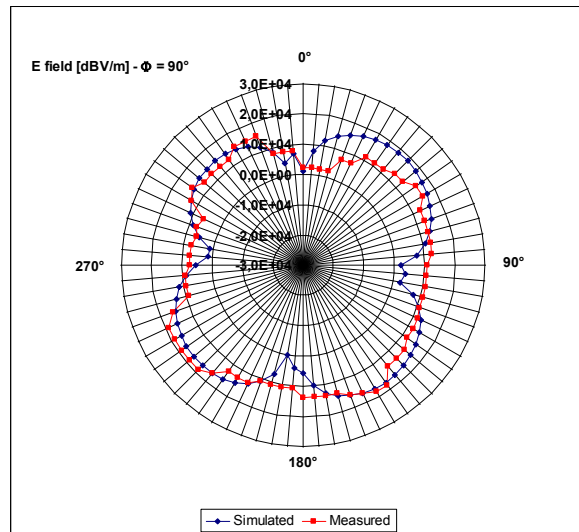
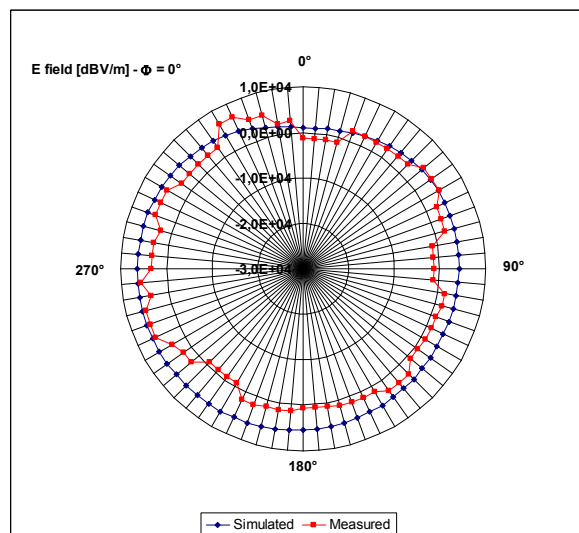


Fig. 51. Comparison of S11 simulation results with measured results.

Fig. 52 shows the E field radiation patterns of the antenna at 2.45 GHz on two principal planes, xz plane ($\Phi=0^\circ$) and yz plane ($\Phi=90^\circ$). The comparison of the radiation patterns shows that simulations and measurements are in a good agreement.



a)



b)

Fig. 52. Comparison of measured and simulated E-field at 2.45 GHz for a) $\Phi = 0^\circ$ and b) $\Phi = 90^\circ$.

Fig 53 shows the current distribution on the antenna. It can be observed that the current is particularly intense at the end of each half meander. Full wave simulations confirm that each half meander can be studied as a transmission line terminating in a short circuit.

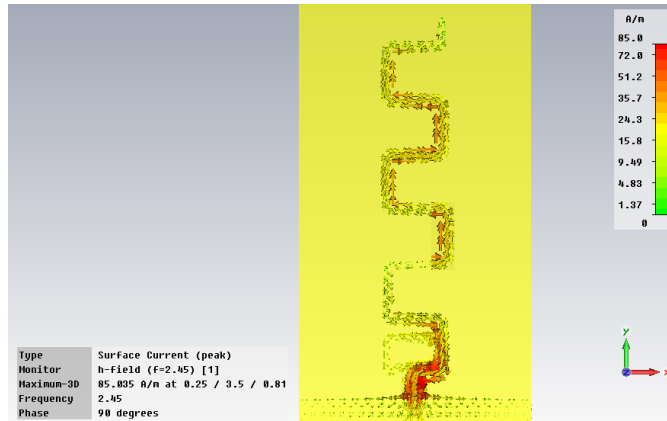


Fig. 53. Current distribution on the meander antenna.

16. Reference

- [1] T. Tamir, \Leaky-wave antennas", ch. 20 in *Antenna Theory, Part 2*, R. E. Collin and F. J. Zucher, Eds., McGraw-Hill, New York, 1969.
- [2] A. A. Oliner, \Leaky-wave antennas", ch. 10 in *Antenna Engineering Handbook*, 3rd ed., R. C. Hansen, Ed., McGraw-Hill, New York, 1993.
- [3] C. H. Walter, \Traveling Wave Antennas", McGraw-Hill, New York, 1965.
- [4] T. Tamir, A. A. Oliner, \Guided complex waves, part I: fields at an interface", *Proc. Inst. Elec. Eng.*, vol. 110, pp. 310-324, Feb. 1963.
- [5] T. Tamir, A. A. Oliner, \Guided complex waves, part II: relation to radiation patterns", *Proc. Inst. Elec. Eng.*, vol. 110, pp. 325-334, Feb. 1963.
- [6] L. O. Goldstone and A. a. Oliner, \Leaky-wave antennas I: rectangular waveguide", *IRE Trans. Antennas and Propagation*, vol. AP-7, pp. 307-319, Oct. 1959.
- [7] A. Hessel, \General characteristics of traveling -wave antennas", ch. 19 in *Antenna Theory, Part 2*, R. E. Collin and F. J. Zucher, Eds., McGraw-Hill, New York, 1969.
- [8] G. Gerosa and P. Lampariello, \Lezioni di Campi Elettromagnetici I", Edizioni Ingegneria 2000, 1995.
- [9] F. Frezza, *Lezioni di Campi Elettromagnetici II*, March 2004.
- [10] Pozar, David M. and David H. Schaubert, \Microstrip Antennas: The Analysis and Design of Microstrip Antennas and Arrays", John Wiley, New York, NY, 1995.
- [11] Pozar, David M., \Microwave Engineering", John Wiley, New York, NY, second edition, 1998.
- [12] Kumar, Girish and K. P. Ray, \Broadband Microstrip Antennas", Artech House, Boston, MA, 2003.
- [13] Menzel Wolfgang, \A New Travelling-Wave Antenna in Microstrip", *Archiv fur Elektronik und Ubertragungstechnik (AEU)*, Band 33, Heft 4, pp. 137-140, April 1979.
- [14] Yau, D., N. V. Shuley, and L. O. McMillan, \Characteristics of Microstrip Leaky-Wave Antenna Using the Method of Moments", *IEE Proc. Microwave Antennas and Propag.*, Vol. 146, No. 5, pp. 324-328, October 1999.

- [15] Lee, Kun Sam, \Microstrip Line Leaky Wave Antennas", Ph.D. thesis, Polytechnic Institute of New York, 1986.
- [16] N. Marcuvitz, \On ω -eld representations in terms of leaky waves or Eigenmodes", IRE Transactions on Antenna and Propagation, Vol. AP-4, pp. 192-194, July 1956.
- [17] T. Itoh, R. Mittra, \Spectral domain approach for calculating the dispersion characteristics of microstrip lines", IEEE Trans. Microwave Theory Techn., Vol. MTT-21, pp. 496-499, 1973.
- [18] Mesa, Francisco and David R. Jackson, \Investigation of Integration Paths in the Spectral Domain Analysis of Leaky Modes on Printed Circuit Lines", IEEE Trans. Microwave Theory Techn., Vol.50, No. 10, pp. 2267-2275, October 2002.
- [19] Kuester, Edward F., Robert T. Johnk, and David C. Chang, \The Thin-Substrate Approximation for Reection from the End of a Slab-loaded Parallel-Plate Waveguide with Applications to Microstrip Patch Antennas", IEEE Transactions on Antennas and Propagation, Vol. AP-30, No.5, pp. 910-917, September 1982.
- [20] A. Oliner, "Leakage from higher modes on microstrip line with application to antennas," Radio Scienze, Vol. 22, pp. 907-912, 1987.
- [21] O. Losito, "A New Broadband Microstrip Leaky-Wave Antenna" *Applied Computational Electromagnetics Society Journal*, Vol.23, n.3, Pg. 243-248 September 2008.
- [22] O. Losito, "A Simple Design of Broadband Tapered Leaky-Wave Antenna" *Microwave and Optical Technology Letters*, vol. 49, pp.2833-2838, 2007.
- [23] W. Hong, T. L. Chen, C. Y. Chang, J. W. Sheen, Y. D. Lin "Broadband Tapered Microstrip Leaky-Wave Antenna", IEEE Trans. Antennas and Propagation, Vol. 51, pp. 1922-1928, 2003.
- [24] Y. Qian, B. C. C. Chang, T. Itoh, K. C. Chen and C. K. C. Tzuang, "High Efficiency and Broadband Excitation of Leaky mode in Microstrip Structures", *IEEE MTT-S Microwave Symposium Digest*, vol. 4, pp. 1419-1422, 1999.
- [25] P. Burghignoli, F. Frezza, A. Galli, G. Schettini "Synthesis of broad-beam patterns through leaky-wave antennas with rectilinear geometry" *IEEE Antennas and Wireless Prop. Letters.*, vol. 2, pp. 136-139, 2003.
- [26] Christophe Caloz, Tatsuo Itoh, 'Electromagnetic Metamaterials: transmission line theory and microwave applications', Wiley-IEEE Press December 2005, Chapter 3, pp. 122-124.
- [27] A. Rahman, Y. Hao, Y. Lee and C.G. Parini, "Effect of unit-cell size on performance of composite right/left-handed transmission line based leaky-wave antenna", *Electronics Letters*, 19th June 2008 Vol. 44 No. 13.
- [28] C. Caloz and T. Itoh, "Novel microwave devices and structures based on the transmission line approach of meta-materials", in *IEEE-MTT Int. Symp. Dig.*, June 2003, pp. 195 - 198.
- [29] C. Caloz, I. Lin, and T. Itoh, "Orthogonal anisotropy in 2-D PBG structures and metamaterials," in *IEEE-APS Int. Symp. Dig.*, vol. 1, June 2003, p. 199.11.
- [30] C. Caloz and T. Itoh, "Application of the transmission line theory of lefthanded (LH) materials to the realization of a microstrip LH transmission line", in *IEEE-APS Int. Symp. Dig.*, vol. 2, June 2002, pp. 412 - 415.
- [31] A. Rahman, Y. Lee, Y.Hao and C. G. Parini "Limitations in bandwidth and unit cell size of composite right-left handed transmission line based leaky-wave antenna". *Proceeding of EuCAP 2007* 11 - 16 November 2007, EICC, Edinburgh, UK.

- [32] L.C. Godara, Handbook on antennas in wireless communications, CRC Press, Boca Raton, FL, 2002, Ch. 12.
- [33] C.W.Harrison, "Monopole with inductive loading", *IEEE Trans. Antennas Propagation* Vol AP-11, pp 394-400, 1963.
- [34] R.C. Hansen, "Efficiency and matching tradeoffs for inductively loaded short antennas", *IEEE Trans. Commun.*, vol. COM-23, pp 430-435, 1975.
- [35] H.Nakano, H. Tagami, A.Yoshizawa, and J. Yamauchi, "Shortening ratio of modified dipole antennas", *IEEE Trans. Antennas Propagation*, Vol. AP-32, pp. 385-386, 1984.
- [36] J. Rashed and C.Tai, "A new class of resonant antennas", *IEEE Trans. Antennas Propagation*, Vol 39, Sett. 1991.
- [37] C. T. P. Song, Peter S. Hall, and H. Ghafouri-Shiraz, "Perturbed Sierpinski Multiband Fractal Antenna With Improved Feeding Technique", *IEEE Transactions On Antennas And Propagation*, Vol. 51, No. 5, May 2003.
- [38] R.P. Clayton, *Compatibilità Elettromagnetica*, *Ulrico Hoepli* Milano, 1992.
- [39] R.K. Hoffmann, "Handbook of Microwave Integrated Circuits", Artech House, Norwood, MA, 1987.

Superstrate Antennas for Wide Bandwidth and High Efficiency for 60 GHz Indoor Communications

Hamsakutty Vettikalladi, Olivier Lafond and Mohamed Himdi
Institute of Electronics and Telecommunication of Rennes (IETR)
University of Rennes 1
France

1. Introduction

Modern multimedia applications demand higher data rates and the trend towards wireless is evident, not only in telephony but also in home and office networking and customer electronics. This has been recently proven by the accelerating sales of IEEE 802.11 family WLAN hardware. Current WLANs are, however, capable of delivering only 30-100 Mb/s connection speeds, which is insufficient for future applications like wireless high-quality video conferencing, multiple simultaneous wireless IEEE 1394 (Firewire) connections or wireless LAN bridges across network segments. For these and many other purposes, more capacity – wirelessly – is needed. Service provided by IEEE 802.11 WLANs fulfills casual internet users and office workers actual needs. But, bandwidth demands are still rising. Millimetre-wave technology is one solution to provide up to multi-Gbps wireless connectivity for short distances between electronic devices. The data rate is expected to be 40-100 times faster than today's wireless LAN systems, transmitting an entire DVD's data in roughly 15 seconds. 60 GHz is ideally suited for personal area network (PAN) applications. A 60 GHz link can replace various cables used today in the office or in home by wireless link as shown in Fig.1, including gigabit Ethernet (1000Mbps), USB 2.0 (480Mbps), or IEEE 1394 (~800Mbps). Currently, the data rates of these connections have precluded wireless links, since they require so much bandwidth. While other standards are evolving to address this market (802.11n and UWB), 60 GHz is another viable candidate. In such a context, 60 GHz millimeter wave (MMW) systems constitute a very attractive solution due to the fact that there is a several GHz unlicensed frequencies range available around 60 GHz, almost worldwide. In Europe, the frequency ranges 62 - 63 GHz and 65 - 66 GHz are reserved for wideband mobile networks (MBS, Mobile Broadband System), whereas 59 - 62 GHz range is reserved for wideband wireless local area networks (WLAN). In the USA and South Korea, the frequency range 57 - 64 GHz is generally an unlicensed range. In Japan, 59 - 66 GHz is reserved for wireless communications (Nesic et al., 2001). This massive spectral space enables densely situated, non-interfering wireless networks to be used in the most bandwidth-starving applications of the future, in all kinds of short-range (< 1 km) wireless communication. Also in this band, the oxygen absorption reaches its maximum value (10-15 dB/km), which gives an additional benefits of reduced co-channel interference. Hence, it is a

promising candidate for fulfilling the future needs for very high bandwidth wireless connections. It enables up to gigabit-scale connection speeds to be used in indoor WLAN networks or fixed wireless connections in metropolitan areas.



Fig. 1. Short range communication.

These new systems will need compact and high efficient millimeter wave front-ends including antennas. For antennas, printed solutions are often demanding for the researchers because of its low profile, lightweight and ease of integration with active components (Zhang et al., 2006). High gain and high efficient antennas are needed for 60 GHz communication due to high path losses at this range of frequencies. Conventional antenna arrays are used for high gain applications. But in these cases for achieving high gain, a large number of elements are needed, which not only increases the size of the antenna but also decreases its efficiency (Lafond et al., 2001), (Kärnfelt et al., 2006) & (Soon-soo oh et al., 2004). It has been reported that for high gain, a superstrate layer can be added at a particular height of $0.5 \lambda_0$ above the ground plane (Choi et al., 2003), (Menudier et al., 2007) & (Meriah et al., 2008).

2. Superstrate antenna technology

In this chapter the authors are explaining how to develop a wideband, high gain and high efficient antenna sufficient for 60 GHz communications using superstrate technology. Also explains the importance of different sources on antenna performance in terms of bandwidth, gain and efficiency.

2.1 Microstrip fed parasitic patch antenna with superstrate

Here the antenna configuration consists of a microstrip feed, patch and a parasitic patch, as the source, which are loaded by a superstrate. Fig. 2 shows the 3D view (a) and side view (b)

of the microstrip fed stacked patch antenna with superstrate. It consists of a lower patch with an optimised dimension of 1.63 mm x 1.6 mm on a substrate RT Duroid 5880 ($\epsilon_r = 2.2$, $t_1 = 0.127$ mm). The upper patch with an optimised dimension of 1.63 mm x 1.63 mm is printed on the lower side of a parasitic substrate RT Duroid 5880 ($\epsilon_r = 2.2$, $t_2 = 0.254$ mm).

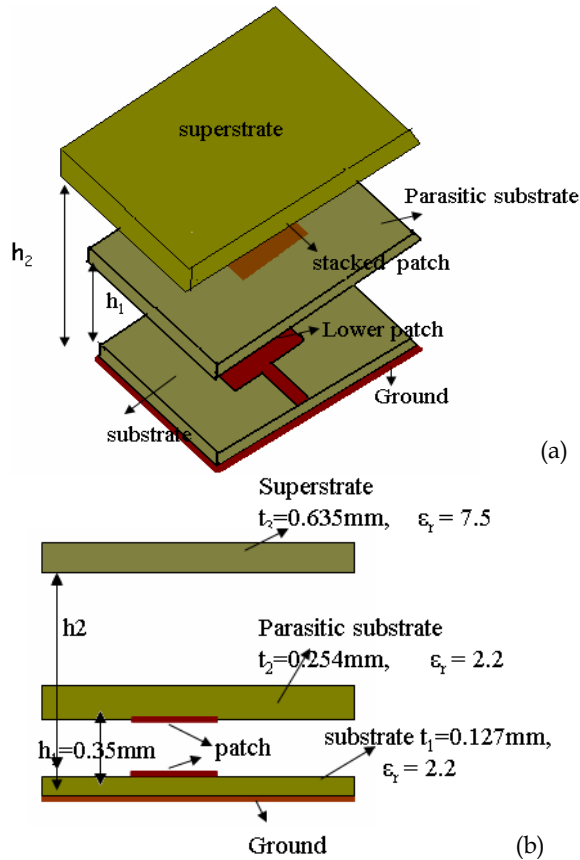
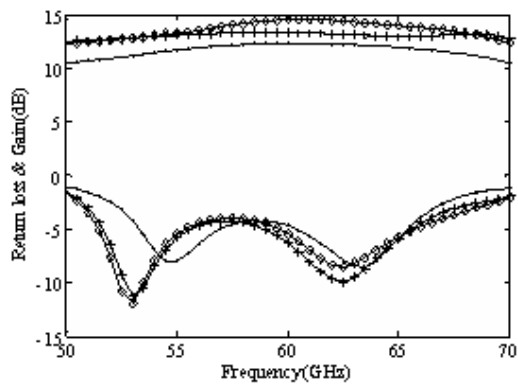
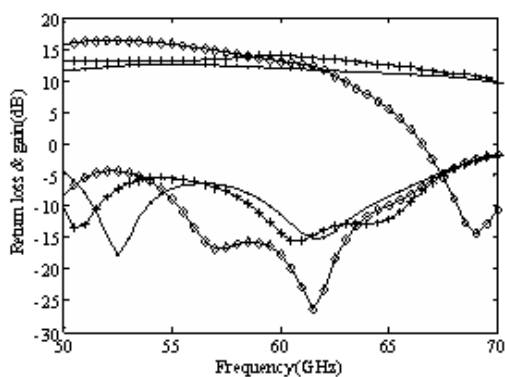


Fig. 2. Cutting plane of stacked patch antenna with superstrate.

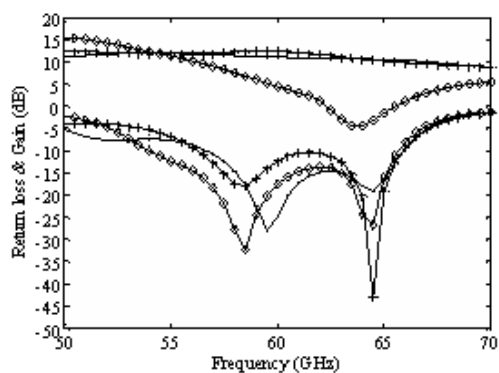
The distance between the lower patch and the upper patch is optimized, by simulation using CST Microwave Studio®, as $h_1 = 0.35$ mm for a resonance at 60 GHz for a larger bandwidth and gain. This antenna is then loaded with superstrate. The material used for the superstrate is Roger substrate RT6006 ($\epsilon_r = 7.5$ at 60 GHz, $t = 0.635$ mm). The dimension of the superstrate and the height from the ground plane are optimized as explained below. The variation of gain and VSWR bandwidth with the variation of superstrate dimension ($0.73\lambda_0$, $1.1\lambda_0$, $2\lambda_0$) for different heights ($0.5\lambda_0$, $0.6\lambda_0$, $0.7\lambda_0$) is shown in Figs. 3 (a-c). It is noted that the maximum gain with good bandwidth is achieved for a superstrate dimension of $1.1\lambda_0$ with a height = $0.6\lambda_0$, and is equal to 13.6 dB with almost flat over a frequency range of 59 GHz to 64 GHz (Vettikalladi et al., 2009). In all other cases either the gain is less than the above value or the VSWR bandwidth is poor. Also noted that when the superstrate dimension is higher than $1.1\lambda_0$, the gain goes down when the height h_2 varies from $0.5\lambda_0$ to $0.7\lambda_0$.



(a)



(b)



(c)

Fig. 3. variation of s_{11} and gain with superstrate dimension and height from ground plane.
 a) $h_2 = 0.5 \lambda_0$ b) $h_2 = 0.6 \lambda_0$ c) $h_2 = 0.7 \lambda_0$ for size = $.73 \lambda_0$ —●— size = $1.1 \lambda_0$ —+—
 size = $2 \lambda_0$ —◇— .

From the literature (Gupta et al., 2005), the theoretical height between the superstrate and ground plane is $0.5\lambda_0$, but in this work it is found to be $0.6\lambda_0$ for maximum gain, it may be due to the stacked patch. Fig. 4 shows the return loss, simulated directivity with theoretical values, and simulated and measured gain of the prototype with superstrate. It is found that there is a gain reduction in the measurement, which is due to the variation of exact heights from the theoretical values as shown in Table I.

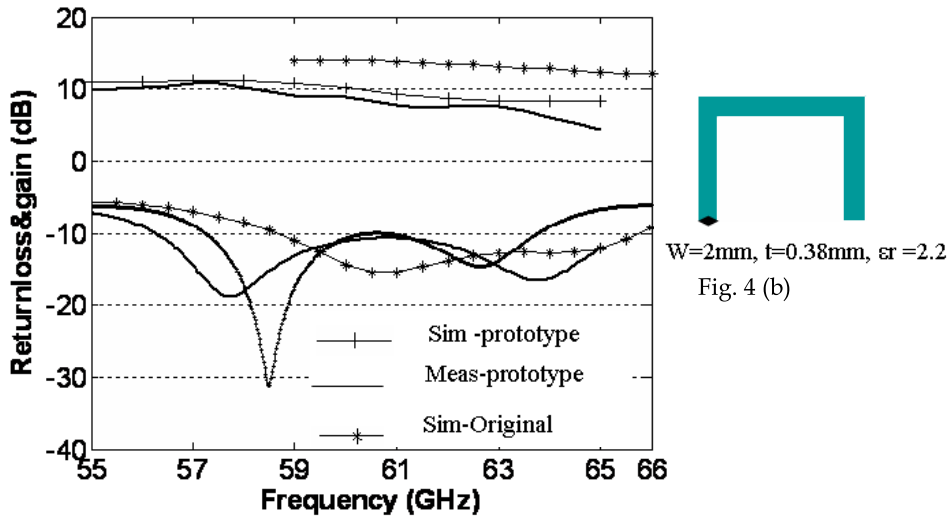


Fig. 4. (a) Comparison of return loss and gain with superstrate.

	Original (mm)	variation while implementation (mm)
Lower Patch	1.6x1.63	1.62x1.65
Upper Patch	1.63x1.63	1.65x1.65
h1	0.35	0.38
h2	3	3.48

Table I. Variation of prototype parameters from exact values.

It is very difficult to maintain the exact thickness h_1 , hence we inserted a substrate cut in the form of a rectangular U shape (Fig. 4(b)), with width 2mm, thickness 0.38mm and permittivity 2.2. Also the thickness h_2 is varied to 3.48mm instead of 3mm ($0.6\lambda_0$) and hence is the reason for the reduction of gain to nearly 10 dB. The E and H planes radiation patterns at 57 GHz and 58 GHz are shown in Figs. 5(a-b). The radiation patterns are found to be broad. There is a cross polar level of less than -20 dB on both the planes. The measured half-power beam widths are found to be 37° for E and H planes at 57 GHz, and 38° and 41° for E and H planes respectively at 58 GHz.

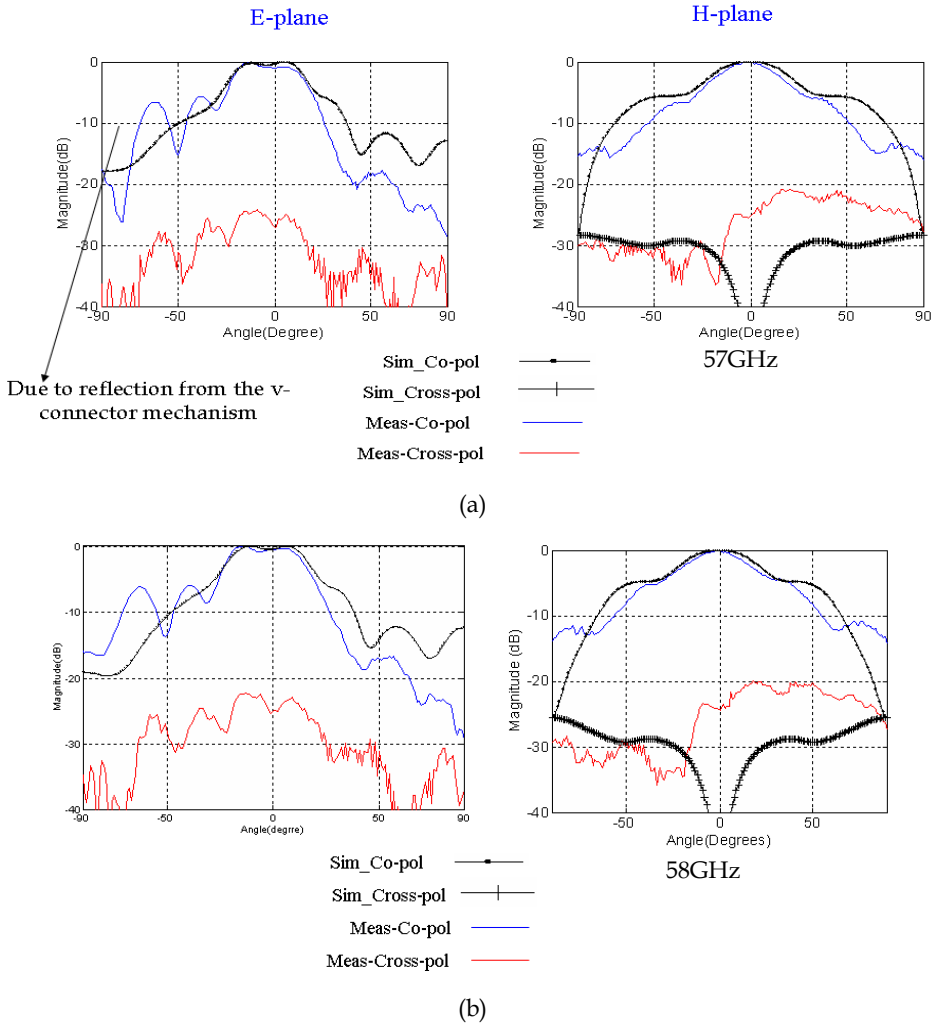


Fig. 5. Measured and simulated E-plane & H-plane radiation patterns of the parasitic patch Superstrate antenna (a) 57 GHz, (b) 58 GHz.

It is noted that for microstrip fed stacked patch antenna, the optimized superstrate size is $1.1 \lambda_0$ for getting maximum gain and broad pattern. This value is considered as the limitation of size in this case. It is also observed from Fig. 3, that when the superstrate size is higher than $1.1 \lambda_0$, and when the height varies from $0.5 \lambda_0$ to $0.7 \lambda_0$ the broad nature of the gain decreases and starts coming down at 60 GHz .I.e. the pattern changes from broadside to sectorial and then to conical for different frequencies in the band as shown in Fig. 6 (for a superstrate size of $2 \lambda_0$ & $h_2=0.6 \lambda_0$), which may suitable for some other application (Vettikalladi et al., 2009b). Here, the small superstrate size is due to the presence of the parasitic patch that disturbs the field in the cavity (thickness = $0.6 \lambda_0$).

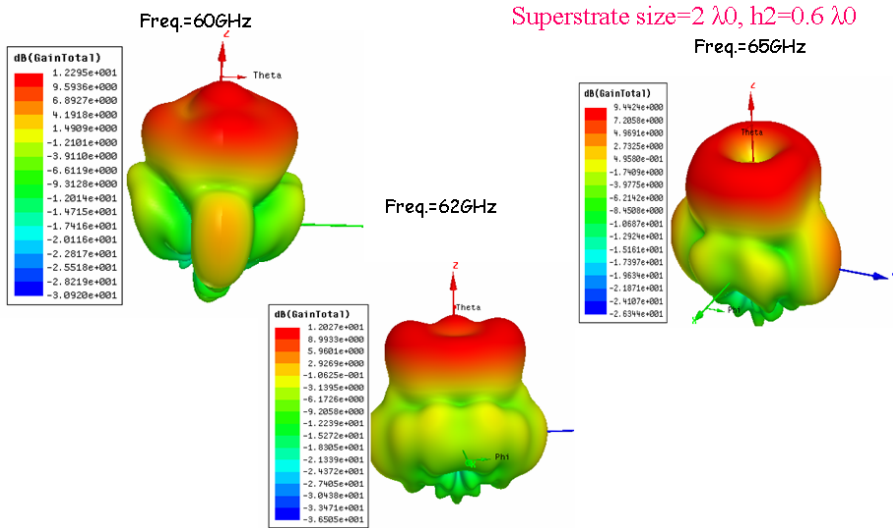


Fig. 6. Gain pattern for a microstrip fed parasitic patch superstrate with a superstrate size = $2\lambda_0$, for different frequencies in the band.

Since this kind of prototype is very difficult to manufacture and hence we are going to discuss with other kind of technology.

2.2 Slot coupled superstrate antenna

In this section, we are explaining a superstrate antenna with aperture coupled source as the excitation. We are also showing the importance of the size of the superstrate for getting maximum gain and also for getting consistent radiation pattern all over the frequency range of interest. Fig.7 shows the side view and the 3D view of a slot coupled patch antenna with superstrate. The slot is optimised to $0.2\text{ mm} \times 1\text{ mm}$ for maximum coupling with a stub length of 0.75 mm . In order to consider the easiness of implementation; we used a thick ground plane of thickness $t=0.2\text{ mm}$. The antenna consists of a patch with optimised dimension $1.3\text{ mm} \times 1.3\text{ mm}$ on a substrate RT Duroid 5880 of permittivity 2.2 and a loss tangent $\tan\delta = 0.003$ with a thickness $t_1 = 0.127\text{ mm}$. Low thickness and low permittivity substrate are used for reducing surface waves. A dielectric superstrate is added above the slot coupled patch antenna (Vettikalladi et al., 2009a). Here we used only one layer to avoid the technological manufacturing problems when many layers are used at 60 GHz. The material used for the superstrate is Roger substrate RT6006 with a relative permittivity of 7.5 at 60 GHz. Theoretically the thickness of superstrate must be $\lambda_g/4$ (0.456 mm), but here we took the thickness ($t_2 = 0.635\text{ mm}$) close to the theoretical thickness available in market for good antenna performance. The distance between the superstrate and ground plane is $0.5\lambda_0$ as per the theory (Gupta & Kumar, 2005). A Rohacell foam layer of permittivity 1.05 is sandwiched between base antenna and superstrate to fix all the layers.

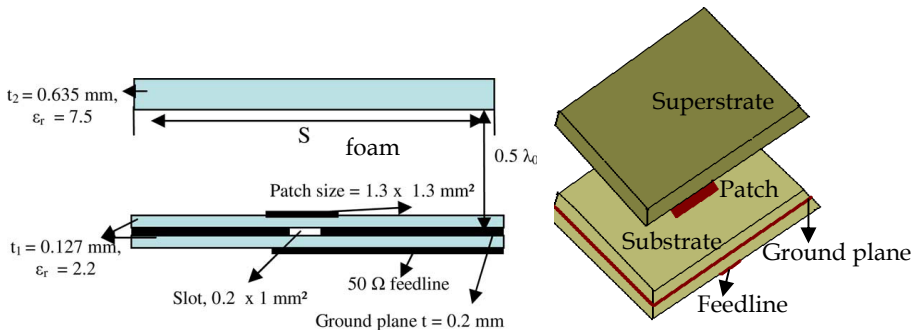


Fig. 7. Cutting plane and 3D view of aperture coupled antenna with superstrate, ground plane size = 30 x 30 mm².

Usually in all the known superstrate antennas large superstrates are used for improving the gain which not only increases the size of the antenna but also decreases the S11 bandwidth. But our objective is different, we want to use a small superstrate for obtaining high stable gain and constant radiation pattern all over the frequency band of interest. To study the effect of superstrate size 'S' and hence to optimize, we considered four square sizes ($1 \lambda_0$, $2 \lambda_0$, $4 \lambda_0$ and $6 \lambda_0$). Simulations are done using CST Microwave studio®. Fig. 8 shows the CST results of S11 and gain variations of the slot coupled antenna without superstrate and with varying superstrate size. It is observed that the S11 and gain vary with various size of the superstrate. When there is no superstrate, the antenna radiates at 60 GHz with a bandwidth of 3.7% over a frequency range of 58.9 to 61.1 GHz with a maximum gain of 5.9 dBi. It is noted that with superstrate the gain is highest for a superstrate size of $2 \lambda_0$. The 2:1 VSWR bandwidth is noted to be BW = 58.7 - 62.7 GHz i.e. 6.7% with a maximum gain of 14.9 dBi. It is also noticed that the gain decreases when the size of the superstrate is above or below $2 \lambda_0$.

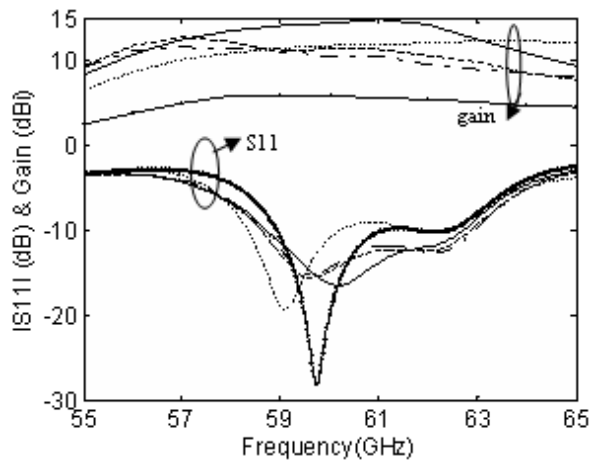


Fig. 8. Variation of S11 and gain without superstrate and with various superstrate dimensions. without superstrate ——— $1 \lambda_0$ $2 \lambda_0$ — — — $4 \lambda_0$ - - - - $6 \lambda_0$ - - - - .

There is a gain enhancement of 9 dB with the superstrate. Fig. 9 shows the comparison of measured and simulated S11 and gain for the optimised superstrate size of $2\lambda_0$. Table II gives the comparison of measured and simulated S11 and gain for the optimised superstrate size. It is noted in S11 that there is a frequency band shift of 2.8% (1.7 GHz), when a V-connector is used and a frequency band shift of 1.5% when a V-band test fixture is used. These frequency shifts are maybe due to the combined effect of connectors and the inaccuracy of the distance between patch and superstrate for the experimental prototype.

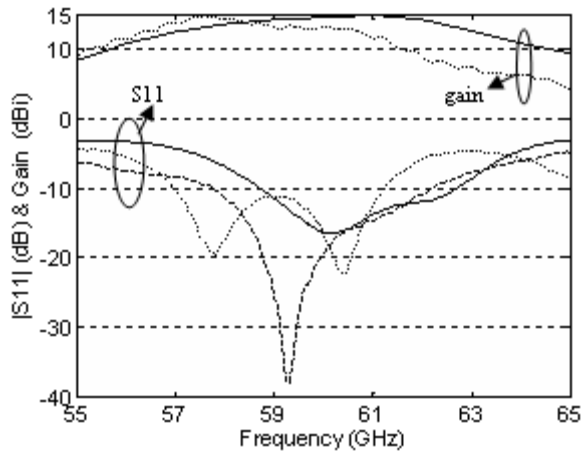


Fig. 9. Variation of S11 and gain with a superstrate dimension of $2\lambda_0$. Simulation — ; measured with V coaxial mounting connector - - - - ; measured with V test fixture - . - .

Return loss bandwidth (simulated)	Return loss bandwidth (measured)	Maximum Gain (simulated)	Maximum Gain (measured)	Efficiency Estimated η
58.7 - 62.7 GHz (6.7%)	57 - 61.1 GHz (6.8%)	14.9 dBi	14.6 dBi	76%

Table II. Comparison between simulated and measured results of aperture coupled superstrate antenna.

Also the gain measured and simulated are in good agreement but with a frequency shift as explained. The gain is measured using comparison technique with a standard horn of known gain. For calculating the efficiency, we compared the measured gain with the simulated directivity. The measured and simulated E plane radiation patterns are shown in Fig. 10a for the optimised superstrate dimension. It is clear from Fig. 9 that the measured S11 and gain are shifted; the measured gain is maximum between 57 to 59 GHz and simulated gain is from 59 to 61 GHz. Hence the radiation patterns are plotted by taking in account of this frequency shifting (e.g.; that is radiation pattern plotted is, 60 GHz simulation and 58 GHz measurement, and so on). It is noted that the radiation patterns are found to be broad and in good agreement with measurements, and there is a cross polar level of less than -28 dB at all frequencies. The radiation patterns are verified to be the same in all the frequencies in the band of interest. The measured half-power beam width is found to be 23° at 58 GHz. Also verified by simulation that the back radiation in this case is below -22 dB as compared to the antenna without superstrate (-12 dB).

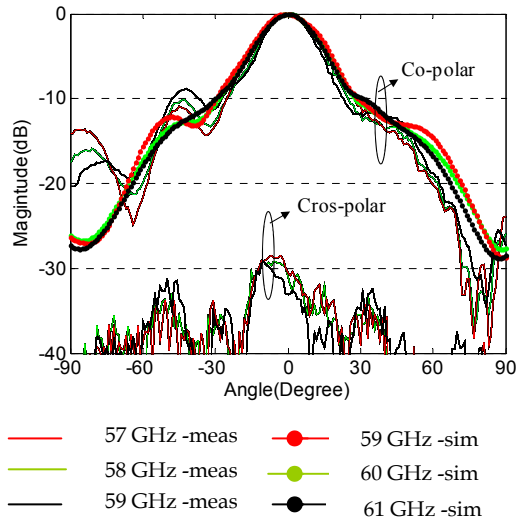


Fig. 10a. Measured and simulated E-plane radiation patterns of superstrate antenna.

The measured and simulated H plane radiation patterns are shown in Fig. 10b for the optimised superstrate dimension. The radiation patterns are also plotted by taking in account of shifting as explained in E plane radiation pattern. It is noted that the radiation patterns are found to be broad and in good agreement with measurements, and there is a cross polar level of less than -28 dB at all frequencies. The measured half-power beam width is found to be 22° at 58 GHz.

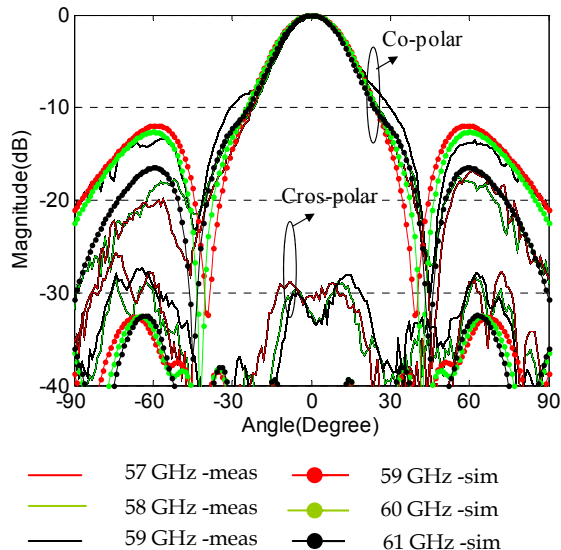


Fig. 10b. Measured and simulated H-plane radiation patterns of superstrate antenna.

When the superstrate size is higher than $2\lambda_0$, the broad nature of the pattern disappeared at 60 GHz. Fig. 11 shows the simulated (60 GHz) and measured (58 GHz) H plane radiation patterns of the antenna with a superstrate dimension of $6\lambda_0$. It is noted that the radiation patterns change from broad side to sectorial / null at 60 GHz, which is also useful for some other applications. It concludes that the dimension of the superstrate is critical for the optimum performance of the antenna. To conclude, the dimension of the superstrate is very important in order to get the consistent radiation pattern for the entire frequency band and it is found to be $2\lambda_0$ in this case. This is the main difference from the already developed superstrate antennas published in the literature.

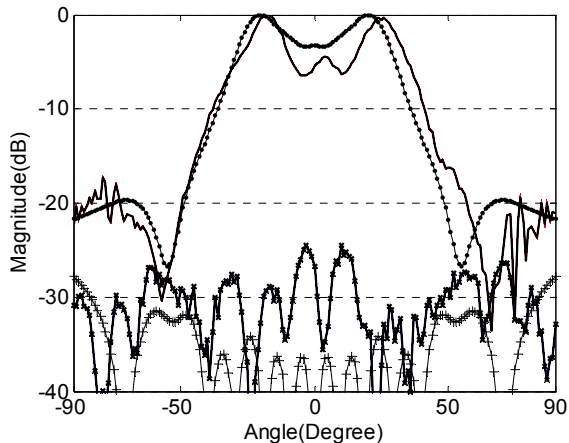


Fig. 11. Measured and simulated H-plane radiation pattern for a superstrate dimension of $6\lambda_0$. Co-simulated \circ — Co-measured \square — Cross-simulated \circ --- Cross-measured \square ---.

2.2.1 Slot coupled 2x2 superstrate antenna array

Fig.12 (a) & (b) show the side view of an aperture coupled 2×2 patch antenna array with superstrate and the feeding network. The distance between the elements in the array are optimized to be $d = 1.3\lambda_0$ for obtaining maximum gain and to minimize coupling. All the base antenna parameters and substrate and superstrate are same as explained in section 2.2.

As explained in section 2.2, usually, large superstrates are used for improving the gain. But our objective is different: we want to use the smallest superstrate for obtaining high stable gain and consistent radiation pattern in the frequency band. To study the effect of superstrate size 'S' and hence to optimize it, here also we considered four square sizes ($2.4\lambda_0$, $3.2\lambda_0$, $4\lambda_0$ and $6\lambda_0$). Simulations are done using CST Microwave studio. Fig. 13 shows the CST results of S11 and directivity variations of the 2×2 slot coupled antenna array with varying superstrate size. The S11 and directivity are affected by the size of the superstrate: the highest directivity of 18 dBi is obtained for a superstrate size of $3.2\lambda_0$. The resulting 2:1

VSWR bandwidth is 5% from 58.6 to 61.6 GHz. It is also noticed that the directivity decreases when the size of the superstrate is above or below $3.2 \lambda_0$ and hence the optimised size of the 2×2 superstrate antenna array is $3.2 \lambda_0 \times 3.2 \lambda_0$. Fig. 14 shows the comparison of measured and simulated S11, and measured gain with simulated directivity for the optimised superstrate size of $3.2 \lambda_0$.

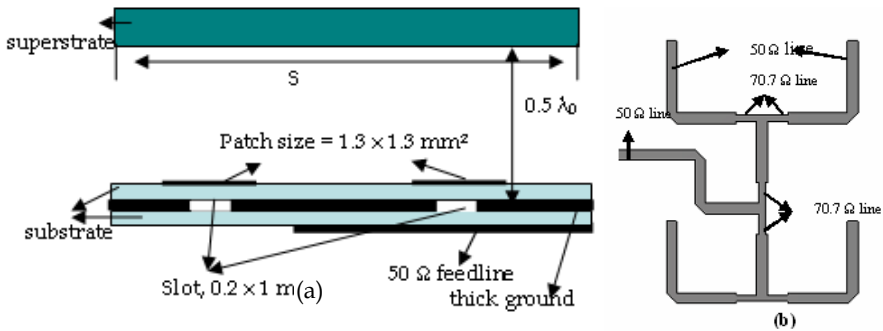


Fig. 12. a) Cutting plane of an aperture coupled 2×2 antenna array with superstrate, ground plane size = $6 \lambda_0 \times 10 \lambda_0$, for connecting V band connector and for ease of measurement purpose, b) 2×2 feeding network.

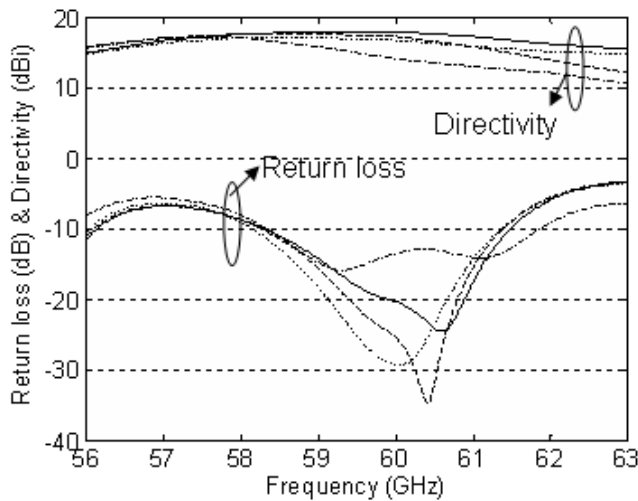


Fig. 13. Variation of return loss and directivity with various superstrate dimensions. $2.4 \lambda_0$ $3.2 \lambda_0$ — $4 \lambda_0$ - - - $6 \lambda_0$ - · - ·

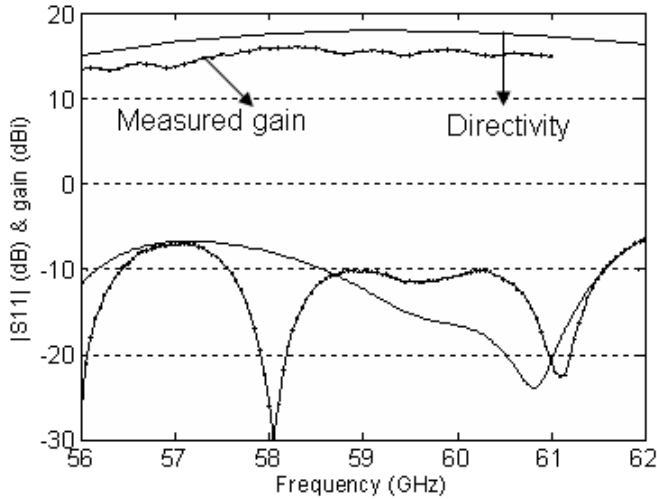


Fig. 14. Variation of S11 and gain with a superstrate dimension of $3.2 \lambda_0$.
Simulation ——— measured - - - - .

Table III gives the comparison of measured and simulated results for the optimised superstrate size (Vettikalladi et al., 2010a). It is found that the measured maximum gain is 16 dBi with S11 bandwidth of 6.7% and an estimated efficiency of 63%. With superstrate there is a gain enhancement of 4 dB compared to the classical 2×2 array (Liu et al., 2009, Book chapter 5, O. Lafond & M. Himdi). The measured gain is maximum at 58 GHz while the simulated directivity is maximum at 59 GHz, which corresponds to 1.7% frequency shift.

Return loss bandwidth (simulated)	Return loss bandwidth (measured)	Maximum Directivity (simulated)	Maximum Gain (measured)	Efficiency Estimated η
58.6 - 61.6 GHz (5%)	57.6 - 61.6 GHz (6.7%)	18 dBi	16 dBi	63%

Table III. Comparison of simulated and measured 2×2 superstrate antenna array.

The simulated and measured E-plane radiation patterns are shown in Fig. 15 for the optimised superstrate dimensions. It is clear from Fig. 14 that the measured gain is maximum between 58 to 59 GHz and simulated is from 59 to 60 GHz. Hence the radiation patterns are plotted by taking in account of this 1.7% shift (e.g. ; that is radiation pattern plotted is, 60 GHz simulation and 59 GHz measurement, etc). It is noted that the radiation patterns are found to be broad and in good agreement with measurements. The measured half-power beam width (HPBW) is found to be 17° at 59 GHz.

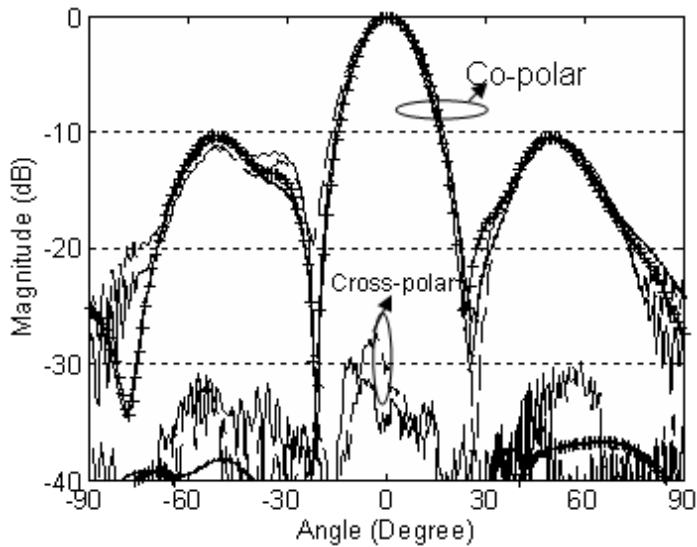


Fig. 15. Measured and simulated E-plane radiation patterns of 2×2 superstrate antenna array.
 58 GHz - measured — — — — 59 GHz - simulated —●—
 59 GHz - measured ———— 60 GHz - simulated —+— .

The measured and simulated H-plane radiation patterns are shown in Fig. 16 for the optimised superstrate dimension. The radiation patterns are also plotted by taking into

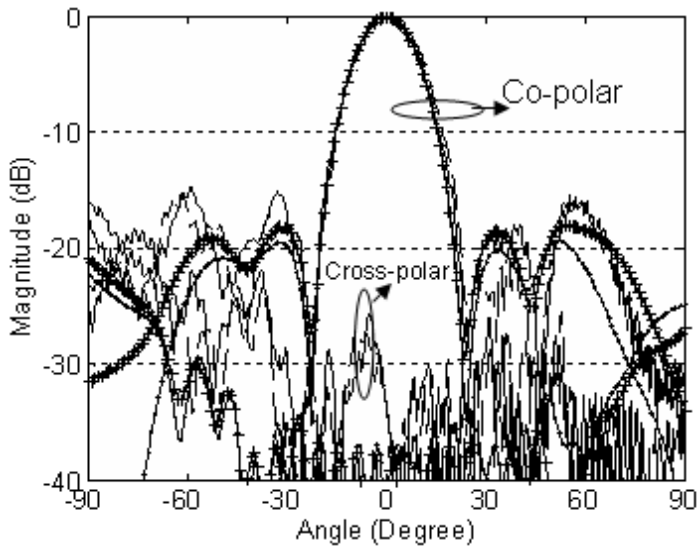


Fig. 16. Measured and simulated H-plane radiation patterns of 2×2 superstrate antenna array.
 58 GHz - measured ———— 59 GHz - simulated —+—
 59 GHz - measured — — — — 60 GHz - simulated —●— .

account the shift as explained for E-plane radiation patterns. It is noted that the radiation patterns are found to be broad and in good agreement with measurements. The measured HPBW is found to be 16° at 59 GHz. The cross polarisation level is lower than -26 dB on both the E and H-plane, and is lower than -19 dB at 45° cut plane in 3D pattern.

2.2.2 Slot coupled 4x4 superstrate antenna array

Fig. 17 shows the photograph of a 4×4 array antenna array with superstrate. The antenna parameters and the distance between the elements are the same as explained for 2×2 superstrate antenna array in Section 2.2. The same substrate for the superstrate is used. Here also the superstrate should be optimised and it is found to be $6 \lambda_0 \times 6 \lambda_0$ which is the total size of the antenna. For manufacturing this prototype, because of the 4×4 array, two metal wedges of 3 mm width are used to position the superstrate at 2.3 mm ($\sim \lambda_0/2 - 0.127$ mm) above the patch array. This mechanical solution is found to be better in this case than foam due to the relation between superstrate position sensitivity and gain increase.

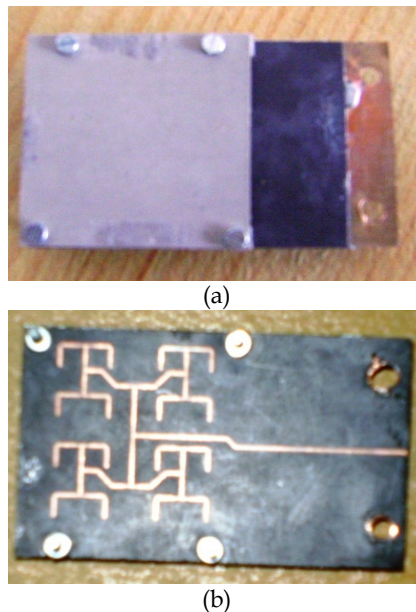


Fig. 17. Photograph of 4×4 superstrate antenna array prototype (a) Top view of superstrate antenna (b) view of antenna feed line network from bottom side. Ground plane taken is $6 \lambda_0 \times 10 \lambda_0$, for connecting V band connector and for ease of measurement purpose.

Fig. 18 shows the measured and simulated S_{11} , and measured gain with simulated directivity. It is found that the maximum gain measured is 19.7 dBi with an efficiency of 51% (simulated directivity = 22.6 dBi) which is far better than a classical 6×6 array antenna of gain 17.5 dBi with an efficiency of 40% at 59 GHz (Lafond et al. 2001), and an 8×8 array antenna (size = $6.5 \lambda_0 \times 6.5 \lambda_0$) of gain 19.7 dBi with an efficiency of $\sim 40\%$ as explained in (Nesic et al., 2001). The S_{11} bandwidth measured for the 4×4 superstrate antenna array is 57.9 GHz to 61.3 GHz (5.7%).

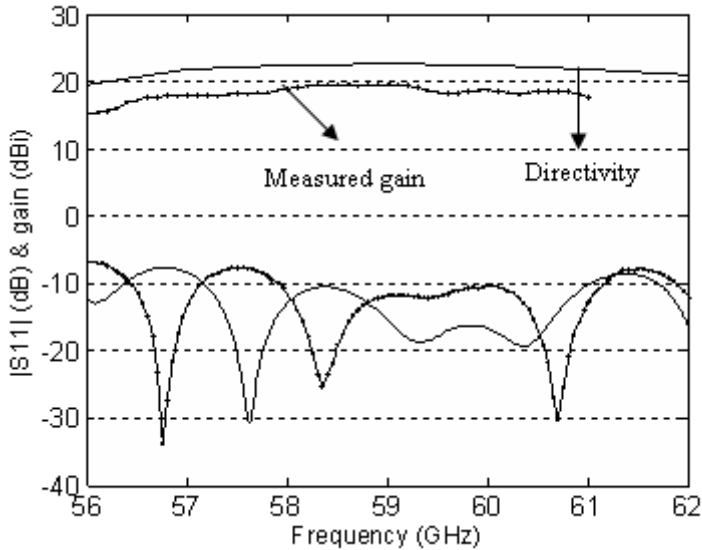


Fig. 18. Variation of S11 and gain with a square superstrate dimension of $6\lambda_0$.
Simulation ——— measured —•— .

It is clear from Fig. 18 that the gain measured and simulated are maximum from 58 GHz to 60 GHz. The simulated and measured H-plane radiation patterns are shown in Fig. 19 (a). It is noted that the simulated patterns are in good agreement with the measured results. The measured HPBW is 8° at 60 GHz.

The measured and simulated E-plane radiation patterns are shown in Fig. 19(b) for the optimised superstrate dimension. The radiation patterns are broad and the agreement between measurement and simulation are quite acceptable. The measured HPBW is found to be 10° at 60 GHz. In this case, the cross polarization level is lower than -25 dB on both the E and H-plane, and is lower than -16 dB at 45° cut plane in 3D pattern.

For both the presented antennas, a distance between the elements of $1.3\lambda_0$ is used for the source array, which induces high ambiguity side lobes for both cases when there is no superstrate: -2 dB for a 2×2 array and -1.9 dB for 4×4 array as shown in Fig. 20. Adding a superstrate will strengthen the main lobe while suppressing the ambiguity side lobes to less than -10 dB for both the arrays without affecting the back radiation as shown in Fig. 20. It also strengthens the front to back ratio as shown in the figure. It is to be underlined that the size of the superstrate is a key point of the design of such structures. In fact the nature of the pattern is conditioned by the choice of this parameter: a broad pattern is obtained for a size limited to $3.2\lambda_0 \times 3.2\lambda_0$ for 2×2 array and $6\lambda_0 \times 6\lambda_0$ for 4×4 arrays.

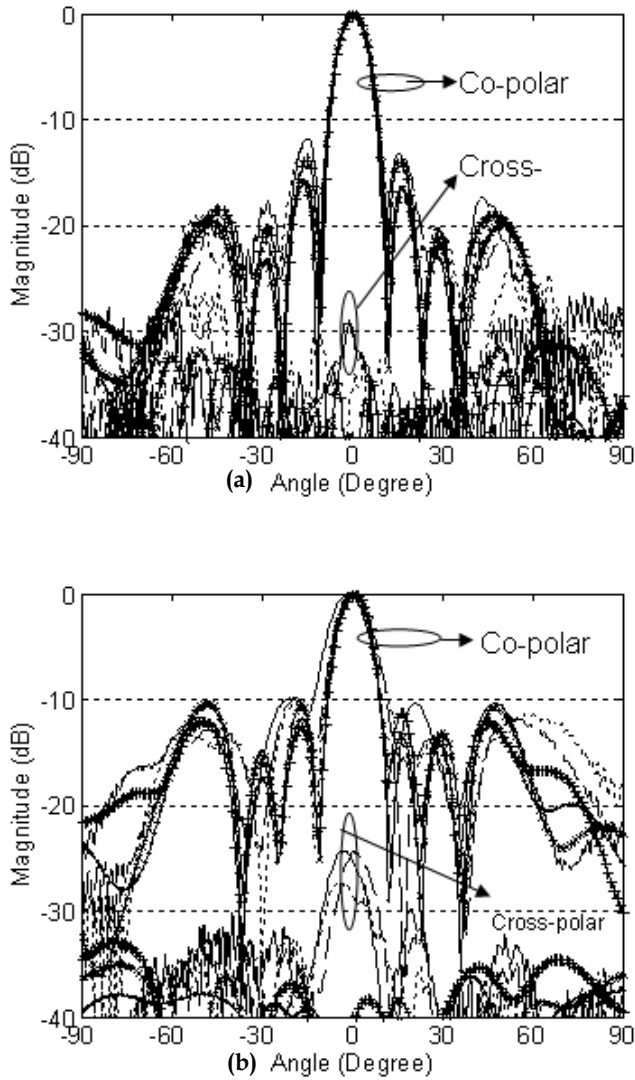


Fig. 19. Measured and simulated (a) H plane & (b) E plane radiation patterns of 4 x 4 superstrate antenna array.

58 GHz - measured 58 GHz - simulated —X—
 59 GHz - measured - - - - 59 GHz - simulated —•—
 60 GHz - measured ——— 60 GHz - simulated —+— .

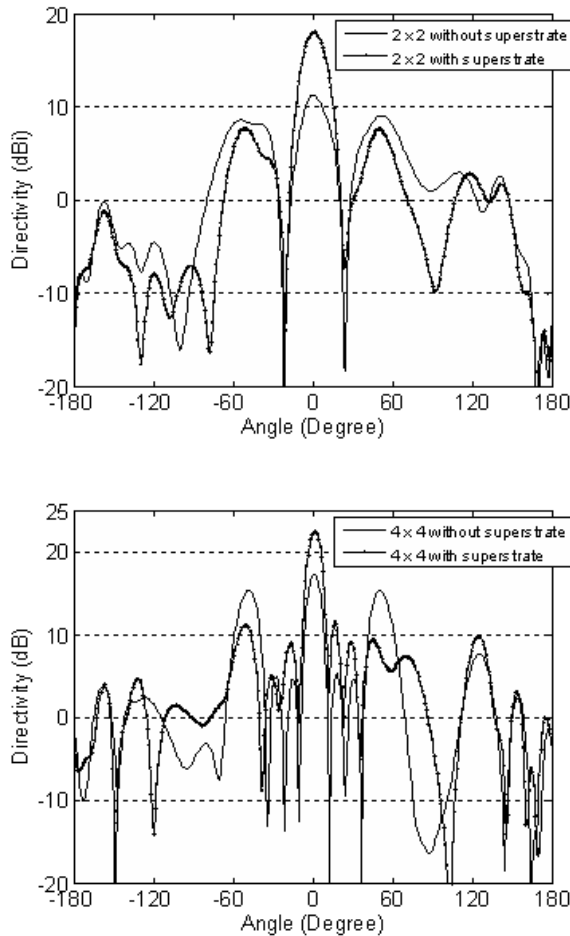


Fig. 20. Comparison of simulated results of 2×2 and 4×4 arrays without and with superstrate in terms of main beam, side lobe and back radiation.

2.3 Superstrate aperture antenna

In this section we are using another source, as aperture, for exciting the antenna. The side view of an aperture antenna with superstrate is shown in Fig. 21(a). The aperture is optimised to $4.4 \text{ mm} \times 1 \text{ mm}$ for maximum coupling with a stub length of 0.4 mm (Fig. 21(b)). To improve the rigidity of antenna, a ground plane of thickness $t = 0.2 \text{ mm}$ is used. For maintaining the exact air thickness in practical prototype, the superstrate is inserted within an air pocket realized in a Rohacell foam block of permittivity 1.05, as shown in Fig. 21(c). All the substrate and superstrate material used are the same as explained in section 2.2.

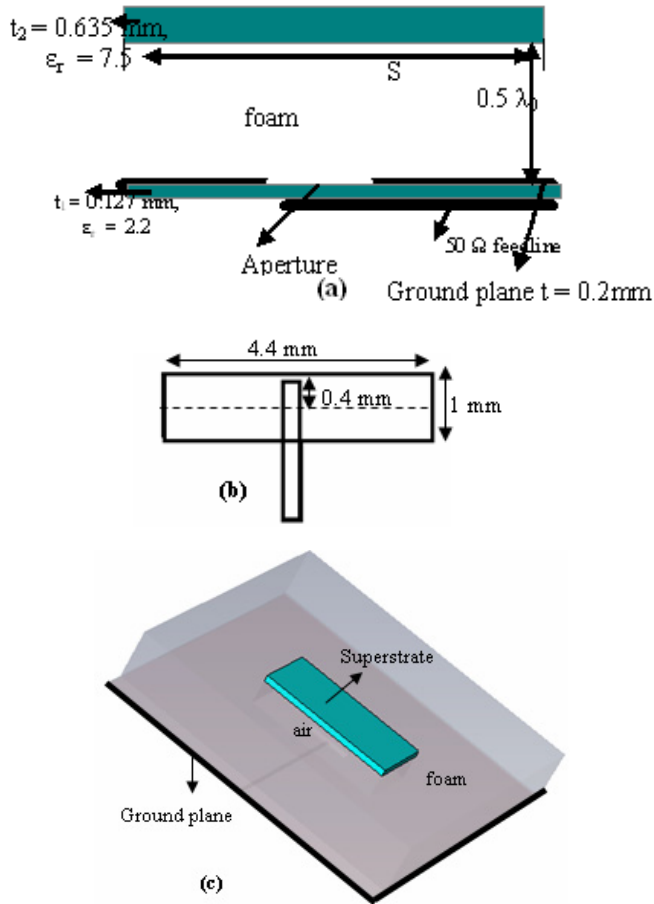


Fig. 21. (a) Cutting plane of aperture antenna with superstrate, ground plane size = $6 \lambda_0 \times 6 \lambda_0$. (b) Aperture and stub in details. (c) Overview of the Prototype: Details of the superstrate and air gap within foam.

In this case also we want to study the effect of superstrate size on antenna performance. To study the effect of superstrate size 'S' and hence to optimize it, a parametric study has been performed, using commercial electromagnetic software CST Microwave studio. To highlight the effects of this parameter, results obtained for four sizes ($1 \lambda_0 \times 2 \lambda_0$, $2 \lambda_0 \times 2 \lambda_0$, $1.2 \lambda_0 \times 2.7 \lambda_0$ and $3 \lambda_0 \times 3 \lambda_0$) are reported in Fig. 22. It is observed that both the S11 and the directivity vary according to the size of the superstrate. A maximum directivity of 14.5 dBi is obtained for a superstrate size of $1.2 \lambda_0 \times 2.7 \lambda_0$. The corresponding 2:1 VSWR bandwidth is noted to be equal to 57.5 - 71 GHz i.e. 22.5%. It is also noticed that the directivity decreases when the size of the superstrate is above or below this optimized value. We plotted directivity only up to 65 GHz because of the decline in the values after that. When the superstrate size is higher than the optimized value, then there is a plunge in directivity as shown in Fig. 22. Hence the broad nature of the pattern moved out at 60 GHz as explained

in (Vettikalladi et al., 2009a), i.e the radiation patterns change from broad side to sectorial / null at 60 GHz, which is also useful for some other applications.

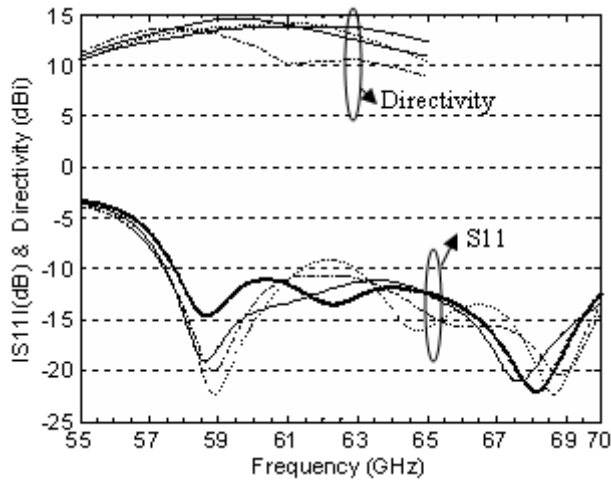


Fig. 22. Simulated results of S11 and directivity with various superstrate dimensions. $1 \lambda_0 \times 2 \lambda_0$ —●— ; $2 \lambda_0 \times 2 \lambda_0$; $1.2 \lambda_0 \times 2.7 \lambda_0$ — ; $3 \lambda_0 \times 3 \lambda_0$ - · - · - .

It concludes that in this case the dimension of the superstrate is critical for the optimum performance of the antenna. Also we can control the shape of the pattern by changing the dimension of the superstrate from broadside to sectorial / null. The comparison of measured and simulated S11, and measured gain with simulated directivity for the optimized superstrate size is shown in Fig. 23 (a). Table IV gives the summary of these results (Vettikalladi et al., 2010b). It is noted that the measured 2:1 VSWR bandwidth is 15% which is larger compared to the superstrate slot coupled antenna (Vettikalladi et al., 2009a), where the bandwidth was only 6.8%. The gain is measured using comparison technique with a standard gain horn. It is found to be 13.1 dBi. Moreover, it is almost flat (ripple ~ 0.5 dB) over a bandwidth of 5 GHz. To determine the efficiency, we compared the measured gain with the simulated directivity. The estimated efficiency is 79%. In order to highlight the effect of the superstrate for this configuration, the simulated comparison of E-plane radiation pattern of aperture antenna with superstrate and without superstrate is shown in Fig. 23 (b). The ripples in the pattern without superstrate are due the diffraction from the edges of the limited ground plane. Also it is clear that aperture antenna is a bidirectional antenna, superstrate technology make this antenna to unidirectional without adding any reflector, which is a highlight of the superstrate with this kind of source. I.e. Superstrate makes the antenna pattern directive and there is a gain enhancement of 8 dB compared to its basic aperture antenna.

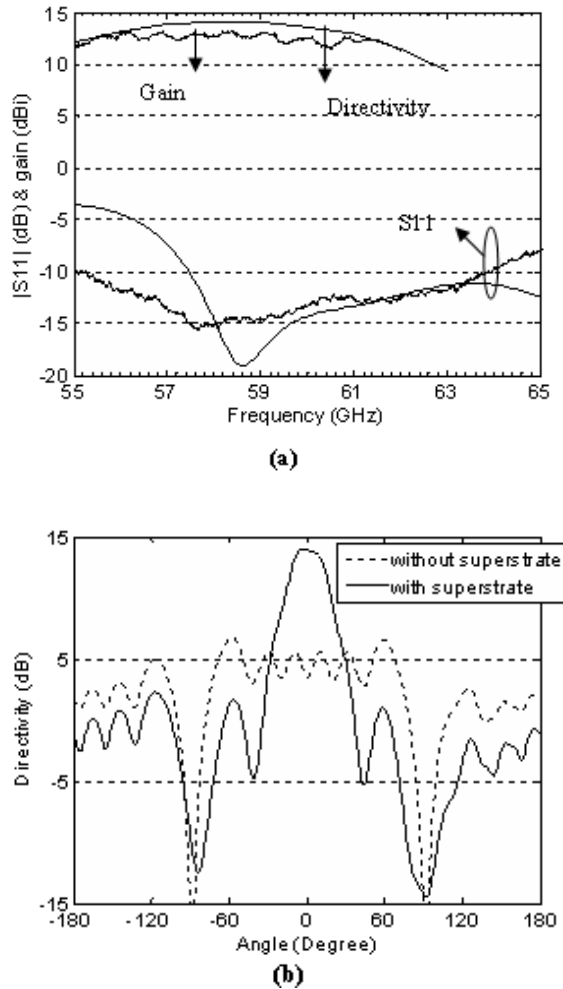


Fig. 23. (a) Results of S11 and gain with a superstrate dimension of $1.2 \lambda_0 \times 2.7 \lambda_0$. Simulation —•—; measurement —. (b) Simulated comparison of E-plane antenna radiation pattern with and without superstrate.

The measured and simulated H- and E-plane radiation patterns at 57 GHz, 60 GHz and 62 GHz are shown in Fig. 24 for the optimized superstrate dimension. It is noted that the radiation patterns are found to be broad and in agreement with simulations. The cross polar level is less than -25 dB for H-plane and -20 dB for E-plane respectively, for all the frequencies in the band. The measured half-power beam widths (HPBW) at 60 GHz are 26° for H-plane and 30° for E-plane respectively. The measured cross polarization level is lower than -17 dB at 45° cut plane in 3D patterns of both planes.

Return loss bandwidth (simulated)	Return loss bandwidth (measured)	Maximum Directivity of the prototype (simulated)	Maximum Gain (measured)	Efficiency Estimated η
57.5-71 GHz (22.5%)	55 - 64GHz (15%)	14.1 dBi	13.1 dBi	79%

Table IV. Comparison between simulated and measured results superstrate aperture antenna.

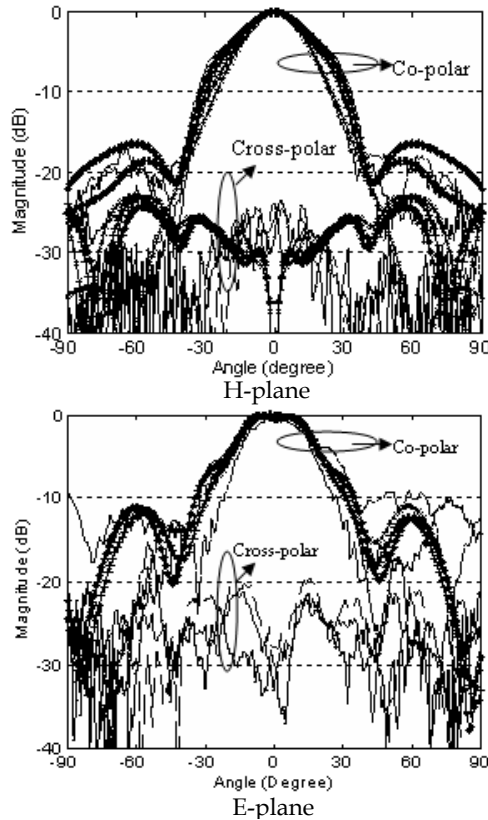


Fig. 24. Measured and simulated H-plane & E- plane radiation patterns of superstrate antenna (Co and Cross polarisation).

57 GHz - Simulated —×— ; 57 GHz - Measured —●— ;
 60 GHz - Simulated —+— ; 60 GHz - Measured —■— ;
 62 GHz - Simulated —*— ; 62 GHz - Measured - - - - .

2.3.1 2x2 superstrate aperture antenna array

The side and 3D view of a 2×2 aperture antenna array with superstrate are shown in Figs. 25(a) and (c). All the parameters of the antenna are the same as explained in section 2.3. For maintaining the exact air thickness, the superstrate is inserted within an air pocket realized in Rohacell foam as shown in Fig. 25(c). The distance between the elements in the array is optimized as $d = 1.3 \lambda_0$ for obtaining maximum gain and to minimize coupling. The 2×2 feeding network is exposed in Fig. 25(b). In a classical array (without superstrate), when the

distance between the patches is $d = 1.3 \lambda_0$, high ambiguity side lobes appear with almost the same level as the main lobe. Adding a superstrate strengthens the main lobe while reducing the ambiguity side lobes to less than -10 dB (E-plane) as shown in Fig. 26. It also strengthens the front to back ratio as shown in the figure. It has to be underlined that the size of the superstrate is a key point for the design of such structures as explained in previous cases.

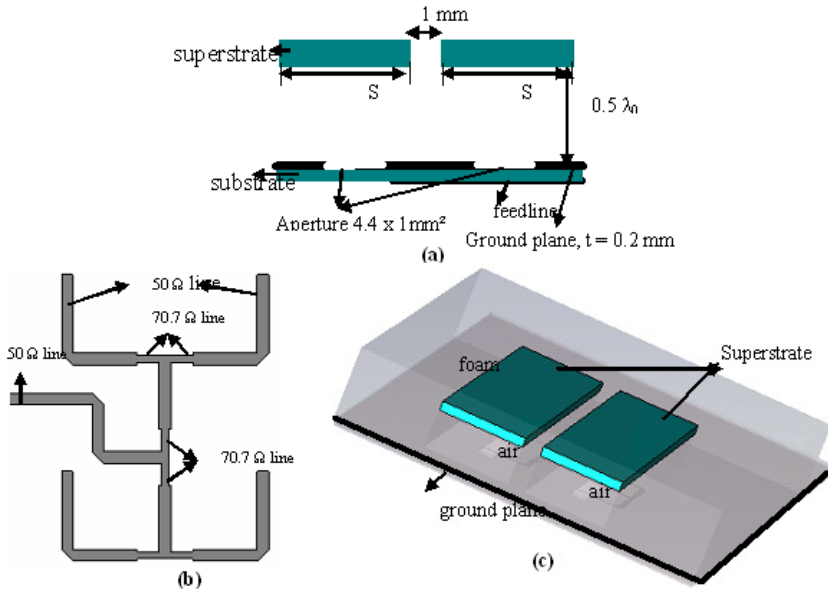


Fig. 25. (a) Cutting plane of a 2 x 2 aperture antenna array with superstrate, ground plane size = $6 \lambda_0 \times 10 \lambda_0$ for connecting V band connector and for ease of measurement purpose. (b) 2 x 2 feed network. (c) Overview of 2 x 2 array prototype: details of the two separate superstrate sheets and air gap within foam.

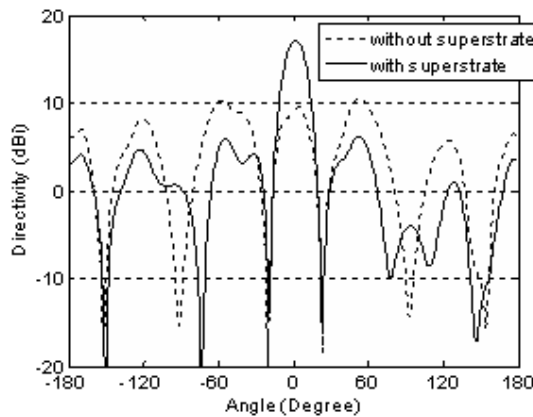


Fig. 26. Simulated comparison of 2x2 antenna array pattern ($d = 1.3 \lambda_0$) with and without superstrate (E-plane).

As did in previous sections, we studied the effect of superstrate size 'S' by simulating different sizes and the optimized solution is found to be two pieces of dimension $1.2 \lambda_0 \times 4 \lambda_0$, one sheet for two aperture antenna, with a spacing of 1mm as shown in Fig. 25(b). If we use a single piece with a size of $2.6 \lambda_0 \times 4 \lambda_0$, then the gain is little lower than in the previous case. The comparison of measured and simulated S11, and measured gain with simulated directivity for the optimized superstrate size is shown in Fig. 27. The highest directivity of 17.9 dBi is obtained for the optimized superstrate size. The resulting simulated 2:1 VSWR bandwidth is 11.3% from 57.2 to 64 GHz.

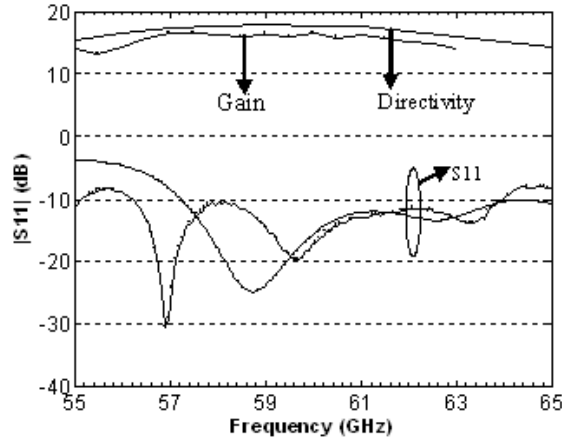


Fig. 27. Variation of S11 and gain for a 2×2 superstrate aperture antenna array. Simulation —; measurement - - -.

Table V gives the comparison of measured and simulated results for the optimized superstrate size. It is found that the maximum measured gain is 16.6 dBi with S11 bandwidth of 13.3% (56 GHz - 64 GHz), and an estimated efficiency of 74%. This gain is comparable to a classical 4×4 array at 60 GHz but with better efficiency (Lafond 2000). Also the measured gain is almost stable (ripple ~ 0.8 dB) over 5 GHz (57 GHz - 62 GHz) in the band of interest (Vettikalladi et al., 2010c).

Return loss bandwidth (simulated)	Return loss bandwidth (measured)	Maximum Directivity (simulated)	Maximum Gain (measured)	Efficiency Estimated η
57.2 - 64 GHz (11.3%)	56 - 64 GHz (13.3%)	17.9 dBi	16.6 dBi	74%

Table V. Comparison between simulated and measured results of a 2×2 superstrate aperture antenna array.

The measured and simulated H- and E-plane radiation patterns at 57 GHz, 60 GHz and 62 GHz are shown in Figs. 28 (a) & (b) respectively for the optimized superstrate dimension. It is noted that the radiation patterns are found to be broad and in good agreement with simulations. The measured cross polar levels are -26 dB for H-plane and -20 dB for E-plane respectively. The radiation patterns are verified to be the same in all the frequencies in the band of interest. The measured HPBW's are 17° for H-plane and 16° for E-plane respectively at 60 GHz.

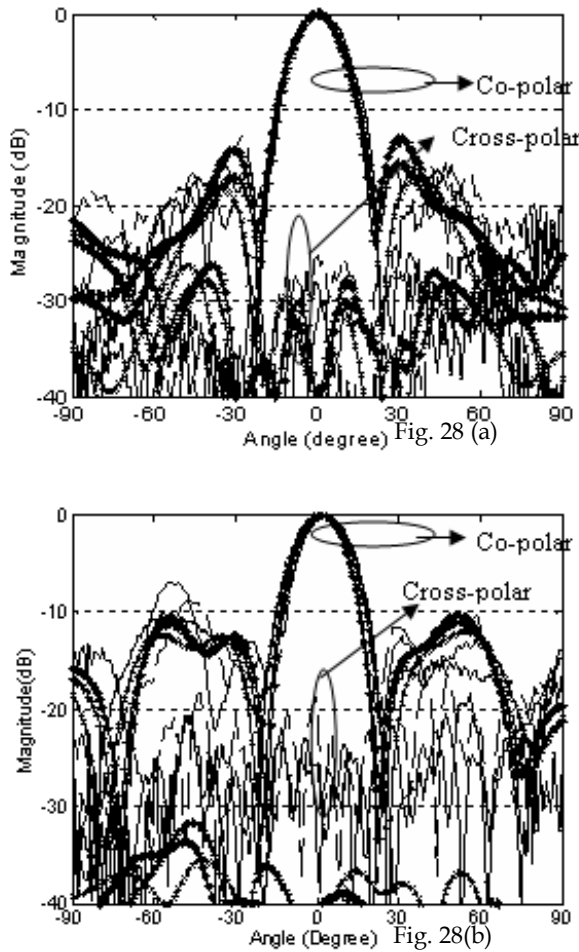


Fig. 28. Measured and simulated H-plane (a) & E-plane (b) radiation patterns of the 2 x 2 superstrate antenna array (Co and Cross polarisation).

57 GHz - Simulated —×— ; 57 GHz - Measured —•— ;
 60 GHz - Simulated —+— ; 60 GHz - Measured ——— ;
 62 GHz - Simulated —*— ; 62 GHz - Measured - - - - .

2.3.2 16 x 16 superstrate aperture antenna array

Finally we developed a big array to obtain very high gain of nearly 30 dBi for 60 GHz outdoor communication, for example from one department to another department inside a university (< 1km). Fig. 29 (a) depicts the 3D side view of the 16 x 16 array prototype. The antenna parameters and the distance between the elements are all same as explained in Section 3.2. For maintaining the exact air thickness, the superstrate is inserted within an air

pocket realized in Rohacell foam as shown in Fig. 29(a). The 16 x 16 feeding network is showing in Fig. 29(b).

As pointed out in previous section , we want to use the smallest superstrate for obtaining high stable gain and consistent radiation pattern in the frequency band. We studied the effect of superstrate size ' S ' by simulating different sizes and the optimized size is found to be 16 pieces of dimension $1.2 \lambda_0 \times 21.8 \lambda_0$, one sheet for 16 aperture antenna, with a spacing of 1mm as shown in Fig. 29(a). If we use a single piece with a size of $20.6 \lambda_0 \times 21.8 \lambda_0$, then the gain is little lower than in the previous case.

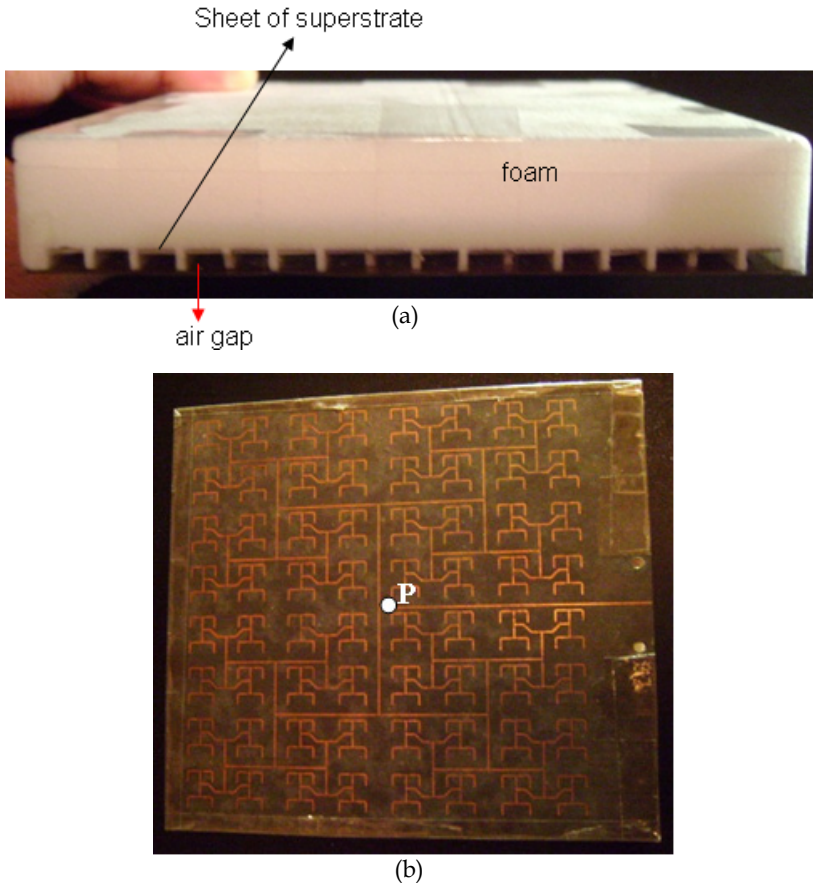


Fig. 29. (a) Side overview of 16 x 16 array prototype: details of the 16 separate superstrate sheets and air gap within foam, total size = $20.6 \lambda_0 \times 21.8 \lambda_0$. (b) 16 x 16 feed network.

The comparison of measured and simulated S11, and measured gain with simulated directivity for the optimised superstrate size is shown in Fig. 30. The highest simulated directivity of 33.3 dBi is obtained for the optimised superstrate size. The resulting simulated 2:1 VSWR bandwidth is 22 %. It is found that the maximum measured gain is 29.4 dBi (at point 'P' in Fig. 29 (b)) with S11 bandwidth of 16.7 % (54 GHz - 64 GHz), and an estimated

efficiency of 41%. The measured and simulated E- and H-plane radiation patterns at 57 GHz, 60 GHz and 62 GHz are shown in Figs. 31(a) and (b) respectively for the optimised superstrate dimension. It is noted that the radiation patterns are found to be broad and in good agreement with simulations. The measured cross polar levels are -28 dB for H-plane and -26 dB for E-plane respectively. The radiation patterns are verified to be the same in all the frequencies in the band of interest. The measured HPBW's are 2.5° for H-plane and E-plane respectively at 60 GHz.

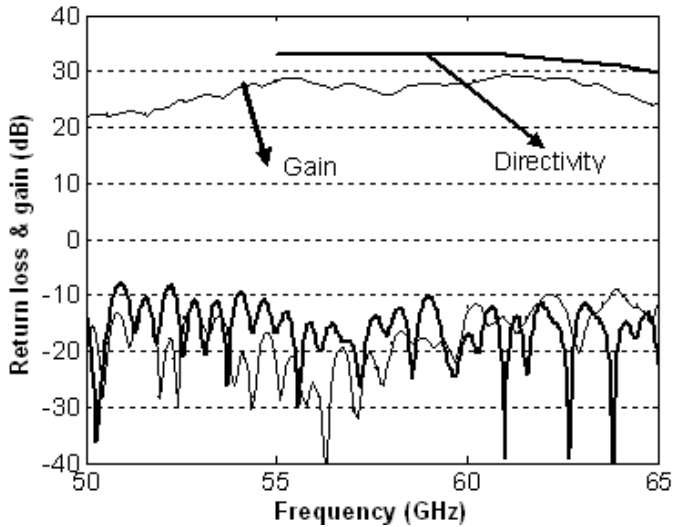


Fig. 30. Variation of S11 and gain for a 16 x 16 superstrate aperture antenna array. Simulation — ; measurement - - - .

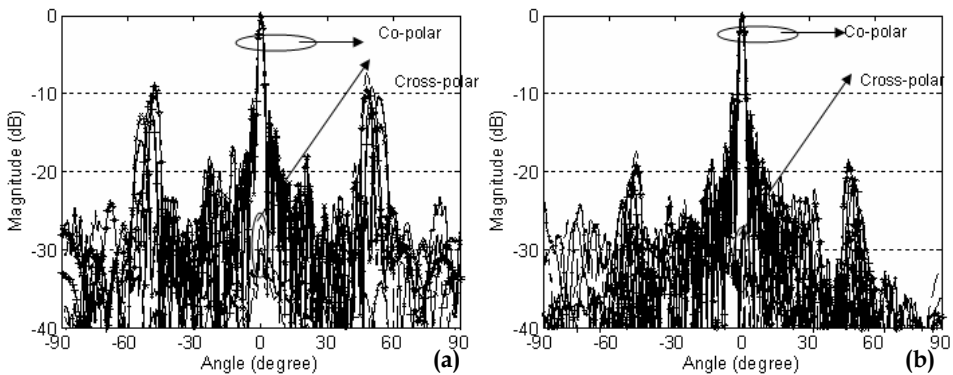


Fig. 31. Measured and simulated E-plane (a) & H-plane (b) radiation patterns of the 2 x 2 superstrate antenna array (Co and Cross polarisation).

57 GHz - Simulated —×— ; 57 GHz - Measured —●— ;
 60 GHz - Simulated —+— ; 60 GHz - Measured —■— ;
 62 GHz - Simulated —*— ; 62 GHz - Measured - - - .

It is noted from the study that single/small array superstrate antenna technology is very good for high gain, wide bandwidth and high efficiency, but is not suggestive for big arrays because of the gain go up is not upto the point but is good in terms of efficiency.

3. Comparison of superstrate slot coupled antenna with superstrate aperture antenna

Table VI gives the comparison of the superstrate aperture antenna element or antenna array presented with slot coupled superstrate antenna element or antenna array. It is clear that superstrate aperture antenna element / antenna array give broad S11 bandwidth of 15% / 13.3% with better efficiency as compared to superstrate slot coupled antenna element / antenna array. Also the antenna size is smaller compared to the other in both the cases. I.e. superstrate aperture antenna element or antenna array gives sufficient bandwidth, gain and efficiency for 60 GHz applications.

Antenna	Return loss bandwidth (measured)	Maximum gain (measured)	Efficiency Estimated η	Size
Single superstrate Aperture	15%	13.1	79%	6 mm x 13.5 mm x 3.48 mm
Single superstrate slot coupled	6.8%	14.6	76%	10 mm x 10mm x 3.48 mm
2x2 superstrate Aperture	13.3%	16.6	74%	13 mm x 20 mm, x 3.48 mm
2x2 superstrate slot coupled	6.7%	16	63%	16 mm x 16 mm x 3.48 mm

Table VI. Comparison between superstrate aperture s antenna and slot coupled superstrate antenna explained in section 2.

4. Conclusion

In this chapter we explained about the significance of superstrate on antenna performance at millimeter wave frequencies. It is found that the size of the superstrate is critical, which is not the case in lower frequencies. Also we studied the antenna performance with different source of excitation. It is noted that superstrate technology is very good for a single patch and also for small array but not that much good for big arrays. As a conclusion, it is found that superstrate aperture antenna element / antenna array is a good candidate for wideband, high gain and high efficiency antenna design in millimetre wave range. Moreover it is easy to integrate with electronics by placing the feed on backside of the substrate where the electronic components are integrated, and the radiating aperture and superstrate (i.e. the radiation part) are on the other side.

5. References

- Cho, W; Yong Hei Cho; Cheol-Sik Pyo & Jae_Ick Choi. (2003). A high gain microstrip patch array antenna using a superstrate layer, *ETRI journal*, 2003, vol. 25, pp.407-411.
- Gupta, R. K. & Kumar, G. (2005). High gain multilayered antenna for wireless applications, *Microw. Opt. Technol. Lett.*, vol. 50, no. 7, pp. 152-154, Jul. 2005.
- Julio-Navarro. (2002). Wide-band, low-profile millimeter wave antenna array, *Microw. Opt. Technol. Lett.*, 2002, vol. 34 , pp. 253-255.
- Kärnfelt, C; Hallbjörner, P; Zirath, H. & Alping, A. (2006). High gain active microstrip antenna for 60-GHz WLAN/WPAN applications, *IEEE Trans. on Microw. Theor. and Techniq.*, Jun. 2006, 54 (6), pp. 2593-2602.
- Liu, D; Gaucher, B; Ullrich, P. & Janusz, G. (2009). *Advanced Millimeter-wave technologies*, (Wiley, 2009), pp. 170-172.
- Lafond, O. (2000). Conception et Technologies D'antennes imprimees Multicouches a 60 GHz, *PhD thesis, University of Rennes1, France*, Dec. 2000, pp. 52-54
- Lafond, O; Himdi, M. & Daniel J. P. (2001). Thick slot-coupled printed antenna arrays for a 60 GHz indoor communication system, *Microw. Opt. Technol. Lett.*, 2001, vol. 28, pp. 105-108.
- Meriah, S. M; Cambiaggio, E; Staraj, R. & Bendimerad, F. T. (2008). Gain enhancement for microstrip reflect array using superstrate layer, *Microw. Opt. Technol. Lett.* , 2008, vol. 46, pp. 1923-1929.
- Menudier, C; Thevenot, M, Monediere, T & Jecko, B . (2007). Ebg resonator antennas state of the art and prospects, *International Conference on Antenna Theory and Techniques*, 17-21 September, 2007, Sevastopol, Ukraine.
- Nesic, A; Nestic, D; Brankovic, V; Sasaki, K. & Kawasaki, K. (2001). Antenna solution for future communication Devices in mm-wave range, *Microwave Review*, pp. 9-17, Dec. 2001.
- Soon-soo oh; John Heo; Dong-Hyeon Kim; Jae-Wook Lee; Myung-sun song & Yung-sik kim. (2004). Broadband millimeter-wave planar antenna array with a waveguide and microstrip feed network, *Microw. Opt. Technol. Lett.*, 2004, vol. 42, pp. 283-287.
- Vettikalladi, H; Lafond, O. & Himdi, M. (2009a). High-Efficient and High-Gain Superstrate Antenna for 60 GHz Indoor Communication, *IEEE Antennas and Wireless Propagation Letters* vol. 8, pp. 1422-1425, 2009.
- Vettikalladi, H; Lafond, O. & Himdi, M. (2009b). High-Gain Broad-band Superstrate Millimeter wave Antenna for 60 GHz Indoor Communications, *5th ESA Workshop on Millimetre Wave Technology and Applications and 31st ESA Antenna Workshop*, 18 - 20 May 2009, ESTEC, Noordwijk, The Netherlands.
- Vettikalladi, H; Le Coq, L; Lafond, O. & Himdi, M. (2010a). Efficient and High-Gain Aperture Coupled Superstrate Antenna Arrays for 60 GHz Indoor Communication Systems, *Microwave and Optical Technology Letters*, 2010.
- Vettikalladi, H; Le Coq, L; Lafond, O. & Himdi, M. (2010b). Wideband and High Efficient Aperture Antenna with Superstrate for 60 GHz Indoor Communication Systems, *2010 IEEE AP-S International Symposium on Antennas and Propagation and 2010 USNC/CNC/URSI Meeting in Toronto, ON, Canada*, on July 11- 17, 2010.

- Vettikalladi, H; Le Coq, L; Lafond, O. & Himdi, M. (2010c). Broadband Superstrate Aperture Antenna for 60 GHz Applications, *European Microwave Week*, 26th sept. to-1st October 2010, Paris, France.
- Zhang, Y.P & Wang, J.J. (2006). Theory and analysis of differentially-driven microstrip antennas, *IEEE Trans. Antennas Propag.*, 2006, vol. 54, pp.1092-1099.

Part 2

Wireless Communication Hardware

Hardware Implementation of Wireless Communications Algorithms: A Practical Approach

Antonio F. Mondragon-Torres
Rochester Institute of Technology
USA

1. Introduction

Wireless communication algorithms are implemented using a wide spectrum of building blocks such as: source coding; channel coding; modulation; multiplexing in time, frequency and code domains; channel estimation; time and frequency domain synchronization and equalization; pre-distortion; transmit and receive diversity; combat and take advantage of fading and multi-path channels; intermediate frequency (IF) processing in software defined radio, etc.

Due to this breadth of different algorithms, the traditional approach has been to create a system model in a high level language such as Matlab (Mathworks, 2011), C/C++ and recently in SystemC (SystemC, 2011). Usually these models use floating point representations, are architecture agnostic, and are time independent, among others characteristics. After the system model is available, then based on the specifications it is manually converted into a fixed point model that will take care of the finite precision required to implement the algorithm and compare its performance against the “Golden” floating point model. The reason to perform this conversion is due to cost and performance. While it is possible to program the algorithm on a floating point Digital Signal Processor (DSP) or using floating point hardware on application specific integrated circuit (ASIC) technology, the resulting: complexity; signal throughput; silicon area and cost; and power consumption among others, usually prohibits its implementation in floating point arithmetic. This is one of the reasons most of the wireless communications algorithms are implemented using a finite precision fixed point number representation.

In the last decade several technologies have made the conversion from floating point to fixed point seamless to a certain point. These technologies rely either on either a high level language such as C or C++ or a set of hardware model libraries for a particular field programmable gate array (FPGA) or ASIC technologies. In addition to these, there are some other electronic system level (ESL) design tools that can take a floating point algorithm and even preserve the same floating point testbench and transform the algorithm into a fixed point representation, where different architectural trade-offs can be made based on the area/power/latency/throughput requirements are in the system specifications.

In this chapter we do not propose a one solution fits all applications methodology, rather we will navigate through the author's encounters with different technologies at different stages in his career and how different applications have been and are currently approached. This is a summary of the last ten years of working with different tools, methodologies and design flows. What has prevailed due the level of integration of current Systems on a Chip (SoC) has been for example: component and systems reusability; fast algorithm and architecture exploration; algorithm hardware emulation; and design levels of abstraction.

2. System level design

By system level design (SLD), we refer to the modeling of the wireless communications systems based solely on the specifications or target standard. At this stage, individual and collective block level performance can be evaluated and also interconnects with other components in the system can be specified. There are two major known approaches for system design, top-down and bottoms-up methodologies.

System level design calls for a top-down methodology. In sophisticated systems such as SoCs, their complexity can be very large and it is a common practice in system level design to create a set of high level specifications with a complete vision of the system including their complete set of interconnects. The next phase is to divide the system into functional blocks, specify all internal interconnects and design each block in the subsystem. This allows the complete system to be simulated using for example a system level language such as SystemC and then be able to replace each block with its Register Transfer Level (RTL) functional equivalent. These techniques are also being heavily used to speed up system verification in which it is not possible to perform in a reasonable amount of time a complete RTL or gate level simulation due to time to market (TTM) constraints or because it is not computationally feasible. SLD methodologies allow performing a complete system level simulation at a higher level of abstraction by just including the key blocks required at the gate level to test interconnectivity and performance.

A system level simulation is in the order of tenths to thousands times faster than gate level simulations, thus assuring that all individual blocks or combinations of blocks will work after being interconnected. In Figure 1 it is shown an ideal case where a system level model or commonly referred as the "running specification" is first generated and creates a "golden" model against all performance implementations will be compared against. Ideally we would like to keep the original testbench for all modeling, design, implementation, simulation and verifications tasks, but this is not always possible. The problem arises when manual or automatic translations could change the behavior of the original testbench. One of the most critical problems in SLD development is that once you descend in the level of abstraction, the system level testbench and models are no longer updated and maintained, then deviating from the original running specification reference.

2.1 System modelling

SLD has been traditionally been done using C language, therefore it is common to refer in industry to the "C-model" as the running specification or "golden" model. The advantage is that C language is particularly fast, runs on all platforms and can represent fixed point precision easily after taking care of the fixed point operations such as rounding, truncation,

saturation, etc. One disadvantage of this methodology is that it is not very straight forward to couple C simulations with RTL simulations and then obtain the complete benefits of system level modeling.

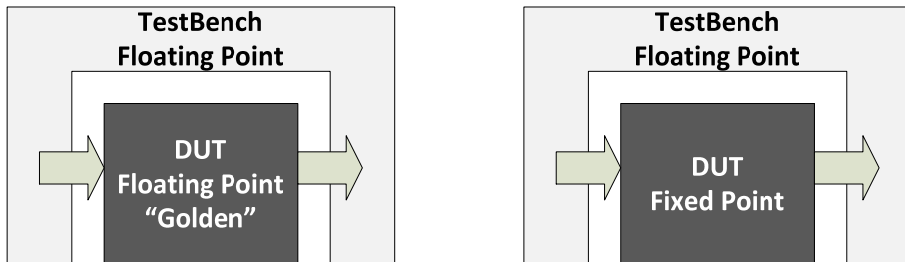


Fig. 1. System Level Modeling approach. Testbench should ideally be reused while verifying the Device Under Test (DUT) at different level of hierarchy. E.g. Behavioral, RTL, gate level netlist and parasitic extracted netlist.

More recently, SLD has been done using C++, since the level of abstraction can be taken one level further and the interfaces and testbenches can be encapsulated and reused. SystemC is a set of libraries that extend C++ to brings capabilities such as fixed point types, transaction level modeling (TLM), parallel event driven simulation compatibility, and testbench reutilization among several other features. Recently SystemC have been used to create complex reusable testbenches that interface directly with RTL code and can be executed using most of the high performance RTL event driven simulators.

A relatively new player in the SLD is System Verilog which in addition to have unique properties to perform verification and design tasks, it can also be used for system level design due to its enhancements comparable with SystemC features. The current belief is that System Verilog can be the "one size fits all" language due to its system and blocks level modeling, system and block level verification, synthesis constructs, and simulation capabilities. One company working in this space is Bluespec that provides high level system modeling, architecture exploration, verification and synthesis using a System Verilog (Bluespec, 2011).

So far, we have talked about languages that are capable of performing SLD, but the drawback of these languages is that they rely on the user knowing the architectural constraints of the design. There is also another very popular complete set of SLD languages that also allow to perform system level modeling at the same time that its users are closer to the algorithm development rather than the language options we just mentioned above. The primary SLD system language for modeling is Mathwork's Matlab and it's time-driven block-based tool Simulink. There are also other tools that also used for system level modeling such as Agilent's SystemVue (Agilent, 2011) and Synopsys' SPW (Synopsys, 2011a) to cite a few used previously by the author.

The author has been exposed to more SLD projects done in Matlab, and in some cases the complete running specification has been kept in Matlab m-code, even the fixed point implementation and test vector generation. Other projects, had Matlab as the main algorithm verification driver, followed by a C model implementation and then by an RTL

implementation. Each tool/language translation can potentially introduce errors in the system level design and verification stages. In an ideal world, we should only deal with one system level language, one system level testbench and multiple implementations at different levels of abstraction. By having different models at different levels of abstraction we can have a different model to resolve efficiently different problems such as interconnection, timing, programming, functional verification, synthesizability, and feasibility of implementation.

2.2 Algorithmic focused system level design

The focus on this chapter will center around Matlab and especially on Simulink. The two major FPGA providers Xilinx (Xilinx, 2011b) and Altera (Altera, 2011a), make available libraries that allow efficient block level modeling of wireless communications algorithms and its automatic conversion to RTL. The code can be either downloaded to the FPGA for standalone algorithm implementation or used with hardware in the loop (HIL) functionality that allows a particular block of the system to be emulated using an FPGA device, this is with the purpose of performing hardware acceleration.

Nowadays the common first step taken by researchers is to test their ideas in Matlab's m-code. Matlab as a system level platform allows a very fast and efficient algorithm implementation of complete systems. Matlab does not include the conception of time; it is more comparable to high level programming languages; has a vast set of libraries or toolboxes in many disciplines; and it is not limited to math or engineering. Matlab has become an indispensable tool in modern electronic design and engineering in general.

If the designer would like to model the system including time as another design dimension, Simulink could be used to design complete dynamic systems that are time aware and also include a large number of libraries or toolboxes for a large number of disciplines.

2.3 System architecture

When evaluating an algorithm, the designer is mostly concerned on modeling a system. One of the problems is that the final implementation cannot be readily extracted from this system level modeling easily. There are different levels of system models, some models can be bit accurate and/or cycle accurate.

In a bit accurate model, the system traditionally has been modeled using floating point precision, and then the algorithm has been converted into fixed point precision for efficient implementation. At this stage the main concern is that the signal to quantization noise ratio (SQNR) will dictate the losses due by the effects of for example: quantization, rounding and saturation. This transformation stage can be performed in Matlab/Simulink, SystemC and C/C++. A bit accurate model will have a very close representation of the final implementation in terms of hardware cost and performance. One problem here is that the internal precision of the operation is difficult to model until the final architecture has been decided.

In a cycle accurate model, the systems are architected such that the generated hardware corresponds one to one to the behavioral model in terms of time execution. The advantage is that a true bit accurate and cycle accurate simulation can be obtained, but at much higher

simulation speed to their RTL or gate level simulations. In the author's experience, this model has not been used much in the past, since it is tied up with a fixed architecture so the conversion to RTL is straightforward with no ambiguities.

After the fixed point precision has been proposed, it is traditionally coded either in a high level language or in a hardware description language. Of course at this stage the model can continue to be modeled in Simulink. Typically an architectural description is being pursued at this level and the model should closely represent the hardware to be implemented.

What is interesting is that at this stage, there are at least from two or more "system models." One very common error is to not update the higher level with architectural changes once high level modeling stage has "finished", this could lead to inaccuracies on the implementation since it is no longer compared with the "golden" model anymore. As we mentioned, the models can get out of synchronization due to lack of communication between the system's team and the implementation's team. It is of extreme importance throughout the life of the project to have all models updated to reflect the latest changes in both SLD and RTL since each one represents a running specification of the system at different levels of hierarchy.

2.4 System testbench

A testbench is created at the behavioral level, what this means is that the testbench is not to be synthesized, that is why the testbench can include language constructs that represent stimulus and analysis rather than processing and are not directly synthesizable. The testbench is designed to test a "black box" or commonly known as the Device Under Test (DUT), generate inputs, measure responses and compare with known "golden" vectors. One very useful feature in Verilog HDL is to be able from the testbench to descend into the design hierarchy and probe on internal signals that are not available at the interface level. VHDL 2008 includes hierarchical names for verification as well.

A rule of thumb says that when a design is "finished", it is just 30% complete and the validation and verification (V&V) stages will start to cover the remaining 70% effort to have a verified finished design. There are different methodologies to accomplish this and unfortunately Verilog HDL and VHDL have not been robust enough to allow complete and efficient design verification. Due to the later, several proprietary verification languages evolved and recently several methodologies such as Open Verification Methodology (OVM)(Cadence, 2011), Verification Methodology Manual (VMM)(Accellera, 2011) and Universal Verification Methodology (UVM)(Synopsys, 2011c) have been developed to fill the gap between HDL and proprietary verification languages including a common framework for verification. The common denominator in all these methodologies is the use of System Verilog as the driver of all three. System Verilog is evolving as the verification and design solution language since it contains the best of design, synthesis, simulation and verification features, the versatility of the HDLs, and it is designed for system level verification.

Talking about levels of design abstraction, another very common approach is to use the popular C and C++ languages to describe algorithms to be implemented in hardware. We

have found that several Electronic System Level (ESL) design tools generate SystemC testbenches that could be used as standalone applications as well as integrated into event driven simulators that are the core when designing hardware implementations. Some examples are Pico Extreme from Synfora (acquired by Synopsys and now is SymphonyC compiler)(Synopsys, 2011b) and CatapultC from Mentor Graphics (CatapultC is more like C++rather than SystemC) (MentorGraphics, 2011).

We have talked about Matlab/Simulink being used at the system level design phase. In order to take full advantage of a common testbench, a hardware design could rely entirely on this platform for rapid prototyping by accomplishing transformations at the level of modeling hierarchy.

Once a design is transformed for example from Matlab m-code to Simulink, or perhaps the design was started in Simulink directly, there are a series of custom libraries that allow the designer to transform their design directly into hardware and keep the original Simulink testbench to feed the hardware design. The design could be verified by generating HDL RTL and by running event driven simulations side by side the Simulink engine and compare with the original Simulink model to verify that the RTL code generated matches the desired abstracted model. Not only a standalone simulation is conceivable, it is possible to download the application directly into an FPGA and generate excitation signals and receive the data back in Simulink. This allows to verify hardware performance at full speed or to accelerate algorithm execution that will take a long time on an event driven simulator. There are several products with similar capabilities such as National Instrument's LabView (NI, 2011) that also allows the option to have "Hardware In the Loop" (HIL) as a way to accelerate computing performance by implementing the algorithm directly in hardware.

The philosophy at this level is to try to reuse the testbench as much as possible to verify correctness of the design at a very high level of abstraction and to code a single testbench that could be used at the system level, while still being able to run the components at single levels of abstraction, namely behavioral, RTL and gate level.

3. Fixed point number representation

This section will cover the different formats used to represent a number using fixed point precision. In addition, the effects of truncation, rounding, and saturation will be covered. SystemC provides a standard set of fixed point types that have been also adopted and adapted by electronic system level (ESL) tools. We will talk about SystemC's fixed number representation. We will talk also about traditional RTL fixed point implementations and the required hardware, complexity and performance.

3.1 SystemC fixed point data types

SystemC includes the *sc_fixed* and *sc_ufixed* data types to represent fixed point signed and unsigned numbers the syntax to include these in a SystemC program is the following:

```
sc_fixed<wl, iwl, q_mode, o_mode, n_bits>
sc_ufixed<wl, iwl, q_mode, o_mode, n_bits>
```

where

wl: total word_length
iwl: integer word length
q_mode: quantization mode
o_mode: overflow mode
n_bits: number of saturated bits

Quantization modes: SC_RND, SC_RND_ZERO, SC_RND_INF, SC_RND_MIN_INF, SC_RND_CONV, SC_TRN, SC_TRN_ZERO

Overflow modes: SC_SAT, SC_SAT_ZERO, SC_SAT_SYM, SC_WRAP, SC_WRAP_SM

For example if we would like to declare a signed integer variable with 16 total bits of which 8 bits are integer, we declare:

```
sc_fixed<16,8> number;
```

As can be observed in Figure 2, the 16 bit number will contain 8 integer bits and 8 fractional bits. The maximum number that can be represented is $2^7 - 2^{-8} \approx 128$ and the minimum number will be $-2^7 = -128$ with a $2^{-8} \approx 3.9 \times 10^{-3}$ resolution. By default, the number will have a quantization mode of $q_mode = SC_TRN$ which means that the number precision will be truncated after each mathematical operation or assignment, and the number will have an overflow mode $o_mode = SC_WRAP$ which means that the number will wrap from approximately 128 to -128. The different modes allow for flexibility in the rounding and saturation operations that are useful to limit the number of bits enhance the SQNR and also to allow infrequent numbers to be saturated and save on the total number of bits. Of course, the price is additional hardware and probably timing to perform these operations.

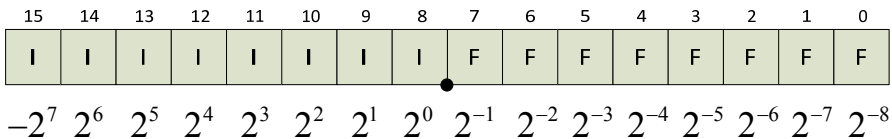


Fig. 2. *sc_fixed*<16,8> representation of a fixed point number.

There are too many ways to describe fixed point notations and representations, but we think that this represents a commonly used format in most of ESL tools that we have explored.

4. Floating to fixed point design considerations

A practical implementation of a wireless communication algorithm involves the conversion of a floating point representation into a fixed point representation. This process is related to the optimum number of bits to be used to represent the different quantities through the algorithm. This process is performed to save complexity, area, power, and timing closure. A fixed point implementation is the most efficient solution since it is customized to avoid waste of resources. The tradeoffs against a floating point implementation are noise, non-linearities and other effects introduced by the processes of: quantization, truncation, rounding, saturation and wrapping among the most important.

Both the floating point and the fixed point solutions have to be compared against each other and one of the most common measure of fixed point performance is the signal to quantization noise ratio (SQNR) (Rappaport, 2001).

Several tools are available to allow the evaluation of a fixed point implementation against a floating point implementation. One of the most important factors is the dynamic range of the signal in question. Floating point adapts to the signal dynamic range, but when the conversion is to be done, a good set of statistics has to be obtained in order to get the most out of the fixed point implementation. The probability density function of the signal will give insight on the range of values that occur as well as their frequencies of occurrence. It may be acceptable to saturate a signal if overshoots are infrequent. We need to carefully evaluate the penalty imposed by this saturation operation and the ripple effects that it could have. This process allows to use just the necessary number of bits to handle the signal most of the time, thus saving in terms of area, power and timing. In section 5.3, we talk about some of the little steps that have to be taken throughout the design in order to save in power consumption. As mentioned, power consumption savings start at the system level architecture throughout the ASIC and FPGA methodologies.

Sometimes the processed signal could be normalized in order to have a unique universal hardware to handle the algorithm. It is very important to take into consideration the places where the arithmetic operations involve a growth in the number of bits assigned at each operation. For example, for every addition of two operands, a growth of one bit has to be appended to account for the overflow of adding both signals. If four signals are added, only a growth of two bits is expected. On the other hand a multiplication creates larger precisions since the number of bits in the multiplication result is the addition of the number of bits of the operands and also it has to be taken into account if the numbers are signed or unsigned.

The fixed point resolution at every stage needs to be adapted and maintained by the operations themselves and specific processing needs to be done to generate a common format. These operations are the truncation, rounding, saturation and wrapping covered briefly for SystemC data type in section 3.

A nice framework of the use of fixed point data types that could be incorporated into C/C++ algorithm simulations are the SystemC fixed point types available in the IEEE 1666™-2005: Open SystemC Language Reference Manual (SystemC, 2011). There are some other alternatives to fixed point data types such as the Algorithmic C Datatypes (Mentor-Graphics, 2011) that claim to simulate much faster than the original SystemC types and used in the ESL tool CatapultC. The ESL tool Pico Extreme uses the SystemC fixed point data types as the input to the high level synthesis process.

Matlab/Simulink also has a very nice framework to explore floating to fixed point conversion. When hardware will be generated directly from Simulink, it is very natural to alternate between floating point and fixed point for system level design. Designs that are targeted for Xilinx or Altera FPGAs could naturally use this flow and reuse the floating point testbench to generate the excitation signals that could be used within the Matlab/Simulink environment in for example Hardware in the Loop (HIL) configuration or fed externally to the FPGA using an arbitrary waveform pattern generator.

Another very useful tool for creating executable specifications in C++ is to use IT++ (IT++, 2011) libraries available for simulation of communication systems.

Each EDA vendor has a different set of tools that allow designers to make the implementation of floating point to fixed point as seamless as possible. This conversion process is a required step that cannot be avoided and traditionally it has been done manually and by matching the results of the Golden model against HDL RTL simulation. Sometimes this comparison is bit accurate, but in some cases the comparison is just done at the SQNR level due to the difficulty to model all the internal operations and stages of a particular hardware implementation.

5. Register transfer level design

Once a system has been verified for performance and has been converted from a floating to a fixed point representation, the specifications are passed to the register transfer level (RTL) design engineer to come up with an architecture that will achieve the desired performance, while consuming minimum power at the right frequency of operation, using minimum area, sharing resources efficiently, reusing as much components as possible, and coming with an optimum tradeoff between hardware and software implementations. We can see that this is not usually an easy task to perform, even for experienced designers.

5.1 Architecture

In this section we will give an overview of the importance of the architecture in RTL design. Examples of different architectures for complex multipliers, finite impulse response (FIR) filters, fast Fourier transforms (FFT) and Turbo Codes will be given comparing their complexity, throughput, maximum frequency of operation and power consumption.

When an efficient architecture is sought, each gate, each register, each adder and each multiplier counts. Sometime it is a good approximation at the system level to count the number of arithmetic operations to get an initial estimate of the silicon area that will be used for the algorithm. While this is a crude approximation it is a very good start point to allocate resources on the System on a Chip (SoC). Many companies have spreadsheets that contain average values for different operations in a particular technology; based on hundredths of designs. The architecture task is to find the optimum implementation of a particular algorithm while accomplishing all the above referred design parameters.

When an algorithm is implemented, what will be the final underlying technology for implementation? ASIC or FPGA; or will it be driven by software and just primitive building blocks will be used as coprocessors or hardware accelerators. Whenever a product needs to be designed on an application area that continues to grow and generate new algorithms and implementation such as video processing, sometime an analytics engine must be architected that will provide co-processing or hardware acceleration by implementing the most common image processing algorithms. This idea could be applied to any communications system or signal processing system where a solution could include a common set of hardware accelerators or coprocessors that realize functions that are basic and will not easily change. One very good example is the TMS320TCI6482 Fixed-Point Digital Signal Processor (Texas-Instruments, 2011) that is used for third generation mobile wireless infrastructure

applications and contains three important coprocessors: Rake Search Accelerator, Enhanced Viterbi Decoder Coprocessor and Enhanced Turbo Decoder Coprocessor.

So the question is: When implementing a particular algorithm, how can we architect it such that it is efficient in all senses (are, power, timing) as well as versatile? The answer depends on the application. That is why hardware/software partitioning is a very important stage that has to be developed very carefully by thinking ahead of possible application scenarios. In some cases there is no option, and the algorithm has to be implemented in hardware, otherwise the throughput and performance requirements may not be met. Let's explore briefly some practical examples of blocks used in wireless communication systems and just brainstorm on which architectures may be suitable.

Finite Impulse Response Filters

An FIR filter implementation can be thought as a trivial task, since it involves the addition of the weighted version of a series of delayed versions of an input signal. While it seems very simple, we have several tradeoffs when selecting the optimum architecture for implementation. For an FIR filter implementation we have for example the following textbook structures: Transversal, linear phase, fast convolution, frequency sample, and cascade (Ifeachor, 1993). When implementing on for example on FPGAs, then we found for example the following forms: Standard, transpose, systolic, systolic with pipelined multipliers (Ascent, 2010).

Most of the FPGA architectures are enhanced to make more efficient the implementation of particular DSP algorithms and the architecture selection may fit into the most efficient configuration for a particular FPGA vendor or family. If we are targeting ASIC, then the architecture will be different depending on the library provided by the technology vendor. When implementing an FIR or any other type of filter or signal processing algorithm, we need to evaluate the underlying implementation technology for tuning the structure for efficient and optimum operation.

Turbo Codes

One interesting example is on Turbo Codes, while the pseudo-random interleaver is supposed to be "random", there has been a pattern defined on how the data could be efficiently accessed. Some interleavers are contention free, while some others have contentions depending on the standard. For example, one of the major differences on the third generation wireless standards namely 3GPP(W-CDMA) and 3GPP-2 (CDMA2000) is on the type of interleaver generator used, this means that to a certain degree it would be possible to design a Turbo Coder/Decoder that could easily implement both standards.

The purpose of an efficient implementation of an interleaver hardware is to have different processing units accessing different memory banks in parallel, some examples on the search for common hardware that could potentially be used for different standards are shown in (Yang, Yuming, Goel, & Cavallaro, 2008), (Borrayo-Sandoval, Parra-Michel, Gonzalez-Perez, Printzen, & Feregrino-Urbe, 2009) and (Abdel-Hamid, Fahmy, Khairy, & Shalash, 2011). The architecture is a function of the standard and sometimes it is very difficult to find a "one architecture fits all" type of solution and in some case to make the interleaver compatible with multiple standards, on-the-fly generation is the best approach, but there can be irregularities or bubbles inserted into the overall computation. This is one of the challenges

in mobile wireless that sometimes is easier to implement complete different subsystems performing efficiently one particular standard, rather than having an architecture that could perform all. This is the case in mobile cellular second generation GSM (Global System for Mobile Communications, originally Groupe Spécial Mobile) and third generation cellular W-CDMA (wideband code division multiple access) that minimum reusability could be achieved and to a certain extent there are two complete wireless modems implemented for each standard.

Fast Fourier Transform

Many of the modern wireless communications algorithms migrated from the CDMA to the Orthogonal Frequency Division Multiple Access (OFDMA) technologies. One of the main reasons to transfer to a completely new technology might have been that the current state of the art on integrated circuit design allowed the efficient implementation of algorithm architectures that were not previously convenient to implement in hardware. This is the case of the Fast Fourier Transform (FFT) which is the core of Orthogonal Frequency Division multiplexing (OFDM) and its derivatives such as OFDMA (Yin & Alamouti, 2006).

OFDM and FFT techniques are not new, as a matter of fact they have been around longer than many of the current wireless technologies. What is new, is the feasibility of the algorithms to be implemented on silicon. An efficient architecture implementation for a pipelined FFT (Shousheng & Torkelson, 1998) has been used as a benchmark for hardware implementation of the FFT algorithms, this technique allows all hardware units to be used at all times once the pipeline is full and is very convenient for FPGA or ASIC implementation.

We will just briefly talk about this on section 10, since it is one example that comes with the FPGA libraries and the purpose of this chapter is not to develop a new FFT form, but rather to see how it can be implemented.

5.2 Maximum operating frequency

While it could be easy to convert an algorithm from floating point to fixed point and to identify architectures for its implementation, the final underlying technology should be taken into account to determine the maximum operating frequency and in some cases the required level of parallelism and/or pipelining. It can be true that an algorithm designed for FPGA will run without major modifications on ASIC, but the reverse is not always true. FPGAs are used widely to perform ASIC emulation, but it does not make much sense to have two different versions of the algorithm running on either technology, since this could invalidate the overall algorithm validation. Sometimes the same code could be run, but in slow motion on FPGAs if real time constraints are not required. If real time is a factor, only some of the low throughput modes could be run on the FPGA platform and simulated for ASIC.

5.3 Power consumption

Power consumption in mobile devices is a crucial part of the algorithm selection and it is tightly coupled to architecture's implementation, frequency of operation, underlying technology, voltage supply, and gate level node toggle rates to give some examples. In this

section we will cover some of the important features to be considered when designing power optimized algorithms implementations.

When designing digital systems we all know that a magic button exists that reduces power consumption to the minimum. Unfortunately this is not the case, the magic button does not exist and power savings start at the system level design, the architecture selection, the RTL implementation, the operating frequency, the integrated circuit technology chosen, the gate clocking methodology, use of multi- V_{dd} and multi- V_{th} technologies, and leakage among some of the most important factors. In reality power savings are being done in small steps starting from efficiency at the system and RTL level design. One power saving criteria is: if you do not have to toggle a signal, don't do it! Power consumption is a function of the frequency of operation, the load capacitance and the power supply voltage. On average, the gate level nodes switch at around 10% to 12%, while an RTL level simulation could have toggles close to 50% meaning that all units are being used all the time and there is no waste in terms of hardware resources.

When deciding the fixed point representation, every bit in the precision counts towards the total power consumption, the number of gate levels between registers the load capacitance of each node. If we decide to include saturation and/or rounding, there are additional gates required to perform these operations. The cost of additional hardware can be worth the gates if the bit precision is reduced from a system with a wide dynamic range that takes into account no overflow for signals that can have very large excursions but are very infrequent. So what could be the best tradeoff between complexity, fixed point precision, internal normalizations, and processing? There is not a single solution to the problem, the best will be to statistically characterize the signals being handled to find out their probability distributions and then based on these determine the dynamic range to be used and if saturation/wrapping and truncation/rounding could be used and within these which methods to apply as mentioned in section 3.

Power consumption depends on the circuit layout as well, while old technologies used to be characterized in terms of gate delays, input capacitance and output load driving capacitance, the end game has changed and modern technologies have to take into account the effects of interconnection delays due to distributed resistance, inductance and capacitance. The solution to the power consumption estimate is not final until the circuit has been placed and routed and transistors are sized. If an FPGA implementation is sought, a similar approach is taken but control is coarser due to the huge number of paths that the signals have to flow in order to be routed among all resources.

Another important factor are the power supply V_{dd} and the threshold voltage V_{th} of the transistors. These two factors control the voltage excursion of the signals and most important the operation region of the transistor. Most of the digital logic design rules assume that the transistors are operating in saturation, power is consumed while transitioning through the active region and this is the region where you want to get out as fast as possible. A transistor operating under saturation regime has a quadratic transconductance relation of the current I and the input gate voltage V_g . When a transistor is not in saturation, it could be in linear region or even in sub-threshold. A transistor in the latter does not have a quadratic, but an exponential transconductance relation. While this is the most power efficient operating regime, it is also the slowest. Many circuits that need

very low power consumption can work in sub-threshold, but there is a huge variability and precision constraints. Most of these designs involve linear analog mode operations.

So what is the secret formula to design power efficient devices? The answer is discipline! Try to save as much as possible at each level in the design hierarchy. If it is in software, set the processor to sleep if there is nothing important to do. If it is hardware, do not toggle nodes that do not require to be toggled, gate the clocks so you can lower power consumption in blocks not used, reduce powers supply V_{dd} to the minimum allowed for efficient operation of the algorithm and design using just the right number of bits. More techniques for low-power CMOS design have been published and good overviews are given in (Chandrakasan & Brodersen, 1998) and (Sanchez-Sinencio & Andreou, 1999).

6. Electronic System Level Design

Electronic System Level Design (ESL) design has come from a promising technology to a reality. Companies such as Cadence, Mentor Graphics and Synopsys have their own ESL tools and have integrated these into their System on a Chip (SoC) design flows. In this section we will address some of the most important features of ESL which are architecture exploration, power consumption estimation, throughput, clock cycle budgets allocated, and the overall integrated verification framework from untimed C/C++ golden model, all the way to gate level synthesis.

One of the advantages of ESL tools is that the same testbench used to design a block could be reused at all levels of abstraction thus minimizing the probability of introducing errors at different levels of the implementation. While RTL design requires thinking very carefully on a target architecture, ESL allows exploring different architectures and taking tradeoffs using a high level description of the algorithm, and avoids the designer to go to the RTL level to verify block's performance. We will go through examples of an OFDM FFT implementation as well as MIMO signal processing. ESL niche applications are hardware accelerators that traditionally are hooked to a microcontroller platform such as an ARM processor and handle data processing intensive operations. This is a common practice in SoC design, several intellectual property (IP) vendors concentrate their products in offering very high performance blocks that interface with a common bus architecture such as AMBA.

7. FPGA implementation

For FPGA implementations we could always resort to the traditional RTL implementation of the algorithm. For this section we will resort to Mathwork's Matlab/Simulink implementations of particular algorithms by the automatic generation of RTL code to be either downloaded to the FPGA and to be tested standalone or to the Matlab/Simulink testbench that could be used to drive the simulation and the actual RTL code will be executed in the FPGA. The latter is referred as hardware in the loop (HIL).

We will give examples of: converting a chaotic modulator/demodulator from Matlab code to a Simulink model; to a Simulink model using Altera DSP builder blocks; and demonstrating the algorithm working on a development board after digital to analog and analog to digital conversions.

In FPGAs the pool of resources is fixed. Depending on the particular algorithm, it could be better placed in one of the different families of FPGAs available by different vendors. Datapath architectures can be very efficiently instantiated on FPGAs since most of building blocks included in these devices are designed for very high performance digital signal processing algorithms. We will talk about the tradeoffs when FPGA utilization is low and high and the effort to place and route (P&R) as well as timing closure.

8. ASIC implementation

Most of the wireless communication algorithms would have two versions: one for wireless infrastructure that needs high performance and power is important but not critical since it is always connected to an external power source, and another for mobile wireless devices in which performance is a requirement but power has to be optimized in order to make the device usable, power efficient and competitive. In this section we will explore these two types of implementation in applications specific integrated circuits (ASIC). We will give an example of a turbo code interleaver/de-interleaver that had been implemented and verified using simulation and an FPGA platform and the changes required to take it to an ASIC implementation.

9. Hardware acceleration

Sometimes it is not possible to evaluate an algorithm using regular simulation techniques due to the computing power that is required to perform these tasks. SoC designs are a good examples of these constraints, not all block could be implemented and verified at the gate level in simulation due to the fact that it will take from hours to weeks to perform these simulations. For these cases it is common to use FPGAs as hardware accelerators or ASIC emulators. ESL tools are very efficient in generating these type of blocks that can be either instantiated for FPGA or ASIC and the only real difference is on the characterized libraries used as well as the system clock frequency.

The basic requirements while designing custom datapath components is to create hardware accelerators that could work as standalone blocks. Normally these components will become part of a large SoC. Many of the current embedded products recently designed are composed of a microcontroller such as an ARM core, a standard bus such as AMBA, and a series of Intellectual Property (IP) blocks that realize specific functions that require high performance and low-power. This is mostly true on cellular mobile devices, while for base stations a dedicated Digital Signal Processor (DSP) could be used since throughput is a more important constraint than power consumption. It is worth mention that these designs could be done in the same technology geometry, but with different characteristics: base station would most likely use a high performance, higher threshold voltage and large leakage process while the mobile device will be constrained to medium performance, very low leakage process and low and probably variable threshold voltages.

Some examples of systems that are designed as hardware accelerators in cellular technologies are:

- Equalizers
- Viterbi, Turbo and LDPC decoders

- OFDM Modems
- Rake receivers
- Correlators
- Synchronizers
- Channel estimators
- Matched filters
- Rate matching filters
- Encryption/decryption
- Modulator/demodulator
- Antenna diversity and MIMO processing

The question is which functions will run on software and which functions will run on hardware. This lies in the gray area of hardware/software partitioning. There are different specifications that need to be considered before taking an educated decision. In theory, anything that could be done in hardware could be done in software and vice versa (of course having an infinitely fast processor with a humongous bus bandwidth and a large number of I/Os). We must carefully evaluate the hardware components to be implemented since no field upgradeability will be possible once an ASIC has been manufactured; we need to find the equilibrium where a firmware patch could potentially get rid of any anomaly not detected at verification and validation time.

In particular, the author worked for many years in teams concentrated on hardware accelerators, but all these components were part of a SoC where traditionally an ARM processor was used with a standard interconnect such as AMBA (ARM, 2011) or OCP (OCP, 2011) and the hardware accelerators were mapped as peripherals in the processor memory space. The ASIC design was first simulated, then emulated on a large FPGA platform at a constrained speed and then the ASIC could finally be developed.

In academia we are more involved with FPGA designs and in particular the platforms being used for teaching include the possibility of a soft core processor. For the author's particular case the platform is Altera and the soft core processor is the Nios II. It is interesting to find that a C to RTL application program exists that allows functions implemented in software could be converted into hardware accelerators. The application is C2H (Altera, 2011b) and even that the author has not been able to test it, it looks promising since it allows the exploration of different hardware/software partitions that could impact the total silicon area, performance, power and cost of a particular application (Frazer, 2088). In the case of FPGA design it could lead to be able to reduce costs or performance by moving back and forth different FPGA migration devices that are pin compatible, but vary in the number of logic elements available, the number of I/O pins available and cost. An equivalent tool exist from Xilinx called Auto-ESL (Xilinx, 2011a) that generates code from C/C++/SystemC.

10. Hardware implementation examples

10.1 MOC digital communications system implementation

In this design example, we will walk through the steps required to implement a mutually orthogonal chaotic (MOC) digital communications system (Glenn, 2009) algorithm architected in Simulink to run on FPGA hardware and the constraints imposed by these steps that were not considered in the original design, that affect the systems performance.

The MOC algorithm was coded first in m-code and later converted into a Simulink model. This is shown in Figure 3. The model allows following what the algorithm does without going deep into the details and the model is time dependent. The data rates at the input and output of each block are not shown and this is one of the most important features to consider in a datapath Simulink model.

After looking at the architecture presented for implementation, each of the blocks was substituted by the equivalent Altera DSP Builder available blocks. Some of the blocks have a direct equivalent while some others have to be converted into an equivalent hardware component. This is shown in Figure 4.

Since this block is originally excited by a binary signal, some digital components were used to group the bitstream into a fixed number of bits that will be used to select the modulation waveform. The original Simulink model does not have time restrictions and could potentially generate a waveform with a very large precision, but for practical reasons the implementation is restricted to a particular clock frequency and thus the number of samples to choose for the modulation waveform has an impact on the algorithm performance. A study of the optimum number of samples and the optimum number of bits to represent each modulation waveform had to be done. Each modulated waveforms also could change in sign and or magnitude, for Simulink the operation is just a simple multiplication, but for a hardware implementation it is more efficient to allocate ROM tables and access the correct magnitude and phase. This is similar to storing one quarter of the phase of a sine wave and generates sine and cosine waveforms out of this reduced table. The difference is that the basis functions for this algorithm are chaotic waveforms, then it is difficult to exploit any symmetry property.

In Simulink it is very convenient to add very high level functions such as the modulators and demodulators observed in Figure 3b and Figure 3c, while this may not be required for a baseband algorithm like the one that are implemented on FPGAs. For implementation and testing we decided to work just at the baseband level.

After the model was converted, we compared the values generated by the Simulink blocks simulation against the one generated by using the Altera DSP Builder blocks. The signals were matched and SQNR was computed to validate the approach as well as rate matching was performed to match the samples. The bit sequence and the modulated waveforms are shown in Figure 4d.

The next step is to generate HDL RTL out of the Altera DSP Builder blocks. This is shown in Figure 5a where RTL code is generated, a Simulink simulation is run, followed by a Modelsim RTL simulation and both simulations are compared and the differences are noted. The generated HDL RTL now can be synthesized and programmed into the FPGA for further development. Since for this particular system the excitation is being generated in the test bench by using a Bernoulli random number generator, we decided to use a pseudo random noise (PRN) sequence generator to embed into the FPGA for standalone testing.

The results for the transmitter are shown in Figure 6, where a) is the Altera Cyclone II FPGA testing board with two 14-bit resolution and data rate up to 65 MSPS analog to digital converters and two 14-bit resolution and data rate up to 125 MSPS digital to analog converters. This configuration is suited for testing communication transceiver applications, digital signal processing algorithms and as a platform for various modulation techniques such as the presented in this implementation example.

Figure 6b shows the modulation operation when an all zero pattern is generated. Figure 6c shows the PRN sequence excitation modulation waveforms and Figure 6d shows a screen capture of the MOC modulated waveforms.

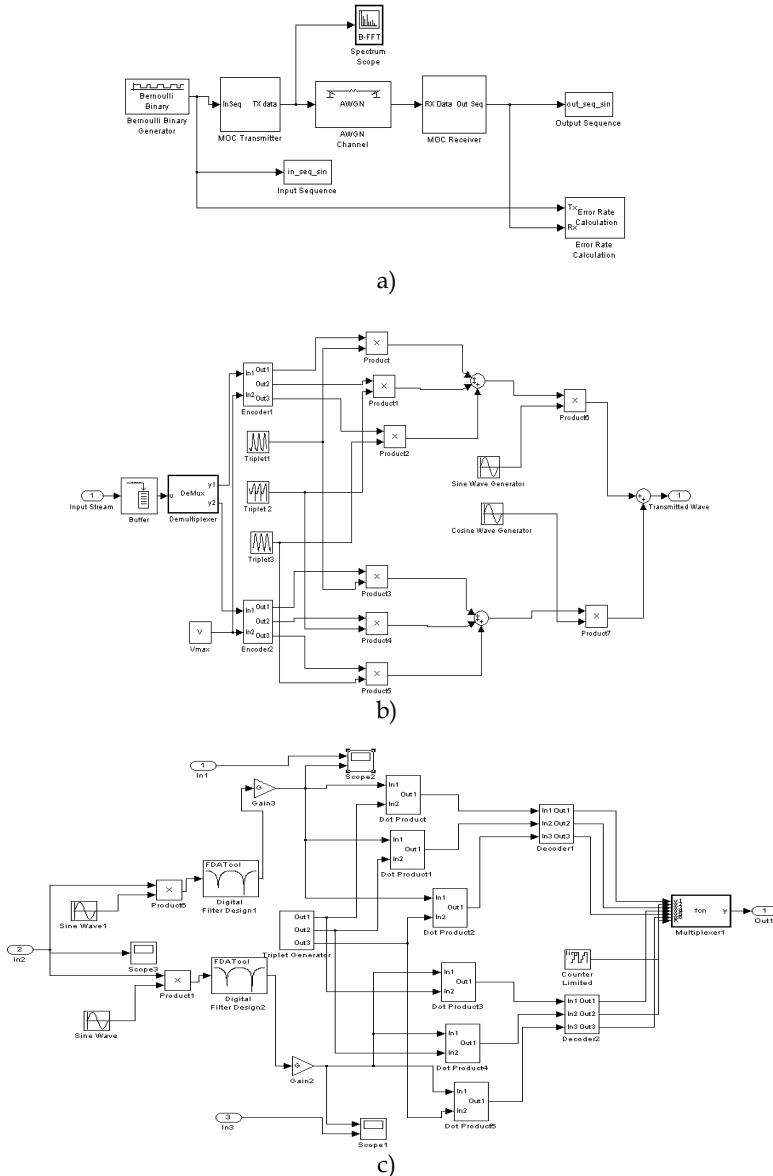
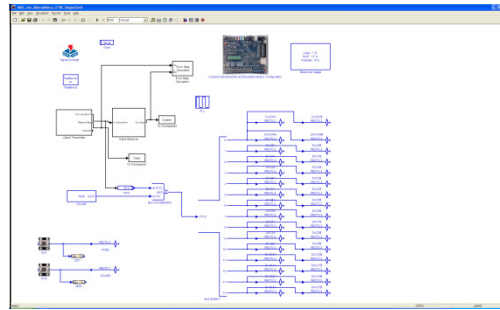
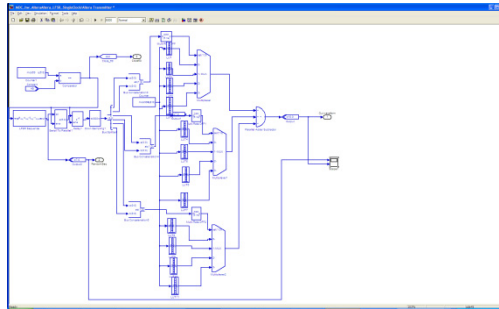


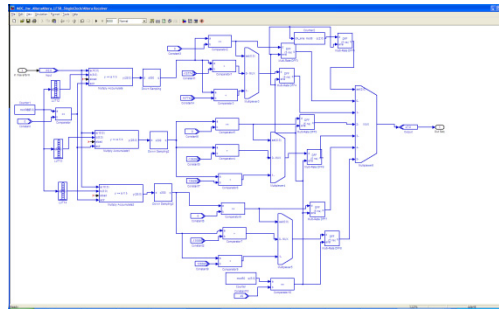
Fig. 3. MOC algorithm architecture implemented as Simulink models.
 a) Complete MOC communications system block diagram including channel modeling.
 b) MOC transmitter block diagram. c) MOC receiver block diagram.



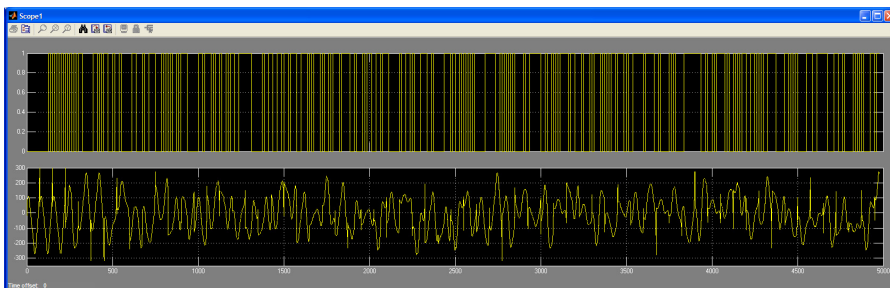
a)



b)



c)



d)

Fig. 4. MOC algorithm transformed to use Altera DSP Builder blocks to automatically generate HDL for FPGA implementation. a) Testbench and interface signals to FPGA. b) Transmitter sub-system. c) Receiver subsystem. d) Simulink simulation waveform.

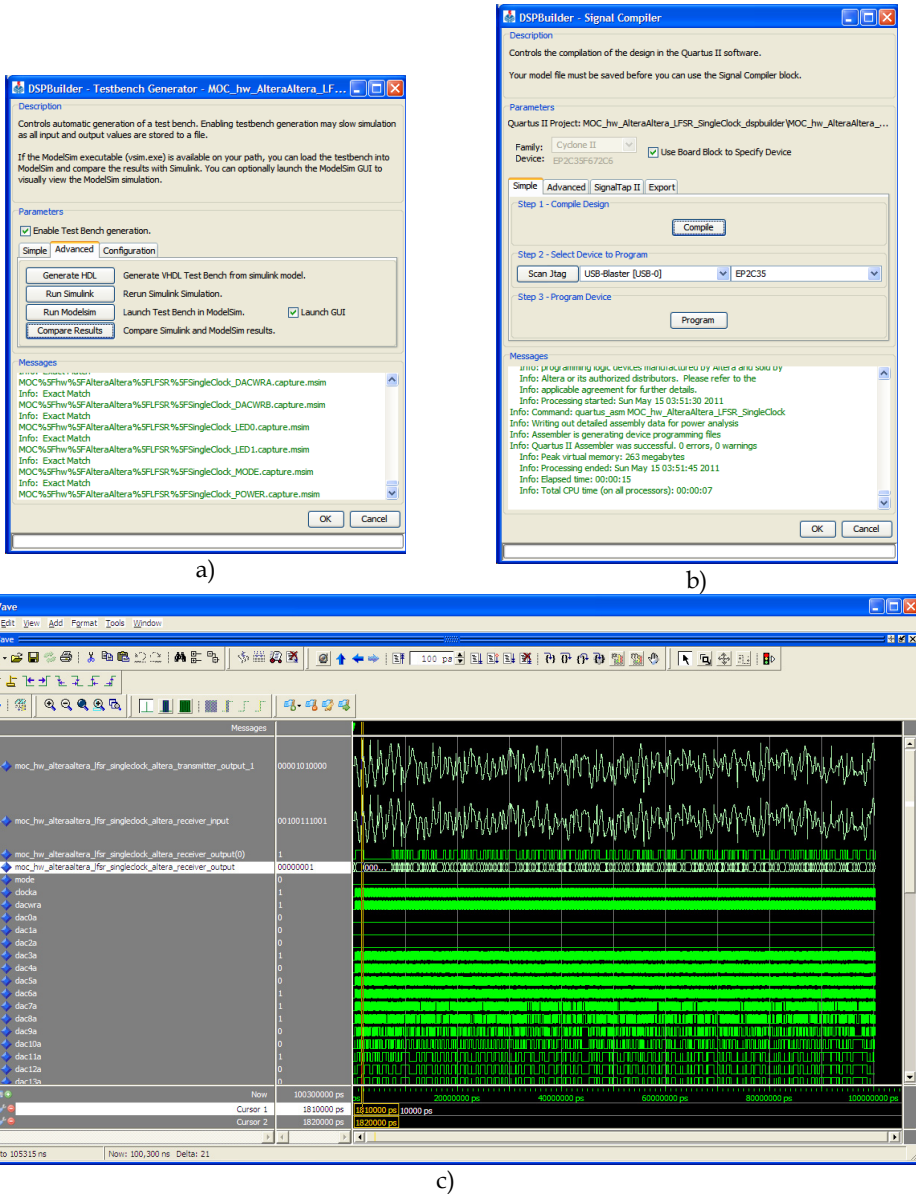


Fig. 5. In addition to a system level simulation within Simulink, it can also control an RTL simulation of the generated HDL code and compare against the system level simulation. a) Test bench generator for RTL simulation. b) RTL HDL simulation of the code generated by DSP Builder. c) Signal compiler for synthesis, place and route, and FPGA programming.

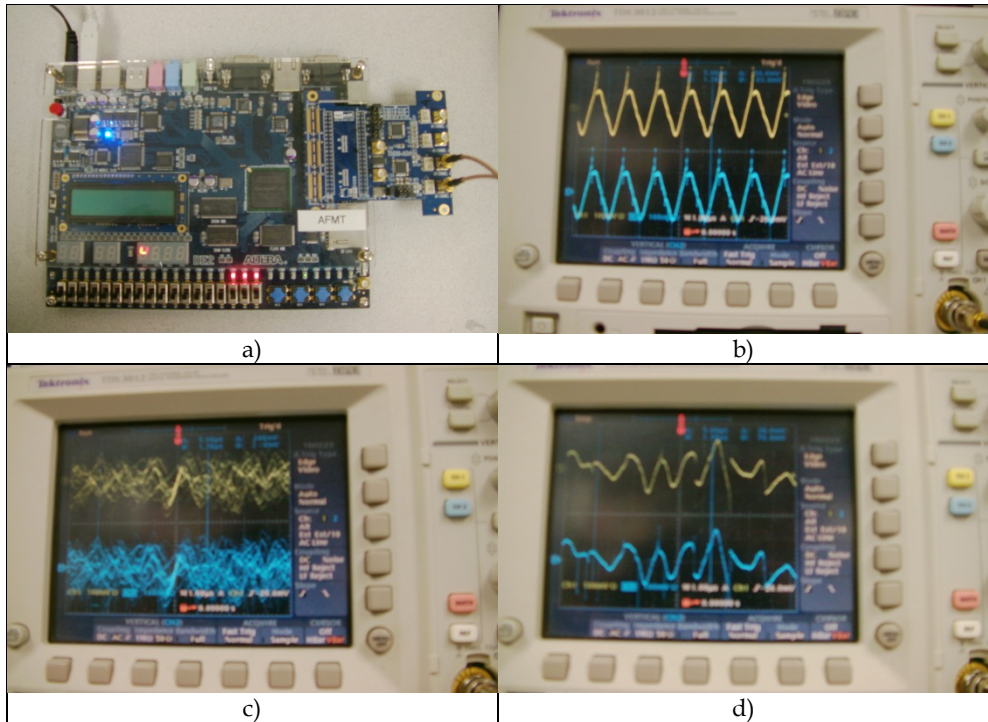


Fig. 6. MOC hardware implementation on an Altera Cyclone II FPGA.

- a) Altera DE-2 with daughtercard dual AD channels with 14-bit resolution and data rate up to 65 MSPS and dual DA channels with 14-bit resolution and data rate up to 125 MSPS.
- b) MOC modulation output when the input is a stream of constant zeros.
- c) MOC modulation output when the input is driven by a PRBN sequence generator.
- d) MOC modulation output snapshot when the input is driven by a PRBN sequence generator.

10.2 Improving the performance of DSP systems for MIMO processing

In the paper “Improving the performance of DSP systems for MIMO processing” (Horner, Kwasinski, & Mondragon, 2011), we explored the efficient implementation of select Multiple Input Multiple Output (MIMO) communications algorithms. Two implementation approaches were considered: adding new instructions to the DSP instruction set and adding a hardware accelerator to the DSP system. Of the two approaches, the second was concluded to be best, as it resulted in notable processing speedups and a more efficient use of the computational resources.

While the research into MIMO algorithms have reached levels of development that show important wireless systems performance improvements, the development of DSP systems to implement them has limited the realization of these algorithms to the simplest and least performing ones. This example addresses this technological gap by studying how to design DSP systems to better handle the increased complexity arising from the particular operations typical of MIMO processing algorithms.

Two hardware co-processors were designed, as shown in Figure 8 one for a Householder decomposition algorithm and one for a Greville pseudo inverse algorithm. These hardware co-processors resulted in a simulated speedup of 2.7 for the Greville algorithm and between 4 and 4.7 for the Householder algorithm.

For the design of the hardware accelerator, Synfora's Pico Extreme (acquired recently by Synopsys) ESL tool was used. The author had previous experience with the tool and the task performed for this work was limited to architecture exploration and to find which ASIC implementation would result in the best compromise between throughput, area, power, and easy of interfacing. The algorithms were written in floating point C code and then converted to fixed point C code by evaluating the impact in performance due to the hardware implementation.

Pico Extreme is a very versatile tool since it is structured as a series of logical steps from running an untimed sequential ANSI C program, to single-to-multi-threaded transformations; to hierarchical block-level resource sharing & scheduling; to automatic retiming and pipelining; to performance and throughput analysis; to rapid exploration of performance impacts of loop unrolling, scheduling, and other optimizations; and to RTL verification among others. The flow methodology is shown in Figure 7.

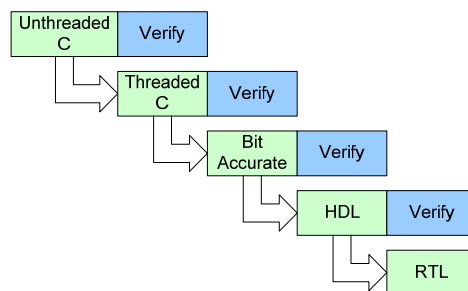


Fig. 7. PICO Extreme design flow.

While this seems to be a dream in which the system designer can implement his design by exploring architectures and trade-offs, then pushing a button and get verified RTL as an output, the reality is that the learning curve of these tools is quite steep and it is not as straight forward as it looks. Even that a very thorough architecture exploration can be performed, the designer still needs to think in terms of hardware when writing the C code to have the same effect as writing in HDL RTL. The C code has to be written in terms of functional units, pipeline stages, memory implementations, operator sharing and general hardware efficiency.

There are two basic methods to specify the design (Synfora, 2009). The number of clock cycles between iteration starts is called II (Initiation Interval) and the number of clock cycles to start all iterations is called MITI (Maximum Inter Task Interval). For this example, MITI can be as small as $N \cdot II$ (where N is the number of loop iterations).

The user is able to provide a target maximum number of clock cycles taken per stage MITI and the tool will select from the library of high-speed components the optimum to achieve higher levels of parallelism at the same time of sharing resources and achieving performance.

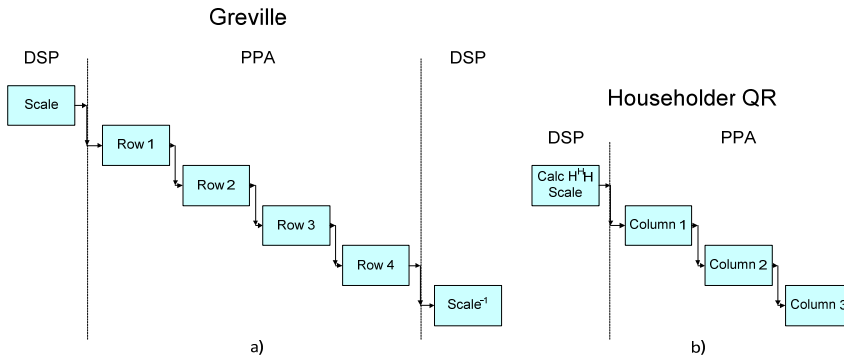


Fig. 8. Processing pipeline for Greville and Householder decomposition methods.

To provide a tradeoff between complexity and speedup, different implementations with different target MITIs were generated. It was noted that as timing constraints tightened, hardware multipliers were switched from two-cycle to one-cycle and the number of multipliers increased to be able to complete complex multiplications (requiring three multiplies) in a single cycle.

MITI timing constraints were used to determine the lowest complexity implementation for each algorithm. The constraints within these ranges of target clock cycles were then used to produce a tradeoff between complexity and resulting speedup. Resulting ranges of targeted number of clock cycles were 230 to 330 for the Householder implementation and 130 to 210 for the Greville implementation.

The resulting speedup was calculated as the ratio of cycles on the DSP-only implementation to the cycles of the DSP-PPA implementation. The resulting silicon area was calculated based on the estimated number of gates given by Pico Extreme and using a characterized CMOS 65nm technology library with an estimate of 854,000 gates per mm². This technology was selected, given that is the one in which the DSP was manufactured and can provide an estimate of the growth of the silicon area for the DSP to enable MIMO processing. A plot of speedup vs. complexity for both clocks and both simulators is shown in Figure 9.

The resulting maximum speedups were close to 2.75 for the Greville algorithm and between 4 and 4.7 for the Householder QR decomposition algorithm. This speedup would result in a large reduction (129 μ s for the Greville implementation and 521 μ s for the Householder implementation) in the amount of time required to compute the channel equalization

matrices for an entire OFDM channel in MIMO communication. There is an upper limit to the speedup, however. Because the DSP is still required for some pre-processing operations, there is an asymptotic limit on the actual speedup achieved. Once the PPA unit is able to compute one stage of the processing pipeline in the same amount of time as the software pre-process, there is little added benefit to faster clock or higher complexity. There is also not a major advantage in the 1 GHz clock over the 500 MHz. While the slower clock would require the more complex implementations to compute faster than the DSP software, the savings on power consumption could outweigh the cost of higher complexity.

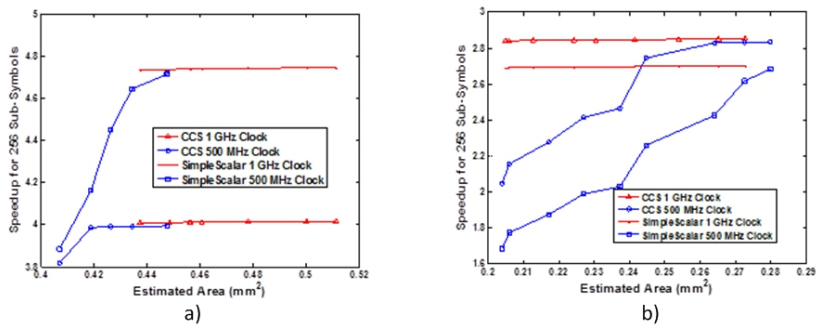


Fig. 9. a) Speedup vs. Complexity for Householder implementation b) Speedup vs. Complexity for Greville implementation.

10.3 OFDM – FFT example

In (Mondragon-Torres, Kommi, & Bhattacharya, 2011), the author proposes the development of an OFDM educational platform that will make use of all the methodologies and tools presented in this chapter with the objective of creating a single system that will allow students to explore different levels of abstraction on hardware design as well as to quantify the effects of the decisions taken on the fixed point precisions as well as all the intermediate signal processing and conditioning through the datapath.

The heart of the OFDM modulation technique lies in the use of the Fast Fourier Transform (FFT), which is a very structured algorithm to convert a time domain signal into the frequency domain and by taking the inverse FFT (IFFT) can be transformed back into the time domain. In Figure 10, a complete digital communication system that employs OFDM modulation is shown (Cho, 2010).

The approach in OFDM systems is to have digital information encoded by traditional phase modulation techniques such as Quadrature Amplitude Modulation (QAM). This modulation technique maps a series of bits into QAM modulated symbols. The number of symbols used for each OFDM frame is traditionally a power of two. Then the IFFT of a block is performed on the frame to convert it back into a time domain representation that can be further processed and sent through the transmitter chain and through the antenna. On the receiver side the process is reversed after frame synchronization by taking the FFT of the received block and obtaining an estimate of the QAM symbols which are mapped back into a series of bits. This sounds pretty straightforward but there are many subtle details that

could be investigated in terms of the effects of: quantization, distortion, channel noise, multipath propagation, fading, Doppler shift, synchronization, etc.

A very simple implementation of a 256 point FFT is presented in this section as shown in Figure 12. No architectural decisions were performed and a regular textbook implementation is used just to demonstrate some of the capabilities of CatapultC. In Figure 11, technology parameters and some common definitions are shown as reference for the reader. Based on the above definitions, we started to change the system parameters to get a feel of their implications.

In Figure 13 it is shown how by unrolling and pipelining the input and output operations we can drastically reduce the latency. What is the price for this? Answer: Memory bandwidth. We can observe that the area has been maintained constant and this is due to the fact that no memories have been considered in these solutions.

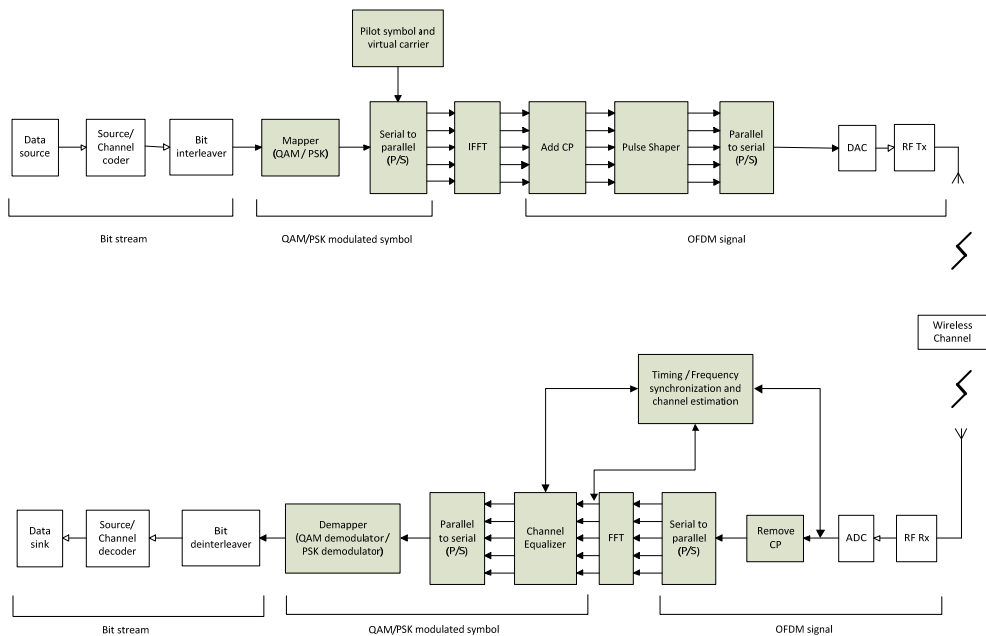


Fig. 10. Digital communications system using OFDM modulation.

Figure 14 and 15 shows the complexity of the solution and we can observe that most of the area is being used in multiplexers to route the signals. On the other hand, more memory will be required for unrolling *printing* and pipelining *reading*. So far we have not touched a single line of code and just by modifying the outer input and output loops we have been able to reduce the latency by 2x at the cost of 2x memory. This is a simple illustration of using the same code to tradeoff performance vs. complexity.

Technology used: Generic CMOS ASIC 90 nm, 200MHz

Definitions

Loop unrolling: Loop unrolling can be used to compute multiple loop iterations in parallel.

Partial unrolling: Computes 'n' copies in parallel

Pipelining: Starts the next loop iteration before the current iteration of the data path contained in the loop has completed

Initial Interval: indicated how often to start a new loop iteration

Latency: Latency refers to the time, in clock cycles, from the first input to the first output

Throughput: Throughput, not to be confused with IO throughput, refers to how often, in clock cycles, a function call can complete.

Fig. 11. Technology used and some common definitions.

```

1  #include <iostream>
2  # include "FixedButterfly.h"
3  # include "Twiddle.h"
4  using namespace std;
5
6  int Bsep, p, Bwidth;
7  int topval, Botval;
8  float1 pi=3.141593;
9  float1 Theta, wnr, wni;
10 float1 Tempr, TempI;
11 float1 xr[N], xi[N];
12
13 #pragma hls_design top
14 void FixedButterfly ( ac_channel<float1> &data_inR, ac_channel<float1> &data_inI,
15                    ac_channel<float1> &data_outR, ac_channel<float1> &data_outI)
16 {
17     //Reading data from the channels bit by bit
18     reading: for(int i=0;i<N;i++)
19     {
20         data_inR.read(xr[i]);
21         data_inI.read(xi[i]);
22     }
23     Stage: for(int s=1;s<=m;s++)
24     {
25         Bsep=Bsep1[s];
26         p=pl[s];
27         Bwidth=Bwidth1[s];
28         coefficients: for(int j=0;j<=Bwidth-1;j++)
29         {
30             wnr=twiddle_real[s][j];
31             wni=twiddle_img[s][j];
32             finalvalues: for(int topval=j;topval<N;topval=topval+Bsep)
33             {
34                 Botval=topval+Bwidth;
35                 Tempr=xr[Botval] * wnr - xi[Botval] * wni;
36                 TempI=xi[Botval] * wnr + xr[Botval] * wni;
37                 xr[Botval]=xr[topval]-Tempr;
38                 xi[Botval]=xi[topval]-TempI;
39                 xr[topval]=xr[topval]+Tempr;
40                 xi[topval]=xi[topval]+TempI;
41             }
42         }
43     }
44     printing: for(int i=0;i<N;i++)
45     {
46         data_outR.write(xr[i]);
47         data_outI.write(xi[i]);
48     }
49 }
50

```

Fig. 12. Program to compute 256 point FFT.

The FFT algorithm itself has not been optimized due to the data dependency among inner and outer loops. Additional pipe stages will need to be implemented in order to break the loop dependency implicit in the direct implementation of the FFT. This probes the point that there the designer has to guide the tool by writing the C code in such a way that the hardware can be inferred.

Another simple tradeoff was executed by increasing the frequency of operation from 100 MHz to 500 MHz as shown in Figure 16. We can observe that the area remained almost constant, while the latency cycles increased by 3% with respect to the 200 MHz implementation baseline, the latency cycles increased by 19%. We can interpret these numbers as the logic required to implement the FFT had a larger critical path, but since the clock was increased 2.5x, the latency time was reduced by 2.0x demonstrating that there is not a linear relationship between the parameters and depends on the implementation given by the particular constraints.

Talking about power, increasing the frequency by 2.5x will have an impact on the power, but at the same time if it is 2.0x faster, we can think for example on reusing the FFT for some other part of the OFDM processor such as computing the IFFT and FFT using the same hardware and sharing it on the time domain rather than have two cores to perform both operations independently.

Solution /	Latency ...	Latency ...	Through...	Through...	Total Area
NoConstraints.v1 (allocate)	1415	7075.00	1417	7085.00	291555.47
UnrollingRead.v1 (allocate)	1176	5880.00	1177	5885.00	292156.30
UnrollingRead & Printing.v1 (allocate)	666	3330.00	667	3335.00	291849.43
Unrolling Print pipeling Read.v1 (allocate)	650	3250.00	652	3260.00	291555....

Fig. 13. Different solutions by selecting different architectural constraints.

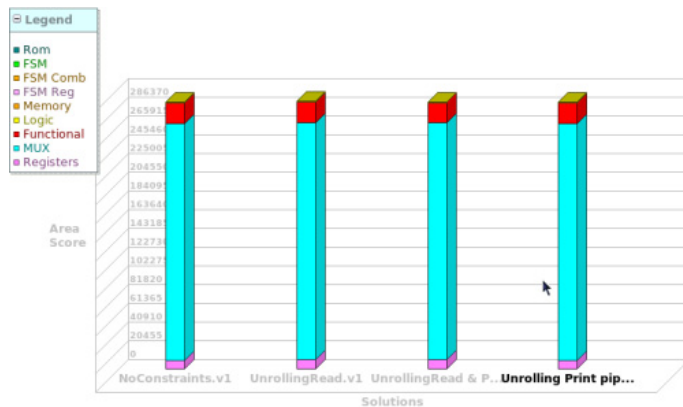


Fig. 14. Graphical view plotting Area.

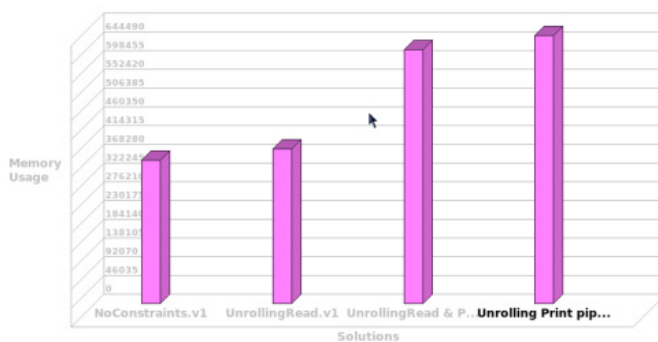


Fig. 15. Graphical View plotting memory usage.

Solution	Latency Cycles	Latency Time	Throughput Cycles	Throughput Time	Total Area
100MHz.v1 (allocate)	1391	13910.00	1393	13930.00	289966.67
200MHz.v1 (allocate)	1415	7075.00	1417	7085.00	291555.47
300MHz.v1 (allocate)	1415	4711.95	1417	4718.61	304308.54
400MHz.v1 (allocate)	1423	3557.50	1425	3562.50	303989.20
500MHz.v1 (allocate)	1695	3390.00	1697	3394.00	300547.22

Fig. 16. Change in performance with change in frequency.

10.4 Hardware In the Loop (HIL)

Hardware in the loop has become a buzz word when designers want to run their algorithm at full speed or at least hundredths or thousands times faster than an RTL or gate level simulation. In SoCs, simulation can take days, weeks and sometimes months, and that depends on the level of detail that is included in the top level simulation. That is why it is important to be able to replace each block by its behavioural, RTL and gate level models in order to refine the level of simulation control and granularity.

Rather than talking about ASIC emulators that are not traditionally available for small companies or universities, we will take a poor’s man approach and show how we can integrate hardware in our computations to able to speed up the testing and processing of algorithms.

Let’s take a closer look at the first level of implementation which is generating automatic HDL code from a Simulink model. Each block or a set of few blocks of the entire communication system can be implemented on hardware this was demonstrated in Section 10.1. So far, we have used an Altera Stratix III FPGA to do system level hardware testing of the Fast Fourier Transform block in the OFDM communication model. For this purpose we have used Hardware in Loop (HIL) block provided by the DSP builder Altera library. This block acts as a link between Simulink and the actual hardware we want to configure.

In modern digital communication systems, the current trend is to implement a pipelined FFT to generate orthogonal sub-carriers. A pipelined FFT generate an output every clock cycle which helps in real-time applications like digital communication systems where data is being continuously fed. We have designed Simulink models to implement FFT using butterfly diagrams which use simple Simulink blocks as well as pipelined FFT which use the advanced block set from DSP Builder. In this section we are going to talk more about the

pipelined FFT for the above mentioned reasons. For more information on the architecture of the pipelined FFT implemented refer to (Shousheng & Torkelson, 1998).

The hardware implementation was done using the Altera's Quartus II version 10.1 and DSP Builder version 10.1. Care must be taken to properly design a Simulink model which would involve block sets from both advanced and standard block sets of DSP Builder. We created this model in layers. The lower level consists of the device block which has the information about the FPGA available in the hardware platform (Stratix III) and the functional blocks that essentially form the FFT. However, on the top level we could only use the signal and control blocks from the advanced block set and other blocks have to be at the lowest level in the design hierarchy.

We make use of the signal compiler and testbench from the standard block set on the top level. The signal compiler is used for creating a Quartus II project, start synthesis, to launch place and route after generating the HDL code. The testbench is used to compare the block level simulations in Simulink and the HDL simulations using Modelsim. Input and output blocks are inserted before and after the subsystem that contains the advanced block set. These blocks have external type parameters to convert from floating or other format handled by Simulink to fixed point as FPGA implementations can only be configured for fixed point. These blocks act as boundaries to the advanced and basic block sets. The procedure to convert the FFT model to HDL, configure the FPGA with the HDL code, and running it from Simulink is detailed below.

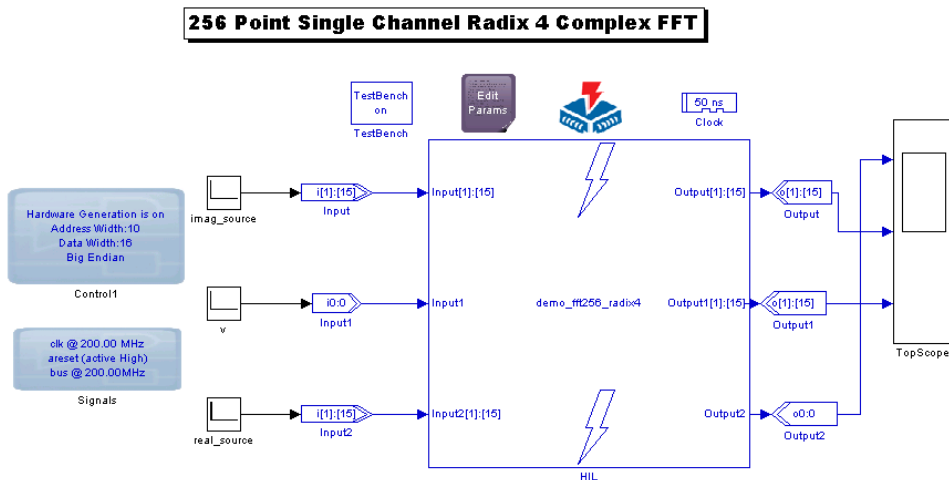


Fig. 17. Hardware In the Loop (HIL) Simulink simulation, actual code runs on the FPGA.

We first run the signal compiler block on the top level to generate HDL code and create a Quartus II project. Then compile the design with Quartus II using the compile option in the signal compiler block. We have now created a Quartus II project for the model and synthesized the HDL code for the same. Now save a copy of this model and instantiate a HIL block on the top layer of the new model from the Altera DSP Builder library found in the standard block set. Open the HIL block and copy the Quartus II project that was earlier

created into the file path. This would generate proper ports for the HIL block. Connect these ports to the appropriate signals. Configure the simulation in burst mode to observe high speed of simulation. In the next menu entry of the HIL block, compile the Quartus II project again, scan JTAG in order to recognize the FPGA device and program it. If we simulate this model it runs at a remarkable speed when compared with the native Simulink simulation. Figure 17 above shows the model which has the advanced block set replaced with a HIL block. This example was modified from the one supplied by Altera to run the FFT on the FPGA platform and to be controlled by the Simulink simulation. We are in the process of converting some other algorithms into hardware following the same methodology to be able to create custom hardware acceleration blocks (Altera, 2007).

11. Conclusions

In this chapter we summarized a few of the methodologies, technologies, tools and approaches that can be taken to convert a wireless communications algorithm into a feasible hardware implementation.

While this chapter is far from being a single methodology to be followed when designing for hardware implementation of wireless communication circuits, we explored many of the practical aspects on how to achieve quick results and also to have a baseline where the final design may compare with.

Push button methodologies are still far from being a reality and even that ESL tools can achieve impressive results and can verify all the way from system level down to gate level against a golden model, there is still some reluctance from the backend teams to rely on automatic tools to do the job. While this approach has been done in automatic place and route in digital systems, ESL has been pushed the level of abstraction one level above RTL design.

What are the advantages of ESL system level design? The most valuable for the author is the ability to explore different architectures and the possibility of generating very complex datapath designs easily with simple constraints and with high hardware reusability.

Can a good RTL designer do it better? The answer is yes if he has all the time to select the best architecture for implementation. SoC design methodologies rely on IP reutilization and to spend the valuable design time just on those blocks that will make the product differentiation.

Due to time to market constraints, design teams cannot spend much time trying to find the best and optimal architecture to implement, sometimes the task are reduced to get the job done on time. One important aspect to remember that most of the products, when the designer announces that the module is ready, it is still no more than 30% of the complete SoC design. Integration, verification & validation, design for testability, design for manufacturability, synthesis, automatic place and route will consume more than 70% of the SoC development time.

Another very important aspect is to be able to run an algorithm on hardware to take advantages of computational speed that for example could be obtained on an FPGA. This is a step required to prove if an algorithm is robust enough. ASIC technologies cannot be

verified using FPGAs, but at least system level emulation can be performed to verify interconnectivity and overall signal flow.

12. Acknowledgements

There are many people that contributed directly and indirectly to the contents of this chapter with their algorithms, ideas for implementation, hard work and enthusiasm. I would like to recognize the following individuals and organizations that contributed in the following areas:

Name	Project
Dr. Chance Glenn Sr.	MOC Digital communications System Implementation
Padma Ragam	MOC Digital communications System Implementation
Nathaniel Horner	Improving the performance of DSP systems for MIMO processing
Dr. Andres Kwasinski	Improving the performance of DSP systems for MIMO processing
Mahesh Nandan Kommi	OFDM - FFT Hardware in the Loop (HIL)
Department of Electrical, Computer and Telecommunications Engineering Technology	Publishing funds.

13. References

- Abdel-Hamid, E. M., Fahmy, H. A. H., Khairy, M. M., & Shalash, A. F. (2011, 15-18 May 2011). *Memory conflict analysis for a multi-standard, reconfigurable turbo decoder*. Paper presented at the Circuits and Systems (ISCAS), 2011 IEEE International Symposium on.
- Accellera. (2011). UVM World: Universal Verification Methodology, 2011, from <http://uvmworld.org/>
- Agilent. (2011). SystemVue ESL Software | Agilent, 2011, from <http://www.home.agilent.com/agilent/product.jsp?cc=US&lc=eng&ckey=1297131&nid=-34264.0.00&id=1297131>
- Altera. (2007). An OFDM FFT Kernel for Wireless Applications (Vol. AN-452).
- Altera. (2011a). Digital Signal Processing, 2011, from <http://www.altera.com/products/software/products/dsp/dsp-builder.html>
- Altera. (2011b). Nios II C-to-Hardware Acceleration Compiler, 2011, from <http://www.altera.com/devices/processor/nios2/tools/c2h/ni2-c2h.html>
- ARM. (2011). CoreLink System IP & Design Tools for AMBA - ARM, 2011, from <http://www.arm.com/products/system-ip/amba/index.php>
- Ascent, S. (Ed.). (2010). *FPGAs for DSP and Communications Course Notes, UCLA Extension, January 24-27, 2011 Course Notes*.
- Bluespec. (2011). Bluespec, Inc., 2011, from <http://www.bluespec.com/>
- Borrayo-Sandoval, H., Parra-Michel, R., Gonzalez-Perez, L. F., Printzen, F. L., & Feregrino-Uribe, C. (2009, 9-11 Dec. 2009). *Design and Implementation of a Configurable Interleaver/Deinterleaver for Turbo Codes in 3GPP Standard*. Paper presented at the

- Reconfigurable Computing and FPGAs, 2009. ReConFig '09. International Conference on.
- Cadence. (2011). OVM-based verification flow 2011, from http://www.cadence.com/products/fv/pages/ovm_flow.aspx
- Chandrakasan, A., & Brodersen, R. (1998). *Low power CMOS design / edited by Anantha Chandrakasan, Robert Brodersen*: Piscataway, NJ IEEE Press, 1998.
- Cho, Y. S. (2010). *MIMO-OFDM wireless communications with MATLAB*: Singapore ; Hoboken, NJ : IEEE Press : J. Wiley & Sons (Asia), c2010.
- Frazer, R. (2008). Reducing Power in Embedded Systems by Adding Hardware Accelerators, 2011, from <http://www.eetimes.com/design/embedded/4007550/Reducing-Power-in-Embedded-Systems-by-Adding-Hardware-Accelerators>
- Glenn, C. M. (2009). MOC Technical brief (ECTET, Trans.). Rochester, NY: Rochester Institute of Technology.
- Horner, N., Kwasinski, A., & Mondragon, A. (2011, 22-27 May 2011). *Improving the performance of DSP systems for MIMO processing*. Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.
- Ifeachor, E. C. (1993). *Digital signal processing : a practical approach*. Wokingham, England ; Reading, Mass. :: Addison-Wesley.
- IT++. (2011). Welcome to IT++! , 2011, from <http://itpp.sourceforge.net/devel/index.html>
- Mathworks. (2011). MathWorks - MATLAB and Simulink for Technical Computing, 2011, from <http://www.mathworks.com/>
- Mentor-Graphics. (2011). Algorithmic C Datatypes - Mentor Graphics, 2011, from <http://www.mentor.com/esl/catapult/algorithmic>
- MentorGraphics. (2011). Catapult C Synthesis Overview - Mentor Graphics, 2011, from <http://www.mentor.com/esl/catapult/overview>
- Mondragon-Torres, A. F., Kommi, M. N., & Bhattacharya, T. (2011). *Orthogonal Frequency Division Multiplexing (OFDM) Development and Teaching Platform*. Paper presented at the 2011 Annual Conference & Exposition, Vancouver, BC, CANADA. http://www.asee.org/search/proceedings?fields%5B%5D=title&fields%5B%5D=author&fields%5B%5D=session_title&fields%5B%5D=conference&fields%5B%5D=year&search=mondragon-torres&commit=Search
- NI. (2011). NI LabVIEW FPGA - National Instruments, 2011, from <http://www.ni.com/fpga/>
- OCP. (2011). OCP-IP : Home Page, from <http://www.ocpip.org/>
- Rappaport, T. S. (2001). *Wireless communications : principles and practice* (2nd ed ed.). Upper Saddle River, N.J. : London :: Prentice Hall PTR.
- Sanchez-Sinencio, E., & Andreou, A. (1999). *Low-Voltage/Low-Power Integrated Circuits and Systems: Low-Voltage Mixed-Signal Circuits*, Wiley-IEEE Press, January 1999
- Shousheng, H., & Torkelson, M. (1998, 11-14 May 1998). *Design and implementation of a 1024-point pipeline FFT processor*. Paper presented at the Custom Integrated Circuits Conference, 1998. Proceedings of the IEEE 1998.
- Synfora. (2009). *PICO USER MANUAL - Writing C Applications: Developer's Guide*. (PE-ASIC-UM-WCADG-VER 09.03-6). Mountain View, CA.
- Synopsys. (2011a). Signal-Processing, 2011, from <http://www.synopsys.com/systems/blockdesign/digitalsignalprocessing/pages/signal-processing.aspx>

- Synopsys. (2011b). Synphony C Compiler, 2011, from <http://www.synopsys.com/Systems/BlockDesign/HLS/Pages/SynphonyC-Compiler.aspx>
- Synopsys. (2011c). Verification Methodology Manual for SystemVerilog, 2011, from <http://vmm-sv.org/>
- SystemC. (2011). Home - Open SystemC Initiative (OSCI), 2011, from <http://www.systemc.org/home/>
- Texas-Instruments. (2011). TMS320TCI6482 Fixed Point Digital Signal Processor, 2011, from <http://www.ti.com/product/tms320tci6482>
- Xilinx. (2011a). AutoESL High-Level Synthesis Tool, 2011, from <http://www.xilinx.com/tools/autoesl.htm>
- Xilinx. (2011b). System Generator for DSP, 2011, from <http://www.xilinx.com/tools/sysgen.htm>
- Yang, S., Yuming, Z., Goel, M., & Cavallaro, J. R. (2008, 2-4 July 2008). *Configurable and scalable high throughput turbo decoder architecture for multiple 4G wireless standards*. Paper presented at the Application-Specific Systems, Architectures and Processors, 2008. ASAP 2008. International Conference on.
- Yin, H., & Alamouti, S. (2006, 27-28 March 2006). *OFDMA: A Broadband Wireless Access Technology*. Paper presented at the Sarnoff Symposium, 2006 IEEE.

Gallium Nitride-Based Power Amplifiers for Future Wireless Communication Infrastructure

Suramate Chalermwisutkul

*The Sirindhorn International Thai-German Graduate School of Engineering
King Mongkut's University of Technology North Bangkok
Thailand*

1. Introduction

Progress in wireless communication technology has enabled applications which were unthinkable as the first digital mobile phone came into the market. Integration of digital camera into a mobile phone was an important step of the convergence between telecommunication and information technology as users started to require transfer of digital pictures besides conventional voice and text information. In addition, fast progress in digital technology has been an immense driving force of the needs for high data rates in telecommunications. Digital multimedia contents e.g. pictures, music, video clips are expected to be available anytime and anywhere which results into tremendous requirements in research and development in wireless technology.

Even though the industry tends to be majorly driven by software applications as well as “look and feel” of mobile devices, enabling hardware technologies in the background also deserve appropriate attention from R&D engineers. As soon as the performance of mobile communication systems cannot fulfil the expectation of users in terms of data rate and error robustness, the importance of the enabling hardware technology becomes obvious.

In order to cope with the rapid growth of the needs in wireless data transmission with constantly increasing data rates, new technical challenges arise perpetually on every layers of the OSI reference model. Whereas new modulation and multiple access techniques e.g. OFDM and OFDMA are introduced to support higher data rates and intelligent network configuration deals with the optimization of routing to increase the capacity and to improve load distribution, progress in hardware components in mobile devices and mobile base stations on the physical layer is also required to serve the needs of the higher OSI layers. Such progress on the physical layer includes techniques and hardware architectures which can enhance power efficiency of the system components while still complying with other specifications regarding linearity, noise, interference, etc.. Also, novel semiconductor device technology provides improved power handling capability resulting in smaller hardware size and high impedance which simplifies the design of matching networks. Moreover, large bandwidth and high impedance offer the possibility to create multiband components by designing the matching networks to be reconfigurable (Fischer, 2004).

This chapter aims to review state-of-the-art research in power amplifiers for wireless communication infrastructure featuring advantages of Gallium Nitride (GaN)-based power

devices including large bandwidth capability, high power density and high output impedance. Regarding the issues of power amplifier design, state-of-the-art power amplifier architectures will be discussed with various prospects. For wireless communication standards with high data rates e.g. WCDMA, WiMAX and LTE, their modulation schemes and multiple access techniques lead to non-constant signal envelope with high peak to average power ratio. As a consequence, power amplifiers in wireless communication infrastructure are required to operate in a wide dynamic range making it difficult to maintain high average efficiency over time. This chapter will discuss widespread techniques for average efficiency enhancement including Doherty power amplifier concept and envelope tracking (ET) with state-of-the-art results. Another possibility for power efficiency improvement is the switched-mode power amplifier where the waveforms of the voltage and current are optimized to achieve low power dissipation at the power transistor. GaN-based power transistors have demonstrated in numerous research works to be suitable power devices for the switched-mode architecture as well as for average efficiency enhancement techniques e.g. Doherty power amplifier and envelope tracking. As examples, results of 2.45 GHz GaN class AB power amplifier and GaN VHF class E power amplifier will be presented in this chapter. The wide band capability of GaN-based devices also supports design of reconfigurable and wideband power amplifiers. With all advantages of GaN-based devices, they are still not a mature technology in terms of reliability and memory effects. Results from investigation on memory effects and parasitics of GaN-based devices will also be discussed in the chapter showing promising improvements in these regards which make GaN-based devices interesting and promising power devices for future wireless communication infrastructure.

2. Power amplifiers in the wireless communication infrastructure

In a mobile communication system, power amplifier is an important component which boosts the transmitted signal power before it is sent via the antenna to the receiving device through wireless channels (see Fig. 1.). In a base station for mobile communication standards e.g. GSM, UMTS or LTE, power amplifier is the part which consumes the largest portion of power. Thus, the efficiency of power amplifier has the greatest influence on the entire system's efficiency. In addition, cooling requirement of a base station is also dominated by its power amplifier. In terms of cost, power amplifier is also the most expensive part of a base station. For the first generation of UMTS base stations, the costs of power amplifier and cooling are about 30%-35% of the cost of an entire base station (Chalermwisutkul, 2007). Besides the efficiency, linearity is also an important specification of power amplifiers which ensures that the transmitted signal is not distorted by the nonlinearity to an unacceptable level causing excessive bit errors.

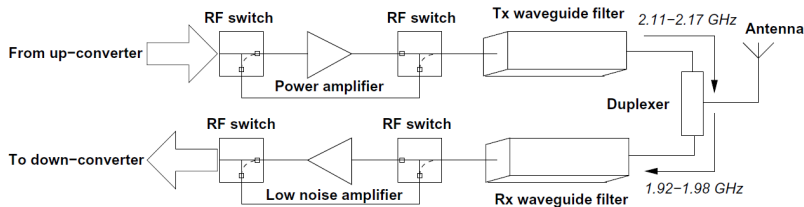


Fig. 1. Block diagram of a UMTS base station transceiver showing power amplifier and other system components.

2.1 Typical architecture and power device for base station power amplifier

In general, power amplifiers in mobile base stations are class AB amplifiers which offer both acceptable power efficiency and linearity. The operating point for the power device of this amplifier class is a compromise between those of highly efficient class B and highly linear class A. The conduction angles, output drain current waveforms, active load-lines and operating points of class A, AB, B, C, E and F amplifiers are depicted below in Fig. 2..

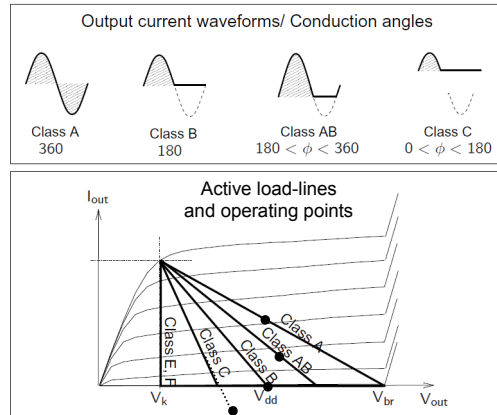


Fig. 2. Conduction angles, output drain current waveforms, active loadlines and operating points of class A, AB, B, C, E and F amplifiers. V_{out} , I_{out} , V_k , V_{dd} and V_{br} are drain output voltage, drain current, knee voltage, drain voltage supply and drain breakdown voltage, respectively (source (Chalermwisutkul, 2007)).

Typically, lateral diffused metal oxide semiconductor (LDMOS) field effect transistors based on Silicon are used as power devices for base station power amplifiers. Silicon LDMOS is considered a mature power device technology for mobile base station amplifiers due to its high efficiency, high power density and high thermal conductivity. However, main reasons which make LDMOS standard device technology for base station amplifiers are its low cost and high reliability. Although it is known that the operating frequency of LDMOS devices is limited to a few GHz, progress in LDMOS technology is still ongoing and new LDMOS devices are continuously introduced into the market with higher operating frequency and other progresses in terms of power efficiency, linearity, etc. (Ma et al, 2005). Due to this fact, the dominance of LDMOS devices in low GHz high power applications has been ensured since the first devices came into the market. However, new challenges in power device technology keep emerging as modern wireless communications are required to cope not only with higher data rates at limited frequency resource, but also with energy saving issues. In other words, there are increasing demands in high power efficiency besides spectrum efficiency for the wireless communication infrastructure. In this regard, there are several cases where it is worth to look for alternative power device to overcome limitation of existing device technologies.

Despite of all advantages of LDMOS, the main drawback of this device is the bandwidth capability. Due to high output capacitance of LDMOS device, the Q factor tends to be high and the bandwidth is small. Also, the operating frequency limit hinders this device from

being used in high frequency applications which are served with other device technology e.g. GaAs MESFET and HEMT. The research interest has been then attracted by wide-bandgap semiconductor materials for high frequency power devices. Silicon Carbide (SiC) is superior in thermal conductivity compared to other wide-bandgap semiconductors. However, the cost of SiC is relatively high. Moreover, this material is not appropriate for applications with very high operating frequencies. For Indium Phosphide (InP), another wide-bandgap compound semiconductor, the focus of research is on extremely high-speed digital applications where high power is not required.

The most prominent wide-bandgap semiconductor is Gallium Nitride (GaN). Comparing with Silicon device technology which is mainly driven by microprocessor and computer industries, GaN found its applications in screen industries enabled by GaN OLED (organic light emitting diode) technology and data storage industries utilizing blue laser produced by GaN laser diode to read out the data from a Blue-ray Disc™. In automotive applications and power electronics, GaN devices are attractive due to high operating temperature and high breakdown field for switching power supply. For RF power amplifiers, GaN-based power devices offer extremely large bandwidth, high power density, high operating frequency and high output impedance. The advantages of GaN-based power devices for wireless communications will be discussed more thoroughly in the next section.

2.2 Techniques for enhancement of average power efficiency

Modulation schemes and multiple access techniques allowing high data rates in wireless communication standards lead to non-constant signal envelope with high crest factor or peak to average power ratio (PAPR). Since a typical class AB power amplifier in mobile base station offers highest power added efficiency (PAE) about at one dB compression area in the power sweep plot, high peak to average power ratio leads to power back-off from the peak efficiency point which leads to efficiency reduction (see Fig. 3.). As a result, average efficiency over time is much lower than the peak efficiency. From the system point of view, reduction of peak to average power ratio can be done with different PAE techniques at the cost of

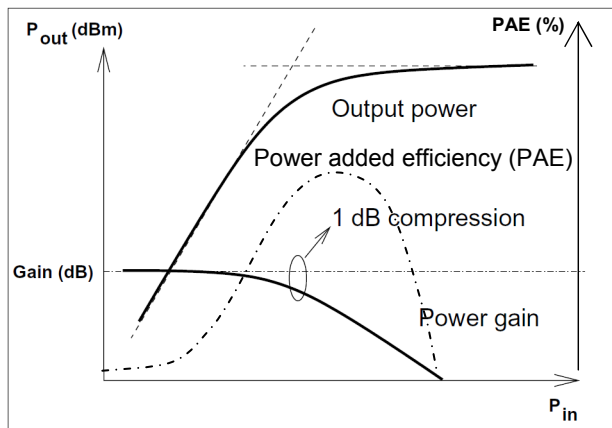


Fig. 3. Typical power sweep plot of a class AB power amplifier showing efficiency degradation when the power is backed-off from 1 dB compression point.

reduced data rate, transmit signal power increase, BER performance degradation, computational complexity increase, and so on (Jiang and Wu, 2008). Independent of the reduction techniques, rest of the peak to average power ratio still exists, so that for further average efficiency improvement, power amplifier architecture which can keep power efficiency high also when the transmitted power is backed-off must be considered.

Envelope elimination and restoration (EER) or Kahn technique

This average efficiency enhancement technique is based on the idea to separate the amplitude modulated envelope from the constant envelope, phase modulated carrier signal. The envelope is amplified with high efficiency envelope amplifier, whereas the carrier is amplified with nonlinear but highly efficient power amplifier. The output of the envelope amplifier is supplied to the carrier amplifier which reconstructs the typical signal with non-constant envelope of modern wireless communication standards (Diet et al, 2004).

Envelope Tracking (ET)

Similar to Kahn technique, supply voltage level of the RF amplifier is dynamically modified depending on the level of the signal envelope. A slight difference is that the input of the RF amplifier is still amplitude and phase modulated. Only with excessive signal power, the supply voltage of the RF amplifier is modified. The RF amplifier of this technique operates also in a linear mode unlike the Kahn technique, where the RF amplifier operates solely in a nonlinear mode.

Outphasing or Chirix technique

Also known as linear amplification using nonlinear components (abbr. LINC), this technique uses two nonlinear high efficiency power amplifier to boost up two signals with differently controllable phases. The two amplified signals are then combined with vector addition and the phase difference between the two signals defines the power level of the resulting signal. Compared to EER and ET, phase is the dynamically changing quantity and not the supply voltage of the RF amplifier (Helaoui et al, 2007).

Doherty technique

The concept of Doherty power amplifier utilizes two power devices which are operated as main and auxiliary amplifiers. As soon as a certain level of input power is reached, main amplifier—normally class B — is running into saturation providing its maximum efficiency. As the main amplifier starts to saturate, the auxiliary amplifier starts to conduct current. The saturation condition of the main amplifier is maintained by load modulation caused by the current from the auxiliary amplifier, so that the main power device acts like a voltage source. At peak output power, the auxiliary amplifier just begins to saturate and high efficiency is ensured for both amplifiers. Block diagram of a Doherty power amplifier is depicted in Fig. 4.. Details about Doherty amplifier can be found in the literature (Raab, 1987).

Compared to other efficiency enhancement techniques, Doherty concept has gained its popularity due to simple architecture which deals with RF circuit design issues only, whereas other techniques make use of digital signal processing to improve average efficiency. Thus, it is more straightforward to design a Doherty power amplifier to cope

with new peak to average power ratio value where high average efficiency is desirable. This can be simply achieved by modifying the input power division ratio between main and auxiliary amplifier. If necessary, three amplifiers can also be used in order to maintain high average efficiency over a high dynamic range.

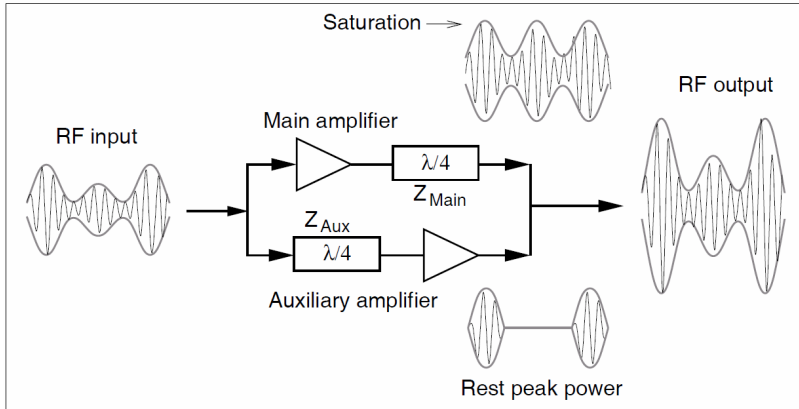


Fig. 4. Block diagram of a Doherty amplifier.

2.3 Switched-mode power amplifiers

In subsection 2.2, average efficiency enhancement techniques with the goal of maintaining high efficiency over a wide range of input power have been described. Considering peak efficiency at peak output power, switched-mode power amplifiers can achieve higher efficiency than widespread class AB power amplifiers. In case of switched-mode, the power transistor operates as a switch so that output voltage and current of the device (drain of FETs and HEMTs or collector for BJTs and HBTs) do not have high values at the same time. For the “off” state, the current is near to zero and the voltage is high and vice versa for the “on” state resulting in theoretical efficiency of 100%. In the following, switched-mode class E, F and D will be briefly described.

Class E

The first class E amplifier has been proposed by Sokal in 1975 (Sokal, 1975). Thereafter, other variations of class E amplifiers have been constantly presented with higher operating frequency where not only class E operation is ensured, but also, practical issues such as small circuit size and simple matching have been taken into account. A good example of such progress in class E amplifier design was represented by the class E amplifier with parallel circuit proposed by Grebennikov (Grebennikov, 2002). Class E offers high efficiency by avoiding simultaneous existence of high drain voltage and high drain current and thus, avoiding power dissipation of the power transistor. Control of the output current and voltage waveforms at drain or collector node of the device is achieved using an output load network. Theoretically, as the transistor turns on, the voltage drops to zero and the current starts to flow so that the output capacitance is gradually charged. As soon as the control voltage of the switch is lower than the switching voltage threshold,

the transistor is turned off and the current drops to zero while the output voltage of the device starts to increase. The ideal class E voltage and current waveforms are depicted in Fig. 5. Variations of class E amplifiers are reported to offer high power as 1 kW for switching applications at low frequency, whereas for RF applications, operating frequency of 10 GHz was already presented (Weiss, 1999). Class E is a promising switched-mode amplifier concept due to its simple architecture and flexibility compared to other switched-mode classes. Combination of a class E amplifier with average efficiency enhancement techniques e.g. EER or Doherty has been reported in the literature (Diet et al, 2004 and Kim et al 2010).

Class F

High efficiency of class F amplifiers is achieved by shaping the wave forms of output current and voltage of the power transistor which operates as a switch. Compared to class E, where load network is required to ensure the ideal switching condition (on state with high current, zero voltage and off state with high voltage and zero current), load network of class F has additional function which attempts to shape the output voltage and current waveforms at the device's drain or collector node. For conventional class F, odd harmonic peaking of the device's output voltage is realized by providing high impedance (open circuit condition) at the odd harmonic frequencies. As a result, the voltage waveform approximates a square wave. For the drain current, even harmonics are provided in addition to the fundamental by offering the device a short circuit condition at even harmonic frequencies. As a result, the current waveform approximates a half wave signal. Ideal current and voltage waveforms of a class F amplifier are shown in Fig. 5. Another alternative variation of class F is the inverse class F where the current waveform approximates a square wave, whereas the voltage waveform approximates a half wave signal. Efficiency of class F amplifiers can be increased by offering appropriate termination (open or short) at higher harmonics. However, this occurs at the cost of circuit's complexity. Similar to class E, class F and inverse class F amplifiers can be combined with Doherty technique to obtain high average efficiency for wireless communication signals with high peak to average power ratio. By using class F or inverse class F in a Doherty transmitter, peak efficiency is increased compared to the variation with class B main amplifier (Goto et al, 2004).

Class D

Unlike other switched-mode amplifier classes, class D uses at least two transistors as switches. In case of current mode class D (CMCD), the transistor's output current has a form of a square wave whereas the voltage mode class D (VMCD) shows a square output voltage of the transistor (see Fig. 5.). For both CMCD and VMCD, a tank filter is required to obtain the sinusoidal signal at the load. For CMCD, additional BALUN is also required, whereas for VMCD, two supply voltage sources are needed (see Fig. 6.). When one of the switches is turned on, the other one is turned off, so that high current and high voltage cannot exist at the same time. Theoretically, 100% efficiency can be achieved. In practice, the efficiency is compromised by limited switching speed and device's output capacitance. Due to these reasons, frequency of operation is limited for class D amplifiers. Experimental, state-of-the-art RF class D power amplifiers can operate at frequencies in the region near to 1 GHz (Aflaki et al, 2010).

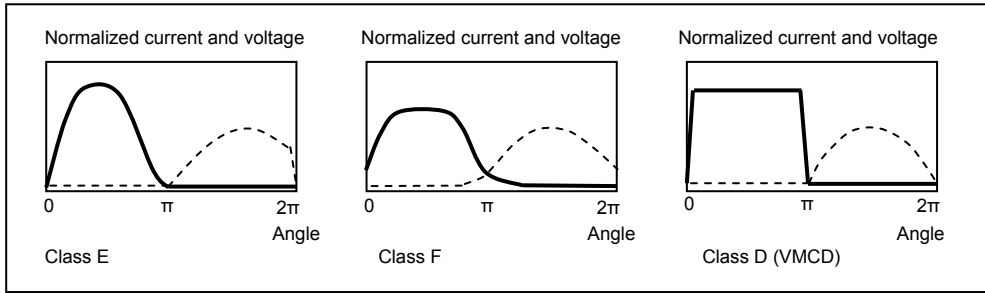


Fig. 5. Ideal current and voltage waveforms of class E, F and D switched-mode amplifiers (Raab et al., 2002). Broken lines represent the device's output current and the solid lines represent the device's output voltage.

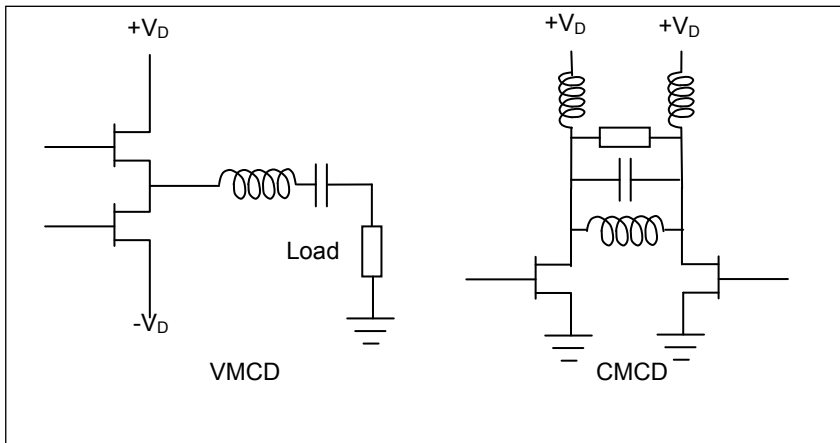


Fig. 6. Configurations of voltage mode class D (VMCD) and current mode class D (CMCD) amplifiers.

2.4 Linearization techniques

In general, a trade-off exists between efficiency and linearity of power amplifiers. For conventional transconductance amplifier classes e.g. class A, AB and B, it is obvious that high efficiency classes are nonlinear. In subsection 2.2, average efficiency enhancement techniques aiming to keep the efficiency high over a wide dynamic range have been discussed. Even though efficiency is the main goal of such techniques, linearity was also taken into consideration so that none of such techniques would have severe impact to linearity. However, when the desired efficiency profile is achieved, linearity might not comply with wireless communication standards leading to unacceptable error vector magnitude and bit error rates. In such a case, linearity improvement techniques can be utilized to eliminate the excessive nonlinearity of the amplifier. Widespread linearization techniques are reviewed below.

Feedback linearization

In order to force the RF output to follow the input, feedback of the RF signal is realized using a directional coupler. The simplest variation of this technique subtracts the RF feedback from the input signal. However, the compensation of the non linearity with this technique is not very efficient as the transmitter's gain is reduced. Another variation detects the envelope of the RF feedback and the input signal and subtracts the first from the latter to realize the linearization of the amplitude. For the compensation for both phase and amplitude nonlinearities, another variation called Cartesian feedback was conceived that the feedback signal is down converted to I and Q values which are used to compensate the I and Q of the input signal. For a relatively small bandwidth, the two tone IMD can be reduced by 10 to 35 dB with this technique.

Feedforward linearization

This linearization technique is excellent in terms of bandwidth and IMD reduction. In order to generate the error signal, the power amplifier's output and the input signal are sampled using directional couplers and the first is then subtracted from the latter (see Fig. 7.). The error signal is then amplified and subtracted from the power amplifier's output to obtain the linear output signal. Since this technique utilizes an open loop concept, additional loop control is required in order to compensate the degradation of the power device over time to ensure the right settings of phase shift and gain for maximum linearity. IMD reduction of 20-40 dB can be achieved for bandwidth up to 100 MHz. The drawback of this technique is the complexity of the system.

Digital predistortion

In order to obtain undistorted signal at the transmitter output, the input signal can be intentionally distorted before being fed to a nonlinear power amplifier. The predistorter generates nonlinearities which operate in the opposite way to the nonlinearities generated by the power amplifier, so that the overall response at the PA-output is linear (see Fig. 8). The linearization is done in the digital regime using FPGA which makes the system very flexible and adaptive for changes in power device over time to ensure linear output. As computational power of FPGA is continuously increasing, linearization over larger bandwidth can be realized with this technique. In the literature, linearization with digital predistortion technique which can cope with dynamic nonlinearity caused by electrical memory effects has also been reported (Lee et al, 2009).

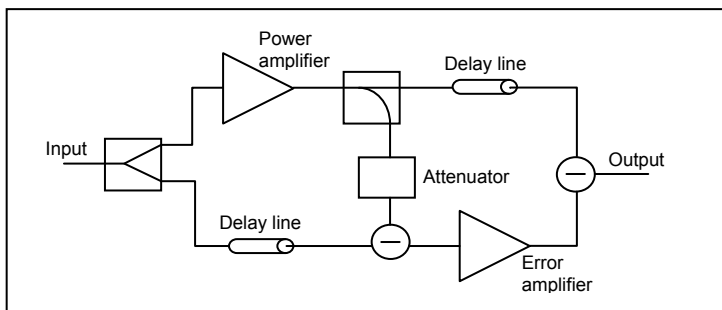


Fig. 7. Block diagram of a feedforward transmitter.

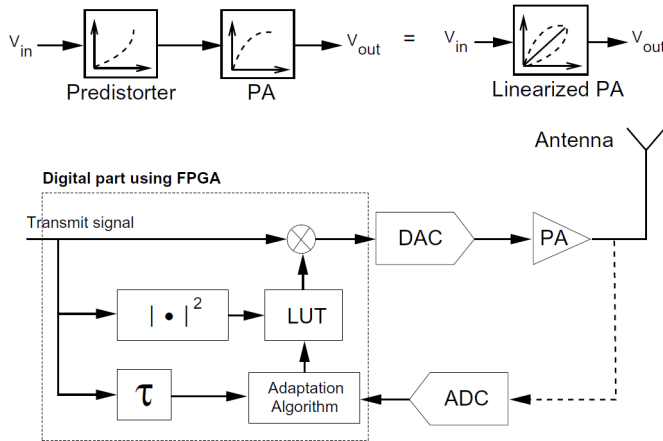


Fig. 8. Principle and block diagram of digital predistortion for linearization of power amplifiers.

In comparison to other techniques, digital predistortion offers higher efficiency and greater flexibility at low cost and represents a mature linearization technique for mobile base stations. Due to the mentioned flexibility and simple architecture, digital predistortion has gained its popularity in the power amplifier design community. In most of the cases where no extremely large bandwidth is required, high efficiency amplifiers e.g. Doherty and switched-mode amplifiers are combined with digital predistortion to improve the linearity.

3. GaN-based power amplifiers

As mentioned in section 2.1, GaN is a promising semiconductor material for high power and high frequency power transistors which are used as power devices in mobile base station power amplifiers. The advantages of GaN originate from physical properties of this wide-bandgap semiconductor. Table 1 shows physical properties of various semiconductor materials including GaN.

Material/Properties	Si	GaAs	InP	SiC	GaN
Bandgap (eV)	1.1	1.4	1.3	3.2	3.4
Saturation Velocity (*10 ⁷ cm/s)	1.0	2.1	2.3	2.0	2.7
Thermal Conductivity (W/cmK)	1.3	0.46	0.7	4.9	1.7
Breakdown Field (*10 ⁶ V/cm)	0.3	0.4	0.7	2.0	2.7
Electron Mobility (cm ² /Vs)	1350	8500	5400	800	1500

Table 1. Physical properties of semiconductor materials for RF and microwave applications.

From table 1, advantages of GaN compared to other semiconductor materials for RF and microwave applications are obvious. GaN offers very high saturation velocity leading to high operating frequency up to 100 GHz or higher. High breakdown field allows GaN-based devices to operate with high supply voltage which is advantageous for the off state of switched mode amplifiers and for obtaining high output power with high output impedance. Due to higher supply voltage, efficiency is also improved due to the reduction of the need for voltage conversion. For extreme operating environment e.g. for automotive applications, GaN offers wide bandgap and high thermal conductivity leading to the capability to operate at high temperature.

The most prominent GaN-based device for RF and microwave applications is GaN-based high electron mobility transistor (GaN HEMT). This kind of device offers extremely high operating frequency due to high electron mobility in the so-called 2DEG channel (Smorchkova, 2001). Moreover, one of the most impressive features of this device is the extremely high power density meaning that the device's size can be much smaller compared to other device technology for the same output power. With size reduction, output impedance becomes larger and parasitic capacitances smaller leading to large bandwidth and uncomplicated matching to 50 Ohm. It was also mentioned in the literature that GaN HEMT can offer better noise performance than that of MESFET's (Mishra et al, 2007).

For wireless communication infrastructure, GaN HEMT has proven itself to be an attractive alternative power device besides LDMOS FET for base station power amplifiers. For WCDMA base station, a GaN HEMT-based transmitter with output power higher than 200 W and supply voltage of 50 V was published in 2004 (Kikkawa et al, 2004). Reliability--one of the biggest concerns regarding GaN HEMT compared to LDMOS--was also presented in that work. However, at this point, it is not possible to foresee when GaN HEMT's will take the place of LDMOS FET's in base station power amplifiers. Even if the frequency of operation is limited to a few GHz for LDMOS, this device technology is continuously developed regarding power, reliability, linearity, etc.. Moreover, LDMOS is considered a cost-effective and mature power device technology with a large LDMOS amplifier designer community. Consequently, knowhow and design experience for this device is available to a great extent. Regarding this consideration, GaN HEMT will find its importance first in applications where large bandwidth is required or high power is desirable at high frequency. Besides reliability, charge carrier trapping in GaN HEMT has been a big issue for device technology improvement. Numerous investigations have been done regarding trapping effects of GaN devices. Charge carrier traps can cause dependency of the pulse-measured I-V characteristic on the quiescent point. This is a phenomenon of the so-called electrical memory effect (Chalermwisutkul, 2008). Other phenomena of memory effects are gate lag and drain lag in time domain where the drain current reaches its final value after some delay as the bias voltages are abruptly changed. In frequency domain, dispersion of output impedance is the consequence of electrical memory effect leading to dynamic nonlinearity with a large bandwidth of spectral regrowth (Fischer, 2004). Improvement of GaN device technology regarding charge carrier trapping and reliability has been reported occasionally e.g. SiN passivation or use of the field plate for traps reduction (Mishra, 2007). In this section, results from the works regarding GaN device modeling and GaN power amplifier design in which the author has been involved will be presented.

3.1 GaN device modeling

Computer simulation of the performance belongs to a typical design flow of power amplifiers. As many as possible components in the amplifier circuit should be characterized and described by models in order to obtain accurate prediction of circuit's performance from the simulations. As the main component of a power amplifier, quality of power transistor model plays a significant role in the accuracy of circuit simulation. Especially for power amplifier design, nonlinearities of the device must also be described by the device model unlike for small signal amplifiers, where it is sufficient to have the device's S-parameter sets of a few bias points of interest.

Even for one device technology, it is not practical to create a universal model which can describe the device's behavior under all operating conditions. In order to describe more effects and dependencies of the device's behavior on dynamic thermal and electrical conditions, more and more model parameters and nonlinear equations are required. In that case, the model would become very complex and long simulation time is needed. Though computational resource can be increased, complex device models suffer from poor robustness, that the simulation would be often terminated without convergence and reasonable results. For switched-mode power amplifiers e.g. class E, F, inverse F or D, a concept of using switch model in combination with the "on" state resistance R_{on} and output capacitance C_{ds} instead of empirical transistor model exists (Negra et al, 2007). This simple model is capable of providing good trend of power and efficiency and of verifying switched-mode operating conditions. At this point, there exist some discussions regarding the accuracy of such switch model for switched-mode power amplifier applications. Especially for power devices with charge carrier trapping and thus, memory effects, the switch model is not able to describe such effects which can have influence in efficiency and output power of switched-mode amplifiers (Chalermwisutkul, 2008).

Electrical memory effects

Even when electrical memory effects of GaN HEMT are still not negligible compared to those of GaAs HEMT, but the benefit of high power density, high output impedance, high frequency, etc. of GaN HEMT can be used, when the device is accurately described including the memory effects by the device model. First of all, the extraction of model parameters should be done using multibias pulsed measurement data. In such a measurement process, the bias voltages of the transistor is pulsed starting from the so-called quiescent point to other bias points in the I-V characteristics and drain current I_{ds} as well as S-parameters of that bias point are measured. Pulsed measurement has a significant advantage which is the isothermal measurement condition. The measured I-V characteristic of a pulsed measurement does not contain the self-heating of the transistor at high V_{ds} and I_{ds} as seen in DC measurement which is more familiar to the realistic operating condition. Moreover, quiescent point of pulsed measurement can be chosen equal to the operating point of the amplifier class of interest in order to create a device model which corresponds to the behavior of the device under realistic operating condition. In particular, the quiescent point dependent device model is necessary for a power device with significant trapping effects (see Fig. 9.). Theoretically, the dependence on quiescent point could be included into the model making the device model a general purpose one. However, as described above, this would increase the complexity and decrease the robustness of the model. Promising results of high power GaN HEMT have been published in 2004 showing the progress in

GaN device technology in term of reduction of trapping effects where the DC measurement of I-V curves shows no significant difference in the level of drain current compared to a pulsed measurement with a quiescent point at high drain voltage region (Kikkawa et al, 2004). In such a case, the quiescent point dependence of the device model would not be so critical. For power transistor manufacturers, normally, only one device model is provided to the circuit designer. As a result, the model of a mature power device regarding trapping will offer more accurate results for arbitrary classes of amplifiers.

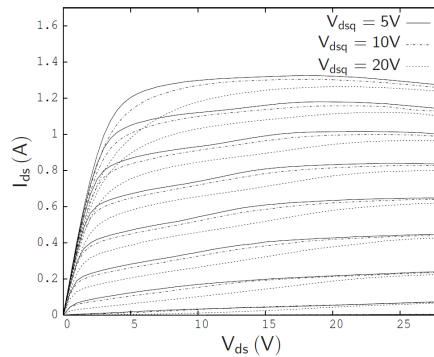


Fig. 9. Dependence of I-V characteristic of a GaN HEMT on quiescent point. The quiescent voltage was constant at a pinch-off value (no quiescent current) whereas the drain quiescent voltage V_{dsq} was varied.

Knee walkout

In contrast to GaAs HEMT and MESFET, the knee voltage of a GaN HEMT depends on the gate voltage and the drain quiescent voltage. With high gate voltage, the knee of the I-V curve becomes more round than at lower gate voltage where the knee is relatively angular. In addition, the knee voltage is shifted to the right toward higher drain voltage when gate voltage is high. This so-called *knee walkout* effect observed only with GaN HEMT and not with GaAs HEMT or MESFET cannot be modeled with standard EEHEMT model. By adding dependency of the knee voltage on the gate voltage and the drain quiescent voltage, the knee region of the I-V curve with high gate voltage can be better described (see Fig. 10.). As a result more accurate power and efficiency simulation can be done (see Fig. 11.) (Chalermwisutkul, 2007).

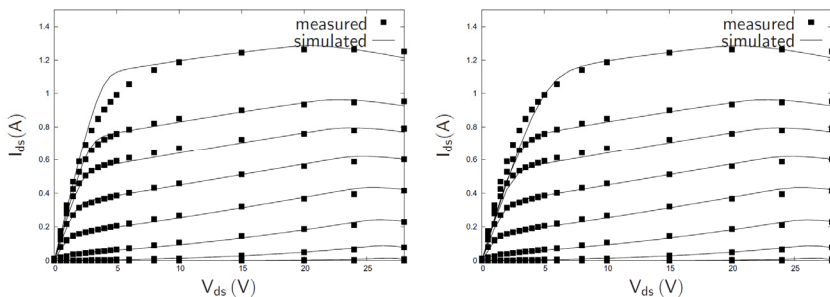


Fig. 10. I-V curves fitting results without (left) and with (right) the description of the knee-walkout.

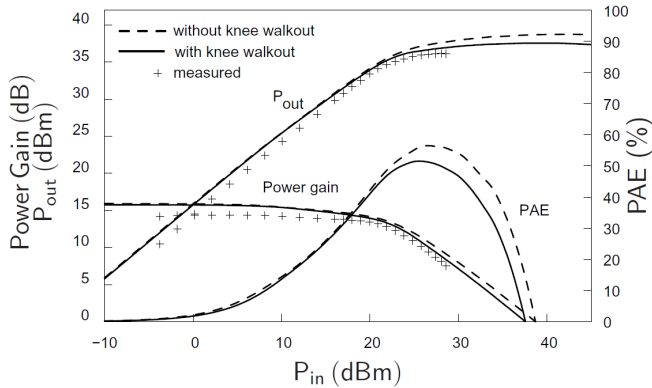


Fig. 11. Power sweep simulation without and with the description of the knee-walkout compared to measured values.

Large signal behavioral model

As discussed before, large signal model is required in order to describe nonlinearities of the power device. However, device modeling is a complex task which requires extensive experience of modeling engineers and special modeling software, so that power amplifier design engineers are mostly forced to rely on the large signal model provided by device's manufacturer. Due to progress in RF measurement techniques, a measurement system has been developed which allows measurement of the so-called X-parameters (Betts et al, 2011). Unlike with S-parameters, not only small signal behavior of the device can be described, but also nonlinearities arising under large signal conditions. In general, the input signal power is swept and the output response at the fundamental as well as at higher harmonics is measured. The measured information is then concluded into the X-parameter set which can be directly used in the circuit simulation software as the device's behavioral model. This kind of device modeling is very convenient and can be combined with source and load tuners to obtain load dependence of the X-parameters. In addition, the extracted behavioral model is accurate, robust and does not require large computational resource. However, the behavioral model cannot provide insights into physical properties of the device and the measurement setup is relatively expensive for small companies and educational institutions with low budget.

Package modeling

Packaged transistors comprise also parasitic components of the package and bond wires. These typical parasitic inductance and capacitance can compromise the performance of the amplifier circuit especially at high frequencies. For example, for class F amplifiers where short or open circuit must be provided at the drain node of the transistor at harmonic frequencies in order to shape the output current and voltage waveforms for high efficiency. Optimization for efficiency can be done best, if the package model of the transistor is known. The current and voltage waveforms which are optimized for minimum overlap should be presented at the internal drain node of the device inside the package and not at the external drain port (Schmelzer and Long, 2007).

Design examples of GaN HEMT power amplifiers

As examples, two GaN power amplifiers are presented. The first one is a 2.45 GHz GaN HEMT class AB power amplifier (Monprasert et al, 2009). This power amplifier is intended for the use in a WLAN system. The power transistor used in this amplifier is NPTS00004 GaN HEMT from Nitronex Corporation. The performance of the 2.45 GHz power amplifier is shown in Table 2. The drain supply voltage was varied with $V_{dsq}=20V$ and 28V. For the drain supply voltage of 28V, the output power is not as high as in the case with $V_{dsq}=20V$ since the drain current was increased as the device started to be saturated. The DC power exceeded the limit of 7 Watts given in the datasheet and the device was damaged. Fig. 12 shows a photograph of the fabricated class AB amplifier.

Drain quiescent voltage	$V_{dsq} = 20 V$	$V_{dsq} = 28 V$
Maximum output power	34.68 dBm	30.93 dBm
Maximum Power Added Efficiency	42.5 %	20.8%
Small signal gain	12.27 dB	13.69 dB

Table 2. Measured performance of 2.45 GHz GaN HEMT class AB power amplifier.

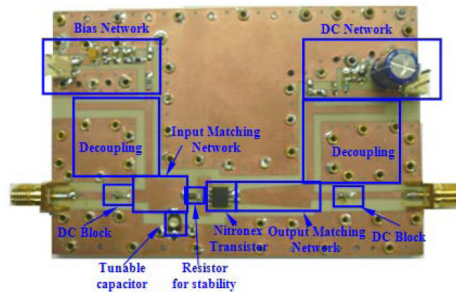


Fig. 12. Fabricated 2.45 GHz GaN HEMT class AB power amplifier.

Another design example is the VHF class E power amplifier (Khansalee et al, 2010). Using the same GaN power device Nitronex NPTB00004, a class E power amplifier for the operating frequency from 140 MHz to 170 MHz has been designed and fabricated. The values of load network L , C , L_0 and C_0 (see Fig. 13.) were determined using equations in the work published by Gebrennikov (Gebrennikov, 2002).

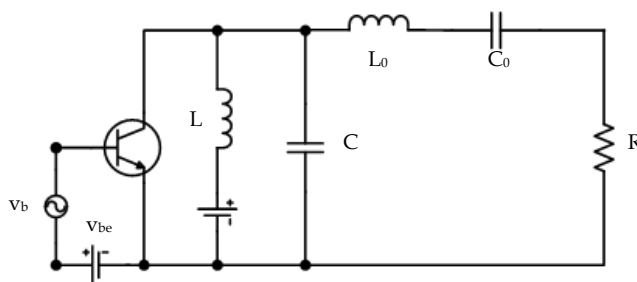


Fig. 13. Schematic of class E power amplifier with parallel circuit.

The optimal load impedance was determined using load pull simulation in Advance Design System (ADS). Simulated drain voltage and current waveforms show that class E operation is achieved (see Fig. 14).

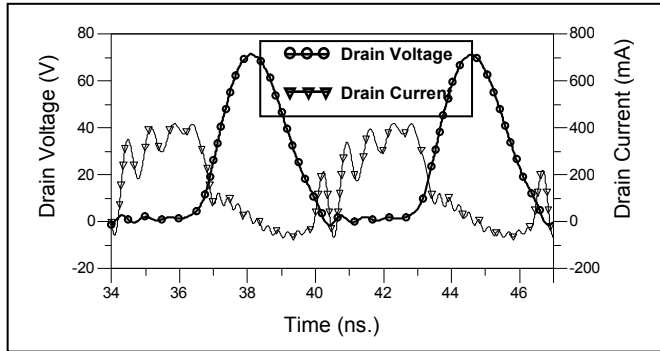


Fig. 14. Simulated drain current and voltage waveforms of the class E amplifier.

The fabricated class E power amplifier delivers maximum output 33.9 dBm, peak Power-Added Efficiency (PAE) of 72.5% and power gain of 16.4 dB at the center frequency of 155 MHz. Fig. 15. shows output power, efficiency and gain over the required operating frequency from 140 MHz to 170 MHz. A photograph of the fabricated GaN class E amplifier is depicted in Fig. 16.

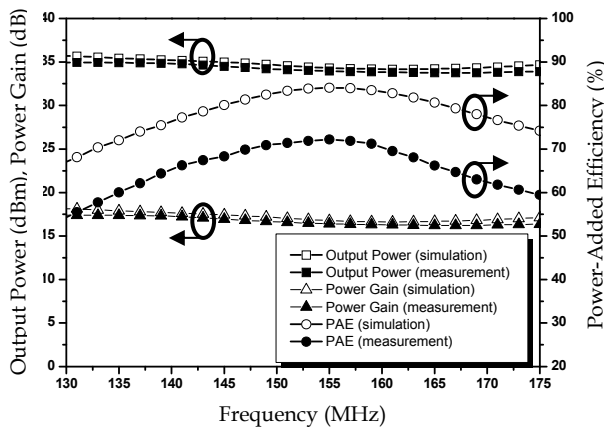


Fig. 15. Simulation and measurement results of power gain, output power, and PAE over the frequency 140 MHz to 170 MHz at input power of 18 dBm with the drain supply voltage of 24 V and gate supply voltage of -1.4 V over frequency.

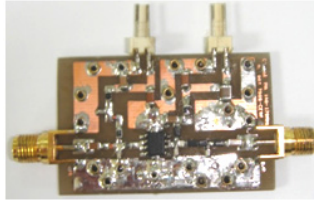


Fig. 16. Fabricated class E GaN VHF power amplifier.

4. Future research in power amplifiers for wireless communications

Needs for high data rates anywhere and anytime while the spectrum resource is limited will be a great challenge for future mobile and wireless communications. In order to utilize the bandwidth efficiently, new approaches on the network layers are being standardized and conceived including opportunistic, software defined and white space radio. The challenge of such frequency agile concepts will be not only be on the network and system layers e.g. spectrum sensing for vacant frequency slots, but also on the physical layer regarding the need of transmitters which can cope with extremely wide band or can be reconfigured for dynamic band migration. Regarding efficiency and power management, issues on every layer must be taken into consideration which would lead to interlayer optimization from network over system to physical layers. Active antenna and multiple inputs, multiple outputs (MIMO) concept will also be important topics which will require co-design and integration of amplifiers and antennas.

Energy saving is and will be a big issue not only in automotive and electrical power areas but also in wireless communications. To fulfil the intention for the “green transmission”, high efficiency must be provided by all infrastructure components e.g. base stations. Also, the trend of modern wireless communication standards is going in the direction of low power and small base stations will small cell size. This means that not only the mobile devices e.g. smart phone or tablets require aesthetic design but also the infrastructure components which should be well integrated into the environment. High efficiency will contribute to this requirement by offering small size of base stations. Regarding efficiency, research and development efforts will be spent in high efficiency signal transmission including design of switched-mode high efficiency power amplifiers with modulated input for improved efficiency e.g. class S amplifiers for delta sigma modulated signal (Pivit et al, 2008). Considering the demand of wide bandwidth and the capability to deliver high switching speed at high power, GaN-based devices are promising device technology for future wireless communications.

5. Conclusion

In this chapter, GaN-based power amplifiers for wireless communication infrastructure have been discussed. GaN HEMT's offer superior performance compared to state-of-the-art power devices for base station power amplifiers e.g. LDMOS. Especially high power density and high supply voltage of GaN HEMT's leads to smaller size of the device and thus, to lower parasitic capacitance, higher output impedance and large bandwidth which are advantageous for switched-mode and reconfigurable power amplifiers. In addition, wide

range of operating frequency can be covered by GaN-based power devices. The concerns of GaN transistors regarding charge carrier trapping and reliability is gradually extenuated by the progress in GaN device technology.

Device modelling is another important issue which ensures the power amplifier design community fast design process and accurate simulations. As examples, VHF class E amplifier and 2.45 GHz class AB amplifier have been presented.

6. Acknowledgment

The author would like to thank his family for the support and understanding during the preparation of the manuscript. Also, the author would like to express his appreciation to the research assistants, staffs and students of the RF and Microwave Laboratory, the Sirindhorn International Thai-German Graduate School of Engineering, King Mongkut's University of Technology North Bangkok for their interest in RF and microwave topics as well as for their support.

7. References

- Fischer, G. (2004). Architectural benefits of wide bandgap RF power transistors for frequency agile basestation systems, *Proceedings of the IEEE/MTT Wireless and Microwave Technology Conference*, Clearwater Beach, Florida, April 16, 2004.
- Chalermwisutkul, S. (2007). *Large Signal Modeling of GaN HEMTs for UMTS Base Station Power Amplifier Design Taking into Account Memory Effects*, PhD. Thesis, Faculty of Electrical Engineering and Information Technology, RWTH Aachen University, Germany, April 2007.
- Chalermwisutkul, S. (2008). Phenomena of Electrical Memory Effects on the Device Level and Their Relations, *Proceedings of the 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008*, pp. 229 - 232, ISBN 978-1-4244-2101-5, Krabi, Thailand, May 14-17 2008
- Ma, G.; Qiang Chen; Tornblad, O.; Tao Wei; Ahrens, C. and Gerlach, R. (2005). High Frequency Power LDMOS Technologies for Base Station Applications Status, Potential, and Benchmarking, *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pp. 361 - 364, ISBN 0-7803-9268-X, Washington DC, USA, 5 Dec. 2005
- Jiang, T. and Wu, Y. (2008). An Overview: Peak-to-Average Power Ratio Reduction Techniques for OFDM Signals, *IEEE Transactions on Broadcasting*, Vol. 54, No. 2, JUNE 2008, pp. 257 - 268, ISSN 0018-9316
- Diet, A.; Berland, C.; Villegas, M. and Baudoin, G. (2004). EER architecture specifications for OFDM transmitter using a class E amplifier, *Microwave and Wireless Components Letters, IEEE, Volume: 14 Issue:8, Aug. 2004*, pp. 389 - 391, ISSN 1531-1309
- Helaoui, M.; Boumaiza, S.; Ghannouchi, F. M.; Kouki, A. B. and Ghazel, A. (2007). A New Mode-Multiplexing LINC Architecture to Boost the Efficiency of WiMAX Up-Link Transmitters, *Transactions on Microwave Theory and Techniques, IEEE, Vol.: 55 Issue:2, Feb. 2007*, pp. 248 - 253, ISSN 0018-9480

- Raab, F.H. (1987). Efficiency of Doherty RF Power-Amplifier Systems, *IEEE Transactions on Broadcasting, Vol.: BC33 Issue:3, Feb. 2007*, pp. 77 – 83, ISSN 0018-9316
- Sokal, N. O. and Sokal, A. D. (1975). Class E-a new class of high efficiency tuned single-ended switching power amplifiers, *IEEE Journal of Solid-State Circuits, vol. SC-10, no. 3, June 1975*, pp. 168-176, ISSN 0018-9200
- Grebennikov, A. and Jaeger, H. (2002). Class E with parallel circuit – A new challenge for high-efficiency RF and microwave power amplifiers, *IEEE MTT-S Int. Micro. Symp. Dig., vol. 3*, pp. 1627-1630, June 2002, ISBN 0-7803-7239-5, Seattle, WA , USA, 02 Jun 2002 - 07 Jun 2002
- Kim, I.; Moon, J., Jee, S. and Kim, B. (2010). Optimized Design of a Highly Efficient Three-Stage Doherty PA Using Gate Adaptation, *IEEE Transactions on Microwave Theory and Techniques, Vol. 58, No. 10, October 2010*, pp. 2562 – 2574, ISSN 0018-9480
- Goto, S.; Kunii, T.; Inoue, A.; Izawa, K.; Ishikawa, T. and Matsuda, Y.; Efficiency enhancement of Doherty amplifier with combination of class-F and inverse class-F schemes for S-band base station application, *2004 IEEE MTT-S International Microwave Symposium Digest, Vol.2*, pp. 839 – 842, ISSN 0149-645X
- Aflaki, P.; Negra, R. and Ghannouchi, F. M. (2009). Enhanced Architecture for Microwave Current Mode Class-D Amplifiers Applied to the Design of an S-Band GaN-Based PA, *IET Microwave Antenna & Propagation, Vol.3, No. 6, pp.997-1006, Sep. 2009*, doi:10.1049/iet-map.2008.0282
- Raab, F.H.; Asbeck, P.; Cripps, S.; Kenington, P.B.; Popovic, Z.B.; Potheary, N.; Sevic, J.F.; Sokal, N.O. (2002). Power amplifiers and transmitters for RF and microwave, *IEEE Transactions on Microwave Theory and Techniques, Vol. 50, No. 3, March 2002*, pp. 814 - 826, ISSN 0018-9480
- Lee, M. W; Lee, Y. S.; Kam, S. H.; Jeong, Y. H. (2010), A wideband digital predistortion for highly linear and efficient GaN HEMT Doherty Power Amplifier, *Microwave and Optical Technology Letter, Volume 52, Issue 2, February 2010*, pp. 484-487, DOI: 10.1002/mop.24951
- Smorchkova, I. P.; Chen, L. ; Mates, T.; Shen, L.; Heikman, S.; Moran, B.; Keller, S.; DenBaars, S. P.; Speck, J. S. and Mishra, U. K. (2001). AlN/GaN and (Al,Ga)N/AlN/GaN two-dimensional electron gas structures grown by plasma-assisted molecular-beam epitaxy, *Journal of Applied Physics, vol. 90, no. 10*, pp. 5196-5201, Nov. 15, 2001. doi:10.1063/1.1412273
- Mishra, U.K.; Shen Likun; Kazior, T.E.; Yi-Feng Wu (2008). GaN-Based RF Power Devices and Amplifiers, *Proceedings of the IEEE, Volume: 96 Issue:2, Feb. 2008* pp. 287 – 305, ISSN 0018-9219
- Kikkawa, T.; Maniwa, T.; Hayashi, H.; Kanamura, M.; Yokokawa, S.; Nishi, M.; Adachi, N.; Yokoyama, M.; Tateno, Y.; Joshin, K. (2004). An Over 200-W Output Power GaN HEMT Push-Pull Amplifier with High Reliability, *2004 IEEE MTT-S International Microwave Symposium Digest, Vol.3*, pp. 1347 – 1350, 6-11 June 2004, ISSN 0149-645X
- Negra, R.; Chu, T.D.; Helaoui, M.; Boumaiza, S.; Hegazi, G.M.; Ghannouchi, K. (2007). Switch-based GaN HEMT model suitable for highly-efficient RF power amplifier design, *IEEE/MTT-S International Microwave Symposium, 2007*, pp. 795 – 798, 3-8 June 2007, Honolulu, HI, USA, ISSN 0149-645X

- David Schmelzer and Stephen I. Long (2007). A GaN HEMT Class F Amplifier at 2 GHz with > 80 % PAE, *Compound Semiconductor Integrated Circuit Symposium, 2006. CSIC 2006. IEEE*, pp. 96 - 99 San Antonio, TX, Nov. 2006, ISBN 1-4244-0126-7
- Loren Betts; Dylan T. Bepalko and Slim Boumaiza (2011). Application of Agilent's PNA-X Nonlinear Vector Network Analyzer and X-Parameters in Power Amplifier Design, Agilent Technologies White Paper, May 12, 2011
- Monprasert, G.; Suebsombut, P.; Pongthavornkamol, T. and Chalermwisutkul, S. (2009). 2.45 GHz GaN HEMT Class-AB RF power amplifier design for wireless communication systems, *Proceedings of the 2010 International Conference on Electrical Engineering/Electronics Computer Telecommunications and Information Technology (ECTI-CON)*, pp. 566 - 569, Chiangmai, Thailand, 19-21 May 2010, ISBN 978-1-4244-5606-2
- Khansalee, E.; Puangngernmak, N. and Chalermwisutkul, S. (2010). A high efficiency VHF GaN HEMT class E power amplifier for public and homeland security applications, *2010 Asia-Pacific Microwave Conference Proceedings (APMC)*, pp. 437 - 440 Yokohama, Japan, 7-10 Dec. 2010 ISBN 978-1-4244-7590-2
- Florian Piviti; Jan Hesselbarth; Georg Fischer and Suramate Chalermwisutkul (2008), Radio Frequency Transmitter, Pub. No.: WO/2009/062847 International Application No.: PCT/EP2008/064659, International Filing Date: 29.10.2008

Analysis of Platform Noise Effect on Performance of Wireless Communication Devices

Han-Nien Lin
Feng-Chia University
Taiwan, R.O.C.

1. Introduction

Cloud computation and always-connected Internet attracts the most industrial attention for the past few years. Meanwhile, with the development of IC technologies advancing toward higher operating frequencies and the trend of miniaturization on wireless communication products, the circuits and components are placed much closer inside the wireless communications devices than ever before. The system with highly integrated high-speed digital circuits and multi-radio modules are now facing the challenge from performance degradation by even more complicated platform EMI noisy environment. The EMI noises emitted by unintentionally radiated interference sources may severely impact the receiving performance of antenna, and thus result in the severe performance degradation of wireless communications. Due to the miniaturization of a variety of wireless communications products, the layout and trace routing of circuits and components become much denser than ever before. Therefore, we have investigated and analyzed the EMI noise characteristics of commonly embedded digital devices for further high performance wireless communications design. Since the camera and display module is most adopted to the popular mobile devices like cellular phone or Netbook, we hence focus on EMI analysis of the built-in modules by application of IEC 61967[1][2] series measurement method.

Since the causes of reduction of throughput or coverage due to receiving sensitivity degradation of wireless system could result from decreased S/N via conducted or radiated EMI noises from nearby digital components shown in Figure 1. This chapter discusses RF de-sensitivity analysis for components and devices on mobile products. To improve the TIS performance of wireless communication on notebook computer, we investigated the EMI noise from the built-in camera and display modules as examples and analysed the impact of various operation modes on performance with throughput measurement. We also utilized the near-field EM surface scanner to detect the EMI sources on notebook and locate the major noisy sources around antenna area. From the emission levels and locations of the noisy components, we can then figure out their impact on throughput and receiving sensitivity of wireless communications and develop the solutions to improve system performance. Finally, we designed and implemented periodic structures for isolation on the

notebook computer to effectively suppress noise source-antenna coupling and improve the receiving sensitivity of wireless communication system.

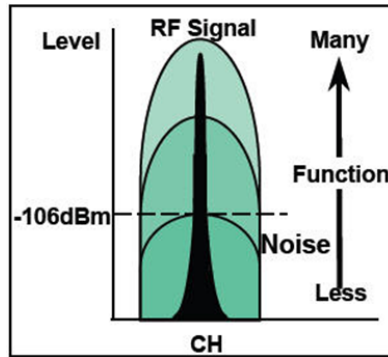


Fig. 1. S/N ratio decreases due to digital components for multi-functions.

2. The noise impact of camera and Touch Panel (TP) modules on product performance

2.1 Performance testing for Wireless Wide Area Network (WWAN) devices

There are two different purposes for the OTA (Over-The-Air) testing[3] on mobile stations. The first testing is for the carrier's cell site coverage which is relative with loss plan and link budget of the cell site. For example, the sensitivity measurement of the W-CDMA receiver is performed by the base-station simulator to determine the receiving sensitivity of EUT (Equipment Under Test) by reporting the minimum forward-link power which resulting in a bit-error-rate (BER) of 1.2% or less at the data rate of 12.2 kbps with a minimum of 20,000 bits. The second testing is the throughput for supporting all kind of the applications for cloud computing. The minimum throughput required will depend on application. For example, the minimum throughput we need to link YOU TUBE for HD video is about 1Mbps at least. Therefore, the mobile station (Smart Phone, Tablet PC, Note book PC, etc.) is required the OTA performance testing on TRP, TIS and De-Sense.

2.1.1 Total Radiated Power (TRP)

TRP measurement is to evaluate the transmitting RF power performance of mobile device by summing the effective isotropic radiated power (EIRP) of complete Theta- and Phi-cut as shown in Figure 2. The procedure is first to measure the radiated power at each Phi degree interval for 360 degree rotation (if interval is 30 degree then it need 12 measurement), and then for the Theta axial. Finishing the 180 degree rotation along Theta axial, the TRP is obtained with following formula.

$$TRP \cong \frac{\pi}{2NM} \sum_{i=1}^{N-1} \sum_{j=0}^{M-1} [EiRP_{\theta}(\theta_i, \phi_j) + EiRP_{\phi}(\theta_i, \phi_j)] \sin(\theta_i) \quad (1)$$

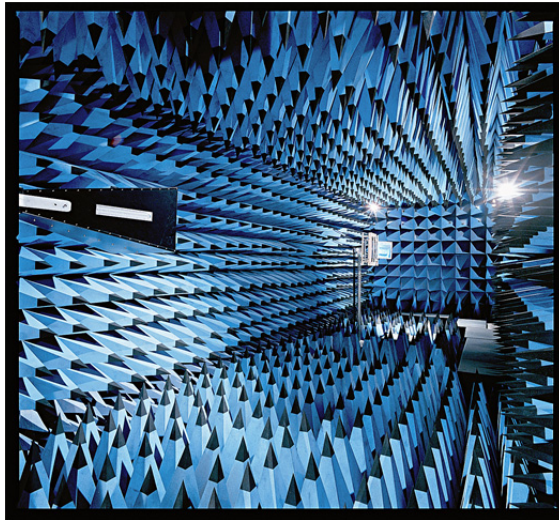
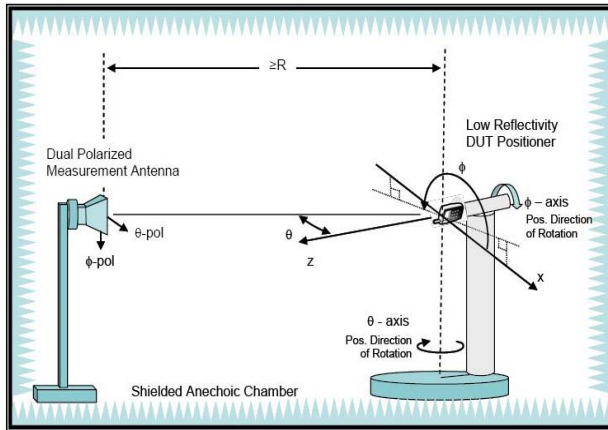


Fig. 2. Total Radiated Power measurement.

For the ideal case, the TRP should be equal to the conducted power (Watts) times mismatching Loss (%) and antenna efficiency as shown in following relationship and illustration. But the antenna efficiency measurement can actually with error resulting from coaxial cable connection as illustrated in Figure 3. When the coaxial cable is connected to the SMA connector, the surface on it could cause measurement error of the antenna efficiency.

$$TRP = \frac{1}{4\pi} \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} (EiRP_{\theta}(\theta, \phi) + EiRP_{\phi}(\theta, \phi)) \sin(\theta) d\theta d\phi$$

$$TRP = P_A \cdot L_m \cdot eff$$
(2)

Transmit Power = Pc (Conducted Power) + Antenna Gain (in dB)

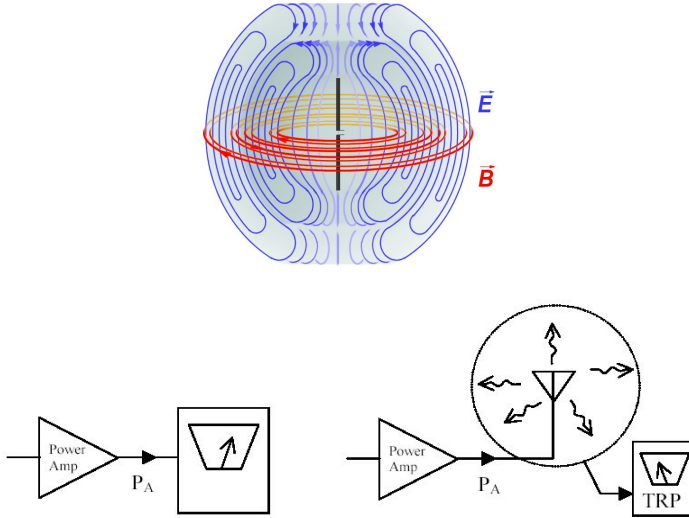


Fig. 3. Illustration of Antenna TRP.

2.1.2 Total Isotropic Sensitivity (TIS)

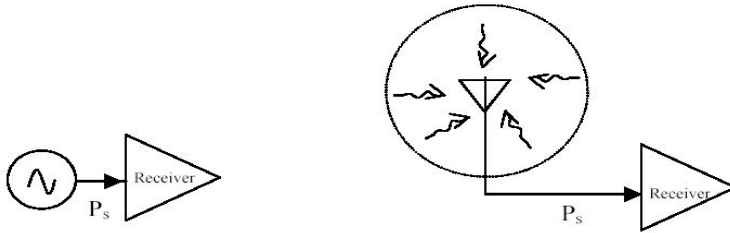
The measurement setup for TIS testing is the same as shown in Figure 2, except with the different calculation. It is analogous to calculate the total resistance form the parallel resistor network. The Effective Isotropic Sensitivity (EIS) is illustrated in Figure 4 and calculated with following formula.

$$TIS \cong \frac{2NM}{\pi \sum_{i=1}^{N-1} \sum_{j=0}^{M-1} \left[\frac{1}{EIS_{\theta}(\theta_i, \phi_j)} + \frac{1}{EIS_{\phi}(\theta_i, \phi_j)} \right] \sin(\theta_i)}$$
(3)

$$G_{x,EUT}(\theta, \phi) = \frac{P_s}{EIS_x(\theta, \phi)}$$
(4)

$$TIS = \frac{4\pi}{\oint \left[\frac{1}{EIS_{\theta}(\theta, \phi)} + \frac{1}{EIS_{\phi}(\theta, \phi)} \right] \sin(\theta) d\theta d\phi}$$
(5)

For the ideal case, TIS should be equal to conductive sensitivity divided by mismatching loss and antenna efficiency as shown in following relationship and illustration. Not only the surface current on coaxial cable would cause the antenna efficiency measurement error, but also the platform noise interference investigated here would de-sense the receiver.



EIS: Effective Isotropic Sensitivity = Received EIRP - Antenna Gain (in dB)
 TIS: Total Isotropic Sensitivity (3D Measurement)

Fig. 4. Illustration of Antenna TIS.

The relationship between receiver performance and platform noise is described by receiver bit error rate and energy per bit (E_b) /Noise(N_0) as shown in Figure 5. For example, the WCDMA receiver sensitivity with QPSK modulation can be determined as following:

Bit Error Rate (BER): 1.2% BER for QPSK demodulation receiver require that $E_b/N_0 = 7.5\text{dB}$ where E_b : measured at base-band output (I/Q output) for each bit.

N_0 : total noise power form RF front end to base-band, include LNA NF, ADC, quantize noise, PLL phase Noise,... with Gaussian system noise representation.

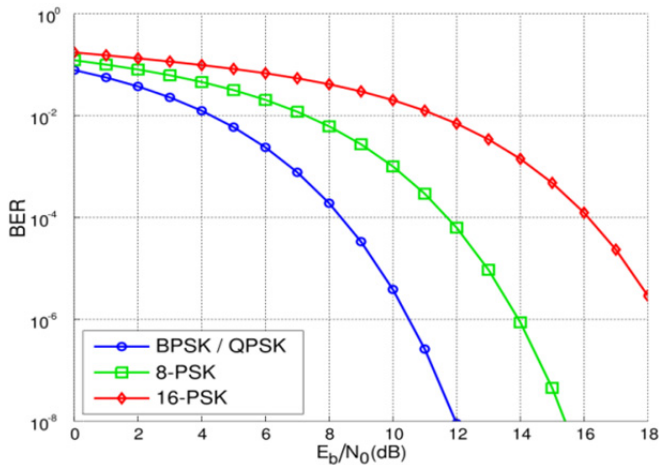


Fig. 5. Receiver Bit Error Rate vs. Energy per bit (E_b) /Noise(N_0).

The BER is measured in time domain after demodulation of receiver. For WCDMA system, it needs $20\text{k bits} / 12.2\text{Kbps} = 1.64$ second at each receiving channel. From communication demodulation theory, N_0 is described as Gaussian noise, and it is the sum of the receiver noise (related to implementation loss) and system noise as illustrated in the following figures.

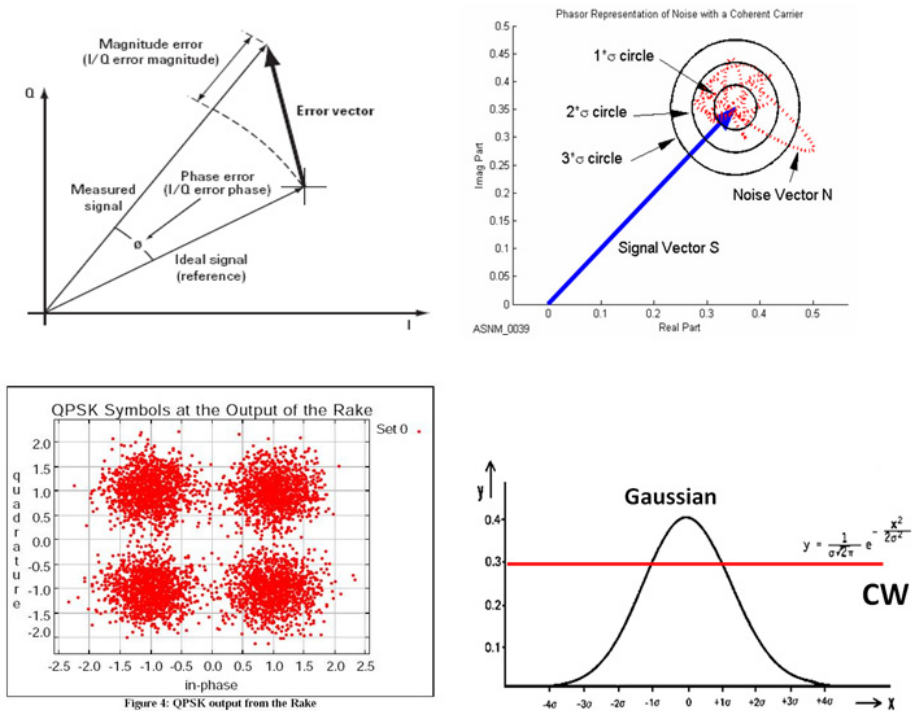


Fig. 6. Illustration of system noise effect.

Based on TIS requirement for WWAN or WLAN throughput, the noise limit of the wireless system should be set to meet the regulatory specification. Figure 7 shows the sensitivity degradation due to self-interference for GSM 1800 and WCDMA systems. The example for WCDMA system is following:

Noise Limit: TIS (dBm) + Antenna Gain (dB) - Eb/N0(dB) + Processing Gain (demodulation dependence) - System Losses (6dB) (depending on chip set, LNA NF, PLL phase noise, ADC.....), Processing Gain (dB)=10 Log(Chip rate/Data Rate) = 10log (3840K/12.2K)= 25dB

WCDMA Noise Limit : Gaussian Probability Density Function Noise Power-103dBm + (-5dB) -7.5dB +25dB - 6dB = 96.5dBm

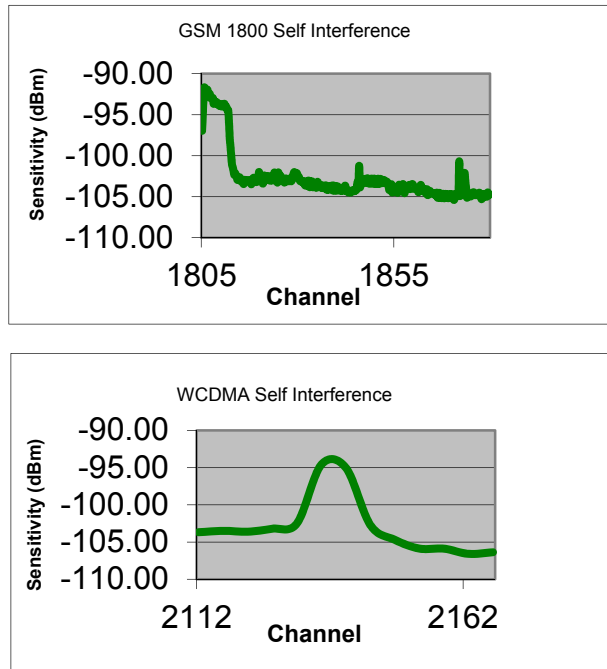


Fig. 7. Sensitivity degradation examples for GSM 1800 and WCDMA systems.

However, since there are more than one thousand Channels for GSM and WCDMA systems, we can't test all receiving channels for those WWAN devices. The alternative way for testing those intermediate channels is to measure the relative sensitivity as following steps and illustration as shown in Figure 8:

1. Move the EUT and position to the location and polarization which results in best radiated sensitivity, then measure for the closest channel (in frequency) and set as Reference Channel. After the 3D testing of the high, middle and lower channels (set as reference channel), we acn then review the 3D graph and find the best sensitivity at theta- and phi-plane of EIS along vertical or norizontal polarization (it means the best radiated sensitivity).
2. Now, there are three bset radiated sensitivity at Theta- and Phi-polarization for Low, Middle and High reference channel in one band. The rest of channels in one band should be then tested as following: The all channels in frequency range from lowest frequency channel (reference channel) to the frequency at $(\text{low} + \text{Middle})/2$ should be de-sensed less than 5 dB to the reference channel. The all channels in middle band of frequency from $(\text{Low} + \text{Middle})/2$ to $(\text{Middle} + \text{High})/2$ should be de-sensed less than 5dB to the reference channel (middle channel). Finally, the all channels in high band of frequency from $(\text{Middle} + \text{High}) / 2$ to (High should be de-sensed less than 5dB to the reference channel (highest frequency channel).

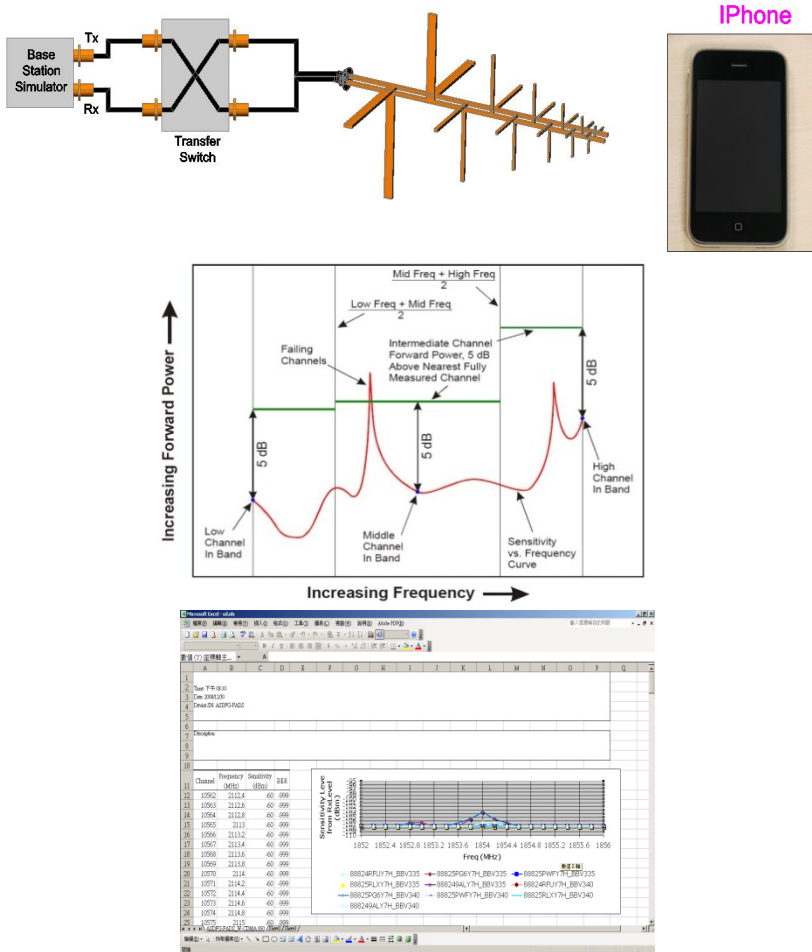


Fig. 8. Alternative way for testing relative sensitivity of intermediate channels.

2.2 De-Sense effect

DeSense is the term representing the noise impact degree on receiver sensitivity. This section will address on the popular TP and camera module about their roles on interference with embedded antenna and discuss the De-sense effect from platform noise coupling.

Nowadays, Touch Panel (TP) and camera module both occupy large part on the smartphone, Tablet PC or NB. Hence no matter where the embedded antennas are placed, the noise emitted from TP and Camera will couple to antenna and thus result in DeSense (Degradation of Sensitivity) problem. On the other hand, all the RF power transmitted by embedded antennas of the wireless products will also couple to nearby TP and camera modules. Those proximate electric (E) and magnetic (H) field coupling will affect normal operation of TP and camera modules.

Figure 9 shows the antenna locations under investigation. In addition, Figure 10 shows the Path-Loss measurement setup and test procedures as following[4]:

1. Put EUT into shielding box.
2. Connect VNA port1 to Tx antenna and laptop antenna to port 2
3. Measure for specific frequencies antenna efficiency of Tx antenna and Rx antenna

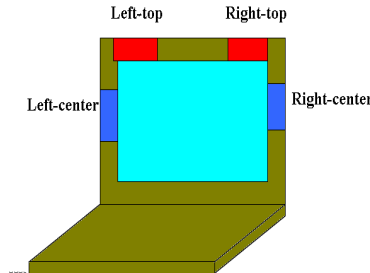


Fig. 9. Antenna locations on laptop display.

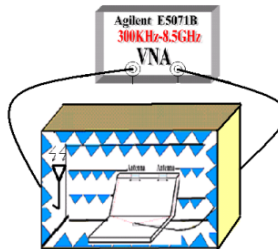


Fig. 10. Test setup to measure Path-Loss.

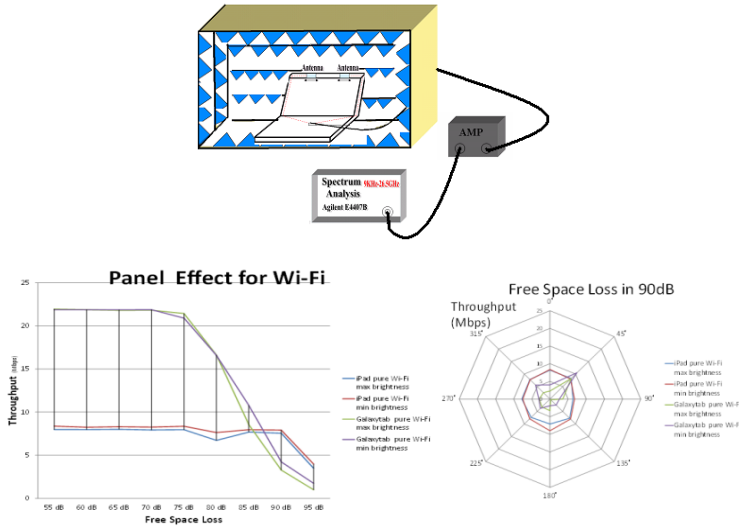
2.2.1 Impact of LCD EMI analysis on 802.11g throughput[5]

In a laptop, there are many interference sources which can be in the form of radiation or conduction. LCD noise is the major interference to the wireless performance. Figure 11 shows the frequency domain measurement setup and measured results for platform noise from LCD. The measurement setup is shown in Figure 11a and the test procedures are described as following:

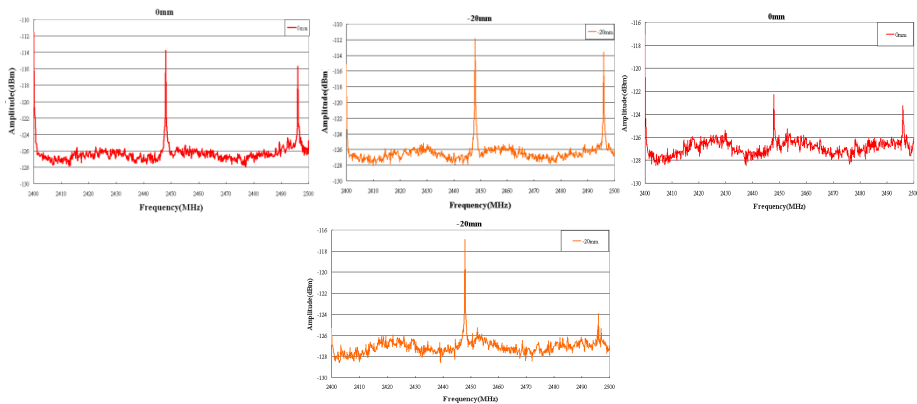
1. Put EUT into shielding box.
2. Connect antenna cable to AMP/Spectrum analyzer via coaxial cable.
3. Power on EUT.
4. Measure noise level for the chosen target frequency.

Figure 11b shows the different antenna placements along the horizontal edge on top of a LCD panel. The measured results show that the noises at 2.400GHz, 2.450GHz, 2.490GHz (harmonics of the pixel clock) are major interference sources that fall into WLAN band. We can obtain 2-5dB noise suppression by simply moving the antenna several millimeters away from its initial location. The comparison for different antenna measurements at different locations shows that the LCD noise might have an significant impact on desensitization to

802.11g. The measurement of antenna positioned towards the left 20mm serves as a reference to quantify the impact of antenna placement on the platform noise measurement. The noise picked up by antennas would desensitizes the receiver and reduced the throughput. Meanwhile, the throughput test procedure and test setup are as follows: the setup consists of an AP (access point), EUT (laptop) and Chariot console throughput software. The AP (access point) and EUT (laptop) are connected through path-loss attenuators to control RFI strength, and the communication traffic is controlled and monitored by a desktop using Ixia Chariot® software as shown in Figure 12a. The system



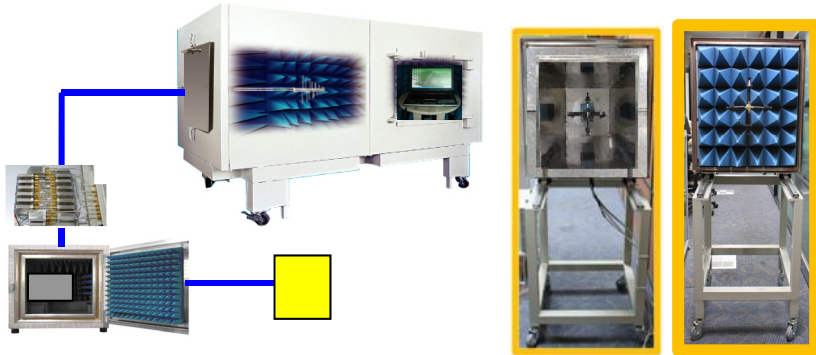
(a) Platform noise measurement setup for antenna port



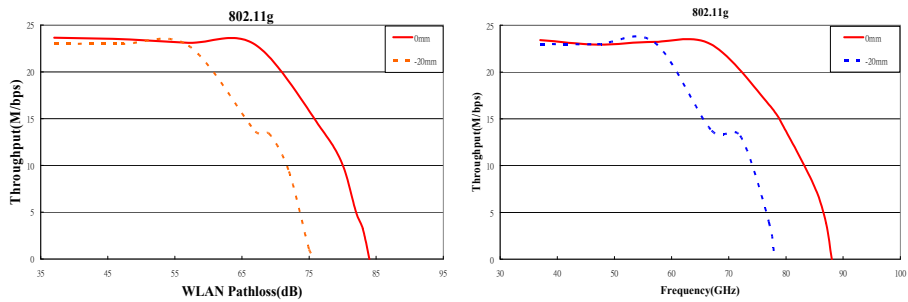
(b) Comparison of the different antennas measurement at different locations on the LCD panel.

Fig. 11. LCD noise measured at antenna port in WLAN band.

path-loss includes cable loss, space loss and attenuators. The results in Figure 12b the real line and dotted line, clearly show that the sensitivity and the throughput decrease as the LCD interference is injected to the communication link between the AP and NIC card. It is found that there is about 10dB desensitization between the two throughput measurements for different locations. It is show that when the LCD noise increases the sensitivity is reduced and performance is also degraded. Moreover, it is again shown that the location of antenna placement is significant to wireless communication performance. Figure 12c shows the photograph of the antenna integrated into a laptop for investigation.



(a) The throughput test procedure and test setup



(b) Throughput comparison for in 802.11g antenna 2 and antenna 3 at two different positions.



(c) Photograph of the antenna integrate into a laptop.

Fig. 12. Throughput comparison in 802.11g for different locations on the LCD panel.

2.2.2 Platform noise analysis of Netbook system[6]

Platform noise analysis can be conducted through the noise floor measurement at antenna port of wireless device. The complete setup for noise floor measurement should at least consist of following hardware instruments: Shielded box, pre-amplifier, spectrum analyser or EMI receiver, high quality coaxial cable, and Netbook EUT.

The frequency-domain noise floor measurement setup for Netbook platform is shown as Figure 11(a). The measuring procedure is as following: Put the EUT Netbook inside the shielded box and connect its antenna port to the pre-amplifier and spectrum analyser. We first measured the ambient noise with Netbook power-off, then powered on the Netbook and measured the noise level within the selected communications bands as interfering platform noise.

Figure 13 shows the noise level captured by the integral WWAN antenna of the Netbook on GSM 850, GSM 900, and DCS 1800 bands. The platform noises on GSM 850, GSM 900, and DCS 1800 bands are shown in Figure 2(a), (b), and (c) respectively. Figure 13(a) shows that the major noise spectrum falls on 864 MHz and 888 MHz, corresponding to the 36th and 37th harmonics of the Azalia Sound Card with 24 MHz fundamental driver frequency.

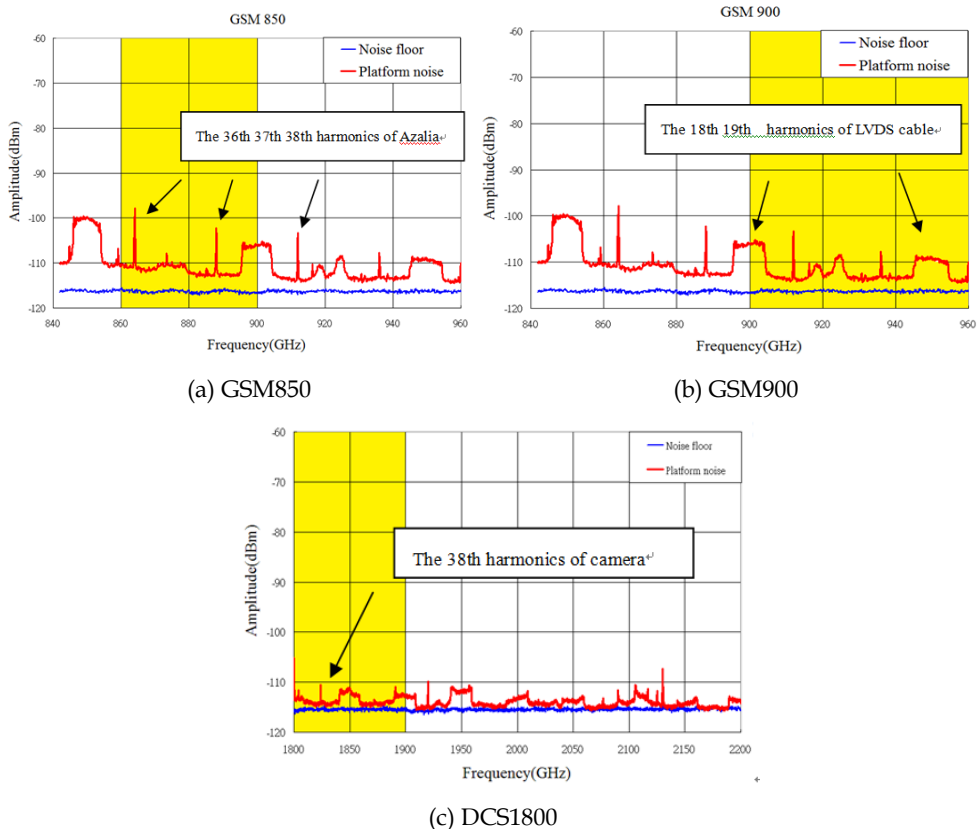


Fig. 13. Measured noise level on WWAN bands of Netbook.

Figure 13(b) shows broadband and regularity characteristics of noise on 900 MHz band. Since the fundamental frequency of the LVDS interconnection cable is 50 MHz and noise spectrum falls between 900 MHz and 950 MHz, we can calculate from system clock map that the noise spectrum received by antenna was originated from 18th and 19th harmonics of LVDS cable. Figure 13(c) shows that the noise measured on DCS 1800 band falls on 1824 MHz, which is 38th harmonic of CCD camera. Figure 13(c) also shows noise occurring around 1850 MHz and 1900 MHz, which are the 38th and 39th harmonics of LVDS cable respectively. From the noise spectrum analysis, we found out that the noise sources mentioned above cause in-band interference on operation bands of GSM 850/950 and DCS 1800 systems frequently. Figure 14(a) shows the noise measurement result on Band-1 of WCDMA, and it appears as lower level and ambiguous. However, because Band 5 of WCDMA almost operate on the same frequencies with GSM 850, it thus suffers interference from the 36th and 37th harmonics of the Azalia Sound Card.

From the platform noise measurement method and clock map analysis, we are able to establish the system design rule for related position and orientation of noise source(s) and antenna(s) placement to suppress in-band interference. We can also utilize various isolation or shielding techniques to effectively prevent the antenna port noise level from platform noise sources, and further reduce the delay time caused by lengthy product debug and speed up testing time.

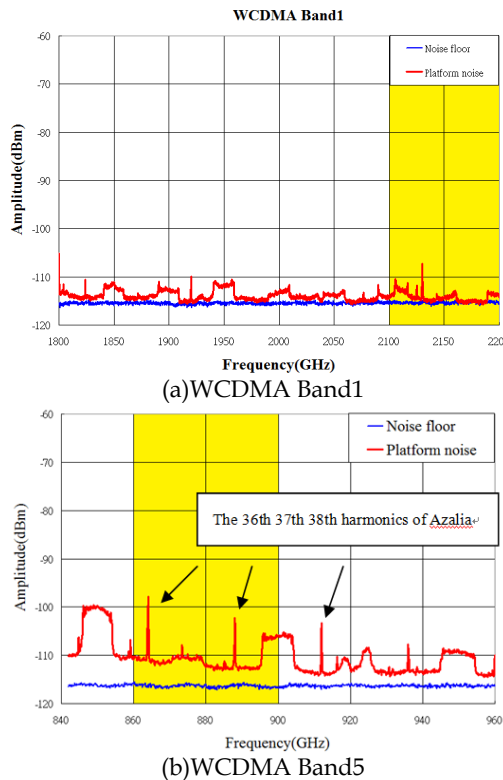


Fig. 14. Measured noise level on WCDMA bands of Netbook.

2.2.3 Impact analysis of platform noise on TIS of GSM/WCDMA systems[6]

TIS is a figure of merit for receiving performance of a mobile or wireless terminal device. Receiver performance is considered as important to over all system performance as is Transmitter performance. The downlink, or subscriber unit receive path is integral to the quality of the device's operation. The receiver performance of the Equipment Under Test (EUT) is measured utilizing Bit Error Rate (BER) or Frame Errasure Rate (FER). This test specification uses the appropriate digital error rate (as measured by the subscriber unit) to evaluate effective radiated receiver sensitivity at each spatial measurement location. All of the measured sensitivity values for each EUT test condition will be integrated to give a single figure of merit referred to as Total Isotropic Sensitivity (TIS). The BER specification of CTIA on GSM and WCDMA systems for optimal transmitted data rate are 2.44% and 1.22% respectively. TIS measurement not only measures the performance of stand-alone antenna, but also takes wireless device itself into account to realize the practical implementation. We evaluated the EIS (Effective Isotropic Sensitivity) by measuring the minimum received power that met the BER requirement on the test position. The TIS result is able to clearly show the 3-dimensional receiving performance of wireless communications device under specific mobile communications environment.

The complete setup for TIS measurement should at least consist of following hardware instruments: fully anechoic chamber, measuring antenna(s), base-station communications emulator, RF relay switch, high quality coaxial cable, control PC, and position controller. The practical setup for TIS measurement is shown as Figure 4. The operation principle of Figure 15 is as following: Connect control PC to the base-station communications emulator and then make the base-station emulator send test signal to transmit antenna. The power level of transmit antenna is set to -60 dBm for TIS measurement. The power level decreasing step specified by CTIA is 0.5 dB for transmit antenna to measure the minimum power level obtained by receiving antenna. When the transmitted power has been attenuated to some lower level and signal received from receiving antenna to base-station emulator with BER worse than 2.44% (GSM)/1.22% (WCDMA), then we have the minimum receiving power for Netbook.

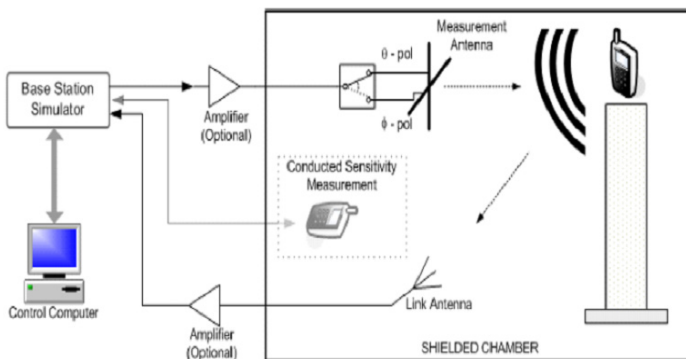


Fig. 15. Setup for TIS measurement.

The results of TIS measurement are shown in Table 1 and 2. The High/Mid/Low channels of GSM850/900 and DCS1800 systems are selected to analyse the platform noise level impact on TIS measurement as shown in Table 1. We found that TIS performance is getting worse as in-band noise increases. Table 1(c) shows the relationship between platform noise level and TIS on CH 512 and CH 698 of DCS 1800 system, it indicates that TIS has 5dB degradation as platform noise increases 2 dB. Table 2 shows the TIS measurement result of WCDMA of Netbook. From the platform noise level and TIS comparison between CH 4357 and CH 4408 of WCDMA Band-5, we found that TIS has 2dB degradation as platform noise increases 1.5 dB. From the observation above, we briefly conclude that TIS performance of the wireless product degrades 2 dB whenever intra-system platform noise level increases 1 dB. It therefore shows that the platform noise is the major factor affecting the receiving sensitivity of wireless devices.

GSM 850	TIS (dBm)	NFS (dBm)
CH128(869.2MHz)	-100.73	-111.39
CH190(882.6MHz)	-102.73	-112.17
CH251(893.8MHz)	-101.19	-112.63

(a) GSM850

GSM 900	TIS (dBm)	NFS (dBm)
CH975(925.2MHz)	-100.74	-108.97
CH037(942.2MHz)	-101.43	-113.18
CH124(959.8MHz)	-100.08	-113.72

(b) GSM900

DCS 1800	TIS (dBm)	NFS (dBm)
CH512 (1805.2MHz)	-103.76	-113.48
CH698 (1842.4MHz)	-98.91	-111.49
CH885 (1879.8MHz)	-102.49	-114.51

(c) DCS1800

Table 1. Measured TIS on WWAN bands of Netbook (a) GSM850 (b)GSM900(c)DCS1800.

WCDMA 1	TIS (dBm)	NFS (dBm)
CH10562 (2112.4MHz)	-105.44	-114.74
CH10700 (2140MHz)	-103.74	-113.12
CH10838 (2167.6MHz)	-106.32	-115.03

(a) WCDMA Band1

WCDMA 5	TIS (dBm)	NFS (dBm)
CH4357 (871.4MHz)	-102.37	-110.63
CH4408 (881.6MHz)	-104.26	-112.17
CH4458 (891.6MHz)	-104.24	-112.56

(b) WCDMA Band5

Table 2. Measured TIS on WCDMA bands of Netbook.

3. RF coexistence problems on product performance [7]

Due to the increasing add-on functions demand for consumer electronics, currently multi-radios, such as WLAN, WWAN, GPS, Bluetooth, and even DVB-H modules, have all been crowdedly embedded and highly integrated in a tiny space of wireless communications platform. Therefore the wireless devices usually have been equipped with more than one antennas, the purpose is to fit for different communication system such as cellular mobile communications, wireless local area networking, and personal area networking. Under this situation, the performance of various kinds of wireless communications is usually degraded by the mutual coupling and interference of closely arranged antennas inside the mobile device. Since the RF modules co-existence has become a critical design problem for wireless communications, we will discuss the RF coexistence problems in this section.

3.1 Isolation required for RF coexistence

Platform noise usually raises the RF receiver noise floor and dramatically degrades system performance by push the E_b/N_0 to the margin when there in-band or out-of-band interference exists. A frequent cause of poor sensitivity on a single channel, or a small number of channels, is due to receiver's in-band noise from broadband digital noise or spurious signals from other coexistent transmitters. We describe in this section the potential coexistence problem for multimode and multiband RF modules, and also illustrate below in Figure 16 the example of isolation required for various RF systems to achieve better service.

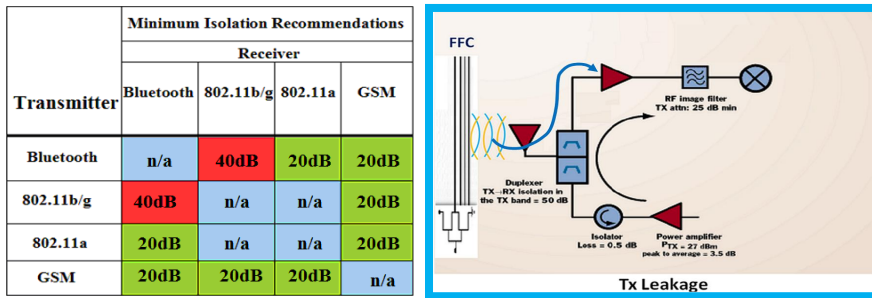


Fig. 16. Tx leakage in FDD system and isolation budget.

3.2 External modulation problem

Even the low-speed digital I/O traces or cables in TP may induce GFSK modulation current, when they are nearby Bluetooth module operating at 2.4GHz. The non-linear ON/OFF switching of digital signal would also play a role as pulse modulation and generate magnetic field through those traces or cables to interfere the DCS and PCS systems. Some extreme case would also happen to WCDMA. The external modulation phenomenon can't be measured by network analyzer until the TP is activated as shown in Figure 17. We also illustrate in this section how the external modulation effect could be found in the final design stage of product with WWAN and BT RF modules ready, however the re-design is needed when platform and TP operate in their normal mode.

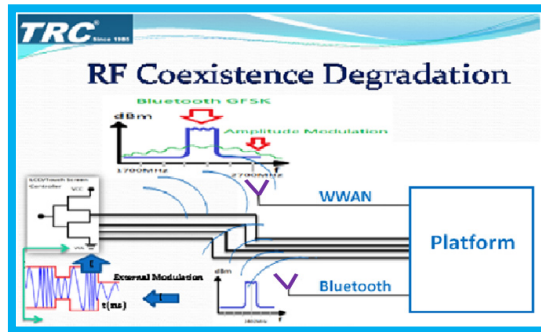


Fig. 17. External Modulation Effect.

4. Platform noise coupling mechanism

The interfering noise sources mentioned above may introduce adverse noise to the nearby wireless modules via conduction or radiation coupling or even both. The digital noise coupled from broadband digital or narrow-band RF devices may result in in-band interference and further degrade the performance of wireless communications, and vice versa. Since the digital noise would cover a wide range of frequency, we will illustrate in Figure 18 and 19 the three potential coupling mechanisms between the noise sources and victim circuit or module: conducted coupling, crosstalk coupling, and radiated coupling.

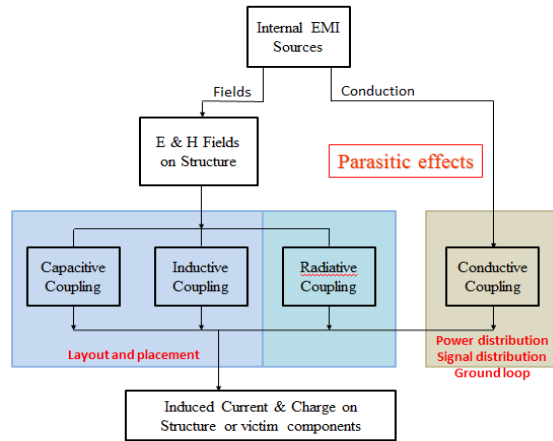


Fig. 18. Different coupling mechanisms.

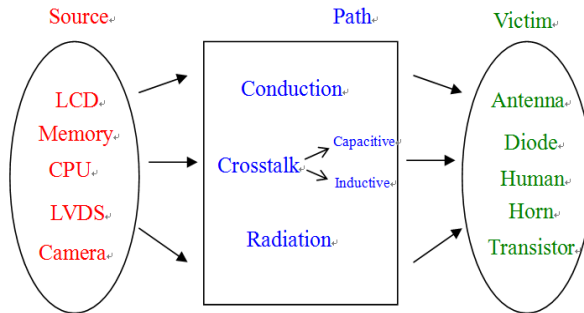


Fig. 19. Illustration of possible noise coupling for wireless device.

The Figure 20 illustrates the EM interaction between TP and embedded antenna, and the photograph of Figure 21 shows the DCS1800 and PCS 1900 transmitted power coupled to LCD panel.

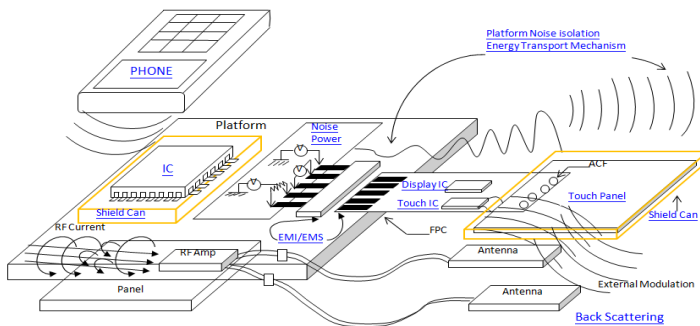


Fig. 20. EM interaction between TP and embedded antenna.

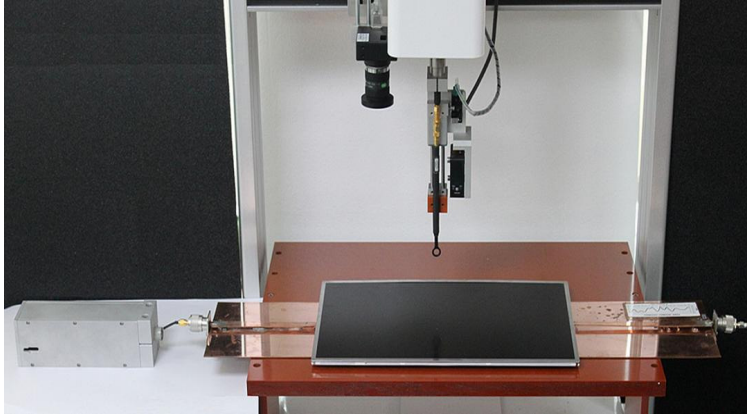


Fig. 21. Illustration of DCS1800 and PCS 1900 coupling to LCD panel.

4.1 Analysis of conductive coupling

The first step of the degradation of sensitivity (De-sense) measurement is conducted testing, because understanding how the interference platform noise conducted to the RF receiver is the most important issue for further analysis. Even the same probability distribution function of noise, there is possible to cause different De-Sense impact depending on the receiver implementation.

The best way to obtain the conducted De-Sense effect is to use the internal WWAN module or chip set of mobile device for RSSI measurement as shown later in Figure 25. The left figure shows a 50 ohm terminated at antenna connector to read the WWAN RSSI data. The figure in the center shows the dummy WWAN module with circuit ground and chassis ground, and the WWAN card inside the shielded box is connect to the dummy WWAN card via coaxial cable to read the RSSI data again. The right figure shows the dummy WWAN module with chassis ground and WWAN card to read the RSSI data. The RSSI data read from the same RF receiver with three different conditions described above will help engineers easily identify the platform noise.

4.2 Analysis of near-field coupling from antenna to embedded devices

Since the antenna is usually implemented in the proximity of TP and camera, the near-field coupling to the nearby TP and camera would sometimes result in malfunction due to the transmitted power (Figure 21). The electromagnetic field distribution on TP due to antennaradiation is shown in Figure 22 below.

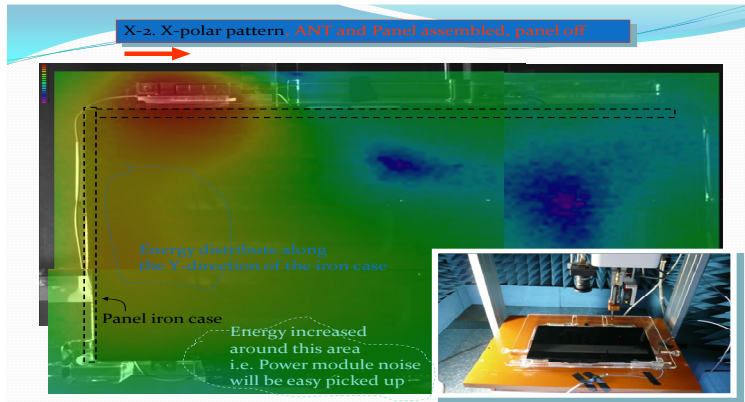


Fig. 22. Antenna radiation coupling.

4.3 Resonance issues of mechanic structure

The LCD panel of Notebook and the enclosure or PCB would usually create a resonant structure. The related resonance issue could be explained with the following measuring methodology.

1. Connect port 1 of network analyzer to LVDS cable and WEB camera cable via a coupling fixture (Balun and low-mu ferrite core as absorbing clamp) for energy transferring to the LVDS cable.
2. Connect port 2 of network analyzer to the embedded antenna of Notebook via its own mini-coaxial cable.
3. Measure the VSWR of port 1 and port 2 respectively. Both VSWR must be less than 2.5 to ensure the testing setup is good enough for efficient energy transfer.

When enclosure of LCD panel was removed, the maximal difference of S_{21} mentioned above can reach upto 20dB. Hence we can conclude that LVDS and Web camera cables, embedded antenna and its mini-coaxial cable, LCD panel all will lead to amplification of coupling between the interference sources and receiving module.

5. Platform noise measurement techniques

Conductive or near-field coupling platform noise will raise the level of RF receiver noise floor and thus cause performance degradation, and thus we need to investigate the isolation required as shown in Figure 23. We will describe the related measurement techniques in this section.

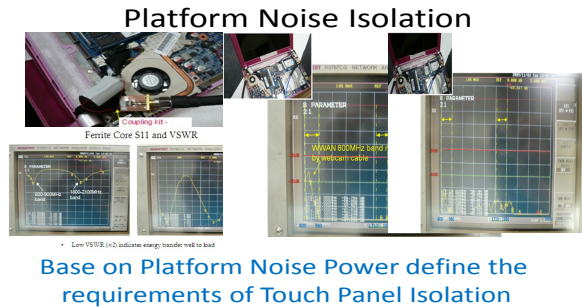


Fig. 23. Platform noise coupling.

5.1 Noise level measurement

The result of noise level measurement represents the de-sense degree caused by IC, component, module, power circuits, PCB layout, interconnect wires, connector, and even the mechanical construction of the product. This section will describe the different methodology for the measuring procedures.

5.1.1 System noise measured with spectrum analyzer

- Measured by frequency domain sweep for multi-bands (GSM 850, 900, PCS1800 etc.) as shown in Figure 24.
- The Noise Distribution Function at area A is close to Gaussian distribution (Maximum Hold - Average = 10dB), area B is close to CW (little amplitude variation with time)
- Noise limit level means the probability density function (PDF) of noise power is equivalent to Gaussian. Limit line for area A applies for the calculation described above, but PDF correlation is needed for area B.

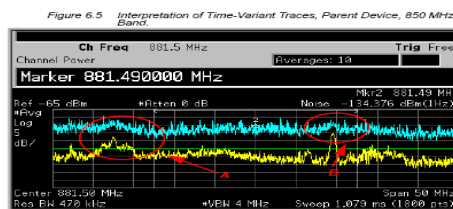


Fig. 24. System noise measured with spectrum analyzer.

5.1.2 System noise measured with WWAN card

In practice, it is more reasonable to measure system noise with a WWAN card because the RSSI is ready to appear at I/Q demodulator output, since it is the same sampling clock and module size for WWAN communication. The noise power measured from RSSI of WWAN add-on card can therefore in compliance with the definition of N_0 . The testing configuration and measurement are both shown in Figure 25.

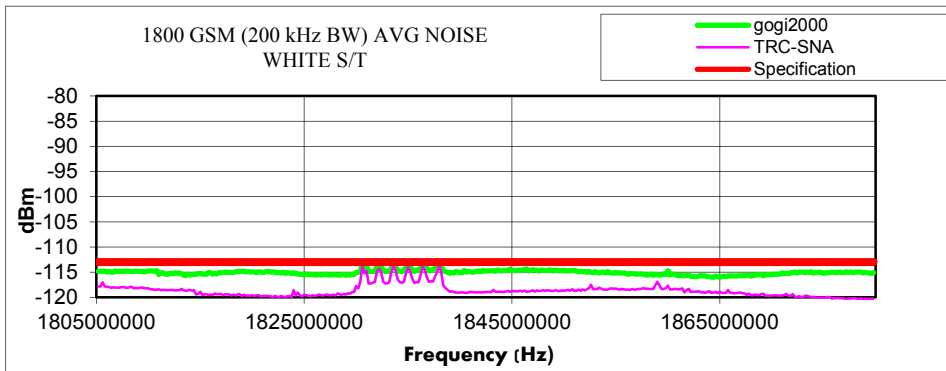
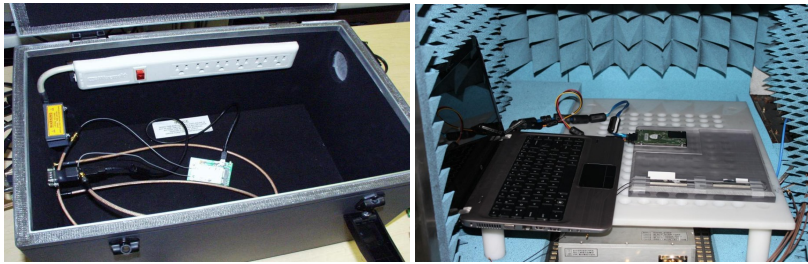


Fig. 25. Test configuration and system noise measured with WWAN card.

The SNA option shown in Figure 26 can also provide the De-Sense Measurement function. The software provides the Base-Station simulator, and provides VSG for WWAN and GPS measurement. The hardware provides switch and combiner for testing signal condition selection.

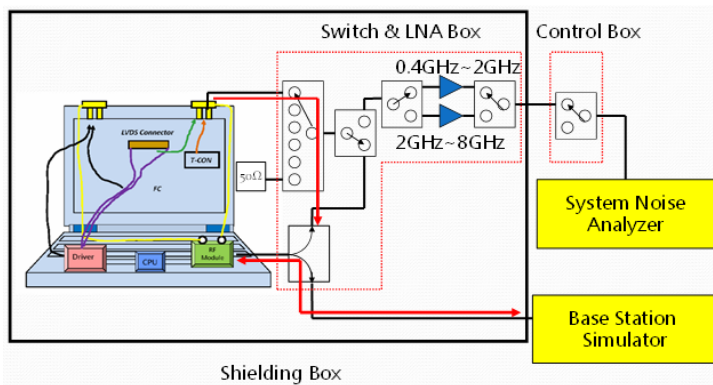


Fig. 26. The SNA test configuration.

The benefit and applications of SNA can be listed as following:

- Provide measurement adapter for fixing the antenna cable routing inside NB and Tablet PC.
- Provide testing fixture for different size LCD display panel.
- Provide testing fixture for SSD noise budget measurement
- Provide measurement adapter for ground isolation between smart phone and SNA (two Balun back-to-Back in series)
- Provide debugging plot to view each band at the same time.

5.1.3 SNA calibration at limit level

Since the WWAN or LTE cards are not originally designed for the noise measurement, and these cards need to provide about 60dB dynamic range for receiver. Therefore the variable-gain amplifier must be utilized for AGC purpose at the same time when the noise floor of receiver reaches up to 6dB. SNA is here designed to measure the platform noise. The dynamic range of receiver is from 4dB under limit line level up to 16dB above limit line level. With front end LNA implemented, the noise floor of SNA system is around 2dB and the average noise floor level is 4dB lower than that of WWAN or LTE card.

The noise floor level for WWAN is -115dBm. However, the limit line level for some particular applications like identifying the main noise source (system noise or panel noise) will be -114dBm. We can use the receiver with -115dBm noise floor level to receive the noise with level of -114dBm, and there is 1dB uncertainty for signal to noise level result. Since the average noise floor level can be obtained for SNA is -119dBm, it can provide more accurate signal to noise ratio for -114dBm measurement in general cases as in Figure 27.

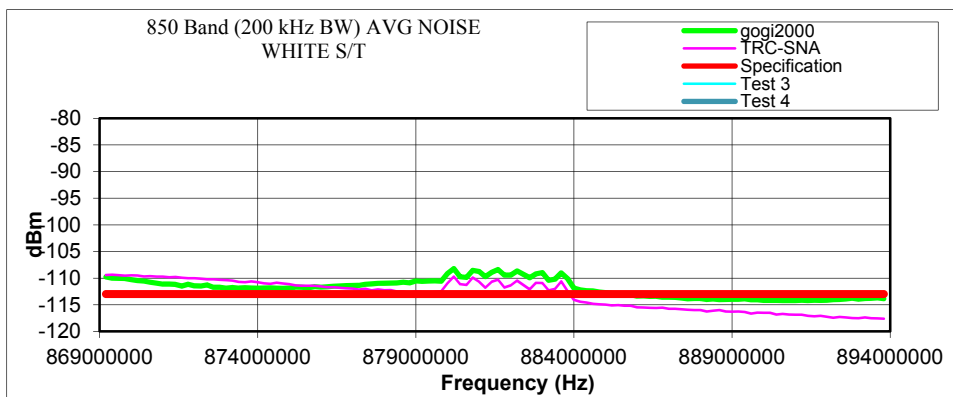


Fig. 27. Noise floor level obtained for SNA.

5.2 Throughput measurement

The throughput standard is widely adopted in different applications like video stream, online game, video conference. For example, you may need 1Mbps or more for YouTube HD video. Throughput is the most common for user experience regarding wireless RF

performance, therefore throughput test need to consider how the user interacts with the RF device. For example, 90dB path loss for laptop 802.11g WLAN may represent that AP is 200 meters away in the free space environment, however the hand and head may cause the range for 90dB path loss much less than 20 meters for smart phone VOIP application. Some real-life products throughput test result has been illustrated in Figure 12 and the Root Cause Analysis (RCA) procedure will be addressed later in section 6.

5.3 Antenna surface current and near-field surface scanning

The current distribution on antenna surface represents the different sensitivity on antenna near field boundary, because the physical geometry of antenna will cause different field intensity coupled via the uniform magnetic flux. Antenna surface current also represents the immunity level for TP and Camera. As to digital noisy components, we can utilize the noise level results to locate the noisy components and identify their noise radiation pattern. We can utilize the near-field surface scanning method to observe the surface current distribution of antenna and locate the noise sources for platform noise analysis.

5.4 De-Sense measurement

De-Sense is self-referred to same subjective device operating in different condition and compared the platform noise impact. For example, a GPS module is first performed conducted test in a shielded box and obtains the $C/N=40\text{dB}$ at -138dBm receiving level. While the same module is then bundled in a wireless platform and performed the conducted test again, the receiving level become -130dBm for keeping $C/N=40\text{dB}$. In this example, we found that the conducted platform noise causes GPS module de-sensed by 8dB. When the antenna terminal and GPS signal is fed to GPS receiver through a combiner, we now need -123 dBm to keep $C/N = 40\text{dB}$ and thus a 7dB de-sense caused by platform noise picking up from antenna. Furthermore, when the internal device, like Bluetooth, is active or hand-held device resting on desk, the interaction between platform noise and antenna resulting in different de-sense effect can be easily observed in all those test cases. The noise current at different locations can also be represented with the de-sense Data. The methodology and measuring technique illustrated below in Figure 28 can be applied for investigating GPS degradation of sensitivity caused by platform noise.

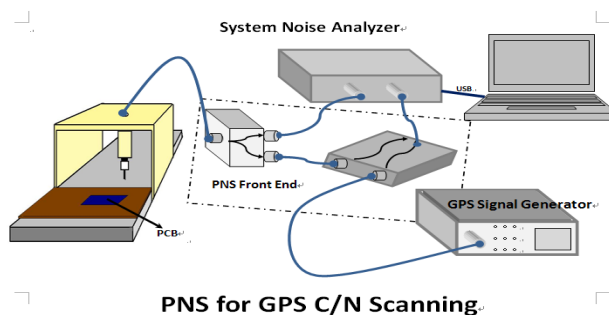


Fig. 28. GPS C/N Scanning.

5.5 Noise measurement at antenna terminal

Since the antennas are the most important port for wireless communications as electromagnetic energy receiving component, they are also susceptible to nearby platform noise. Hence the analysis and measurement of noise level at antenna port is critical for RCA of wireless device to improve link performance. Figures 29-31 show the measurement techniques and configuration for noise measurement at antenna terminal.

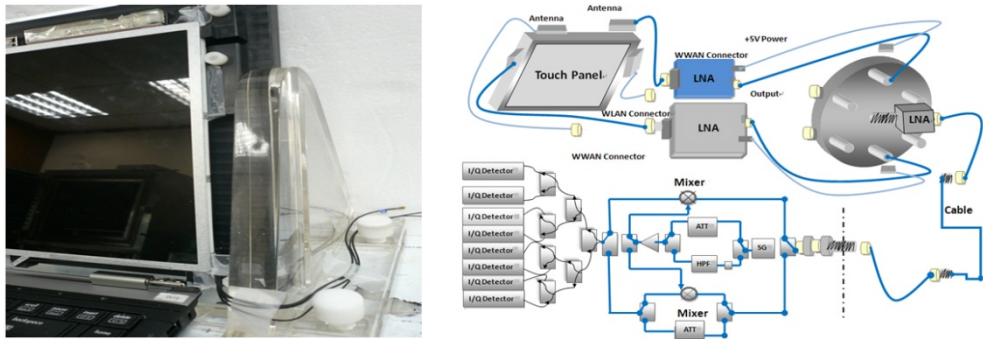


Fig. 29. LCD panel test fixture and the measurement circuits.

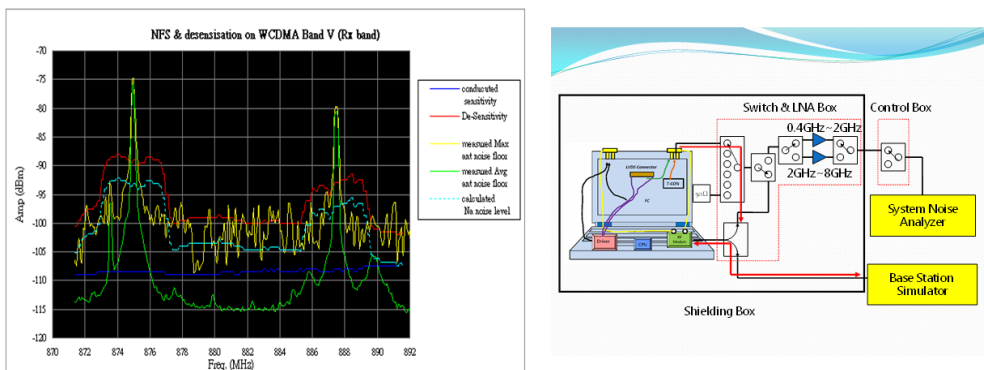


Fig. 30. Platform noise and de-sense measurement at antenna terminal.

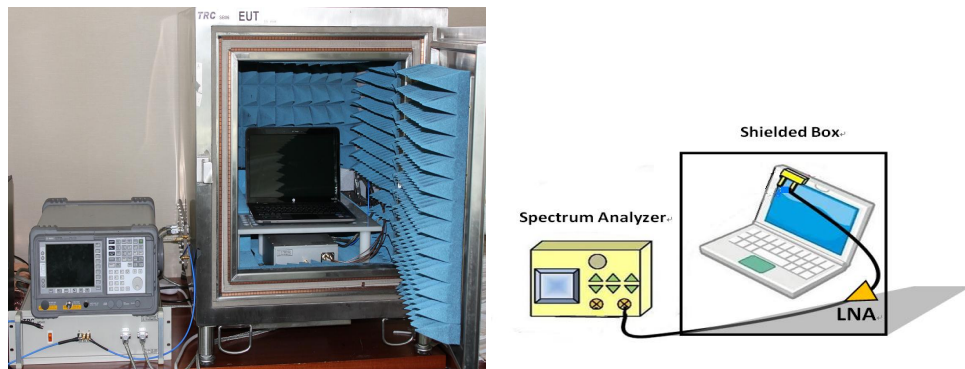


Fig. 31. NFS system and measurement setup.

From the various platform noise measuring techniques described above, we can summarize the comparison as Table 3.

	Measurement Type	Measurement Speed	End to End Calibration	Software User Interface	Correlation
SNA Same LNA and Switch as TRC NFS	Channel	10 times faster than WWAN card	Gaussian Noise Level Cal. At limit line level.	Same as the NFS & Provide the debug mode "View Plot" display test result	Within Limit Line + - 3dB range within 1dB error with WWAN & LTE Card
WWAN Card LTE Card	Channel	Very Slow	N.A .	Depend on Card Vendor of NPT utility	Same WWAN card cab be 2dB error
TRC NFS	Sweep Freq.	Very Faster	LNA & Switch	Well Accepted	Big Difference in Broadband

	Receiver Noise Floor Level	Antenna VSWR Check Before Noise Measurement	Test Fixture, Test Kit and Measurement adapter	De-Sense versus Platform Noise Plot
SNA	Average At 200KHz Channel -119dBm	SNA provide the option for the internal VSWR bridge, Each time debug and close the enclosure should be check the antenna, ensure the antenna is keep the same as original.	Measure the Adapter Insertion Loss for Correction. LCD Test Fixture Can be integrated with LNA and the antenna protect by dielectric material.	Software support the for different brand of base station simulator and VSG, Measure the De-Sense corresponding to the Noise Level.
WWAN Card LTE Card	Average -115dBm	N.A.	N.A.	N.A.
TRC NFS	Depend on Spectrum.	N.A.	N.A.	N.A.

Table 3. Measurement techniques comparison.

6. Design techniques for platform noise improvement

This section will present the Tablet PC as case study to describe the problem-resolving methodology for throughput degradation.

6.1 Identify the main noise source

The first step to solve the interference problem is to identify the main noise source, and then we can further to implement resolving techniques like filtering, shielding, and re-layout, etc. The procedure can be demonstrated below in Figure 32.

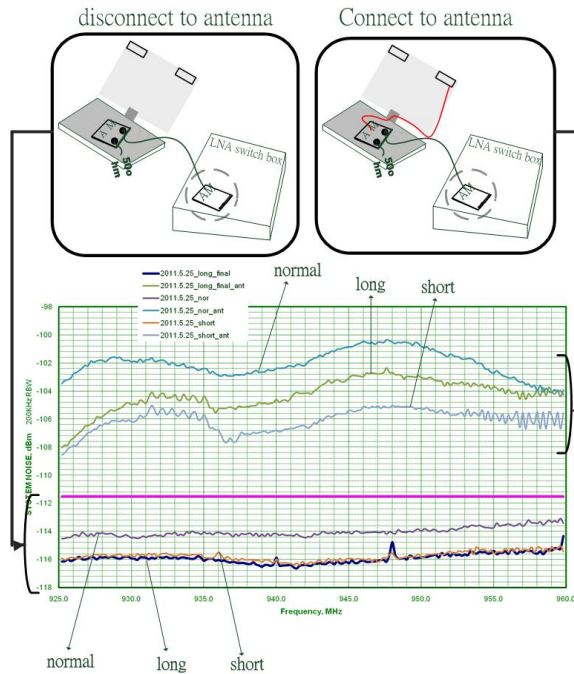


Fig. 32. Use built-in antenna to identify noise source from measured spectrum.

Those engineers who are responsible for resolving noise problem may add the copper foil or put an absorber to cover the IC, routing the antenna cable, re-installing the panel, plugging antenna connector to test port, and they may affect the antenna or noise test results after those activities. Therefore, the antenna VSWR characteristic must be checked before noise measurement. The VSWR measurement of antenna is a quick way to check if the antenna is still kept the same configuration as originally implemented, because 3D radiation pattern and efficiency measurements of antenna usually take 2~3 hours. However, the SNA option described early can provide an embedded VSWR bridge and therefore could automatically check VSWR of each measuring port before noise measurement as shown in Figure 33 for LTE band. There are two kinds of LCD panel testing fixture can be designed for different panel size as shown in Figure 34 and 35.

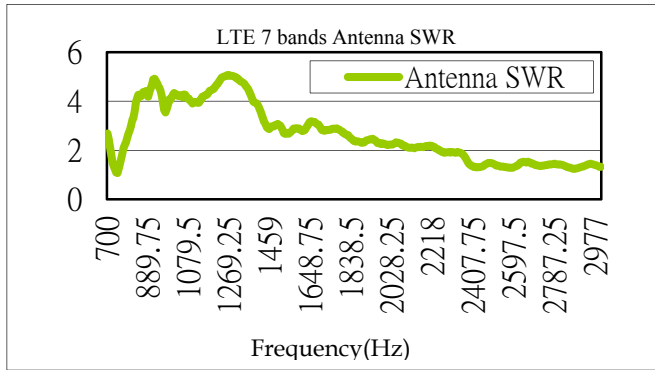


Fig. 33. VSWR of each measuring port before noise measurement for LTE band.

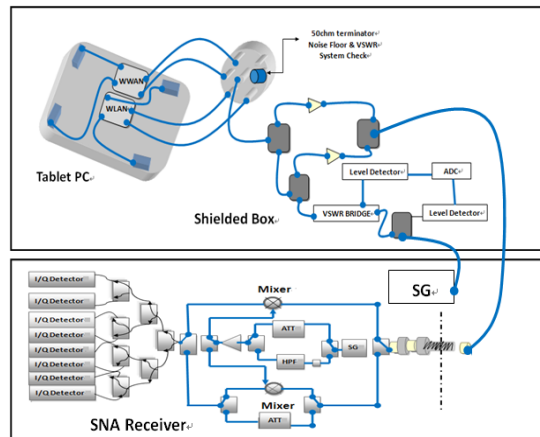


Fig. 34. LCD panel testing fixture integrated with LNA and SNA result for LTE.

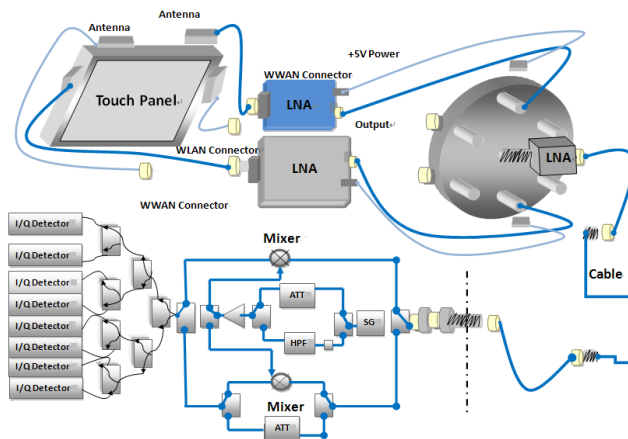


Fig. 35. LCD panel testing fixture with LNA and antenna protected by dielectric material.

6.2 Platform noise isolation

TP and camera are not the only modules resulting in the platform noise problem, but the product's assembly construction such as antenna and its mini-coaxial cable routing, TP and its LVDS cable, camera and its USB cable will also bring up noise problem. We can utilize some design techniques, like component placement and orientation, cable routing, shielding etc. to improve platform noise isolation.

6.3 Antennas isolation [8]

Noise current distribution and antenna surface current are most important message for solving the sensitivity degradation problem. We can use a network analyzer deliver the energy to TP's LVDS or camera's USB lines and measure the insertion loss at antenna port, and that is the most commonly utilized technique to obtain the isolation situation of platform noise.

With the highly integration of powerful computing and multi-radio communications devices in a single product nowadays, multiple antennas are usually implemented to achieve the seamless and convenient communication services. However, the closely placed antennas have resulted in intra-system coupling interference and therefore severely degraded the performance of various kinds of wireless communications. The isolation technique for antenna systems must be implemented to reduce the mutual coupling between coexistent various RF systems. In this section, we will show the optimal isolation achieved from antennas separation, orientation, and utilization of periodic structure to reduce the mutual coupling interference.

The isolation requirement between coexistent RF systems is shown in Table 4. The placement of two chip antennas under investigation for Bluetooth and 802.11b/g WiFi systems inside the mold notebook computer is shown in Figure 36. The chip antennas are fabricated on FR4 with dimension of 1.6 mm thickness and 35mm × 30mm area, and it is fed by microstrip to achieve 50 Ω impedance-matching. The configurations for different spacing and orientation between coexistent antennas are shown in Figure 37 to analyze the mutual coupling effect.

Transmitter	Minimum Isolation Recommendations			
	Receiver			
	Bluetooth	802.11b/g	802.11a	GSM
Bluetooth	n/a	40dB	20dB	20dB
802.11b/g	40dB	n/a	n/a	20dB
802.11a	20dB	n/a	n/a	20dB
GSM	20dB	20dB	20dB	n/a

Table 4. Isolation requirement between coexistent systems.



Fig. 36. Placement of two chip antennas insidemold.

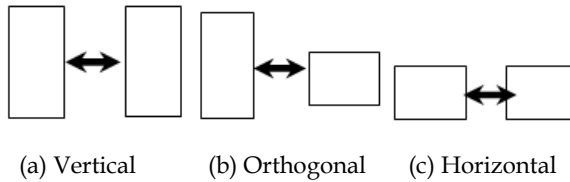


Fig. 37. Three orientation for antennas placement.

The reflection coefficients (S_{11} and S_{22} are reflection coefficients for Bluetooth and 802.11b/g respectively) and isolation (S_{21}) between each other were then measured and shown in Figure 38. The results clearly show that the isolation between antennas is much worse when they are both placed in vertical polarization with main lobes coupling. We then oriented the antennas in orthogonal direction to each other and separated antennas by moving 0mm, 10mm, and 20mm respectively. The measured results of reflection coefficient covering 2.4GHz-2.483GHz in Figure 39 show that the isolation is much better due to orthogonal polarization to each other when they are oriented in orthogonal direction. Finally, we placed both antennas in horizontal direction and adjusted the separation between them, the results in Figure 40 also show better isolation than vertical placement due to main lobes decoupling. Table 5 compares the isolation performance between various orientation and separation for both antennas, and it shows that the orthogonal and horizontal orientations gain almost 6-9dB improvement in isolation except 4 dB difference for 0mm separation.

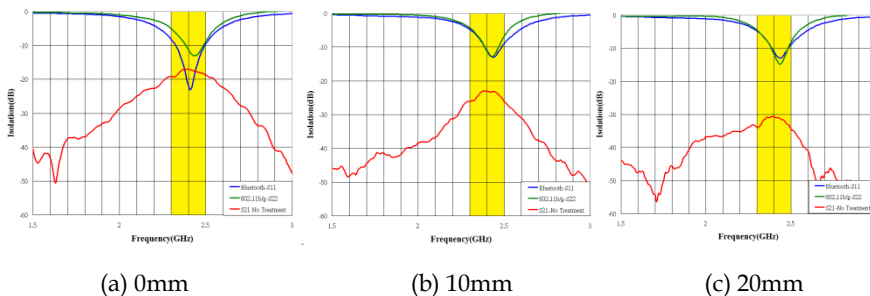


Fig. 38. Measured results for various antenna spacing with both antennas oriented in vertical direction.

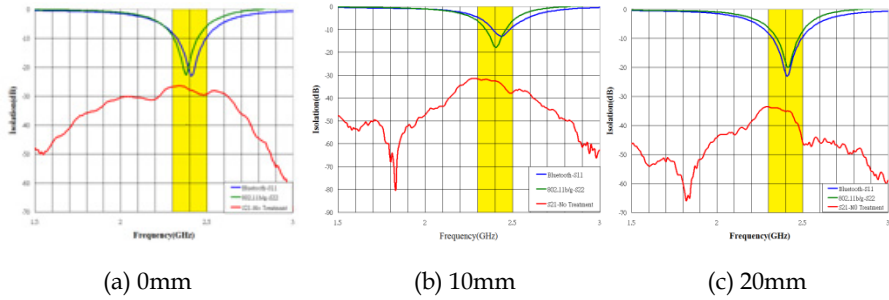


Fig. 39. Measured results for various antenna spacing with both antennas oriented in orthogonal direction.

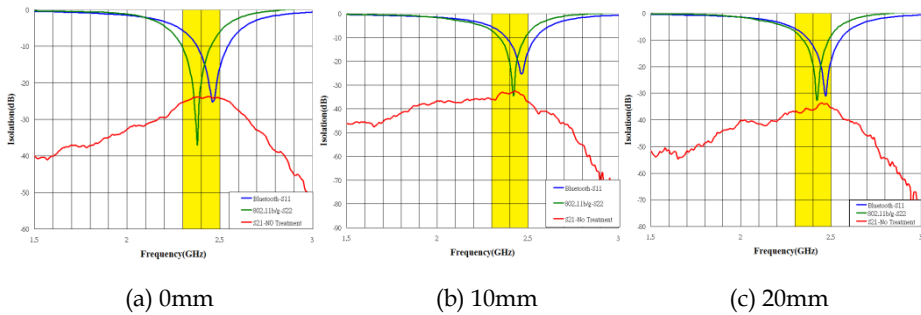


Fig. 40. Measured results for various antenna spacing with both antennas oriented in horizontal direction.

Orientation \ Separation	Vertical	Orthogonal	Horizontal
Isolation			
0mm	16.9dB	28dB	24.1dB
10mm	23.1dB	32.6dB	33.1dB
20mm	30.8dB	36.2dB	36dB

Table 5. Measured results for various antennas placement configuration.

6.3.1 Suppression of mutual coupling interference between coexistent antennas [8]

The applications of EBG structure in antenna not only could improve gain and radiation efficiency, it could also help suppress side lobes and reduce coupling effect. Since isolation requirement could not be met by orientation and separation arrangement between antennas from the above measurement, we therefore chose the best placement configuration with orthogonal orientation and 20 mm separation for further investigation utilizing EBG

structure. We placed the EBG structure beneath the antennas with 7.5mm (less than $\lambda/4$) and 15mm (about $\lambda/4$) distances as shown in Figure 41 and investigated the mutual coupling characteristics. Figure 42 shows the results with 7.5 mm separation between antennas and EBG structure for various antenna orientations. Because of high impedance surface from EBG structure, the 40 dB isolation requirement between antennas could be achieved and parallel-plate guided wave coupling could also be suppressed. Figure 43 shows the results with 15 mm separation between antennas and EBG structure for various antenna orientations.

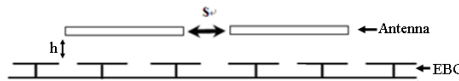


Fig. 41. Configuration of antennas and EBG structure placement.

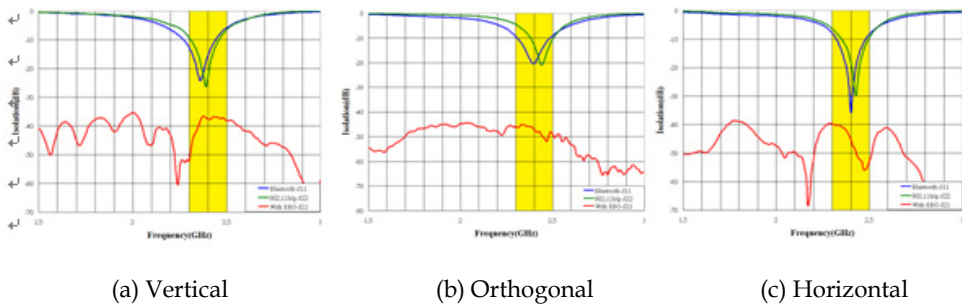


Fig. 42. Isolation and S11 of antennas for various orientation with antennas separation 20mm and EBG-antennas spacing 7.5mm.

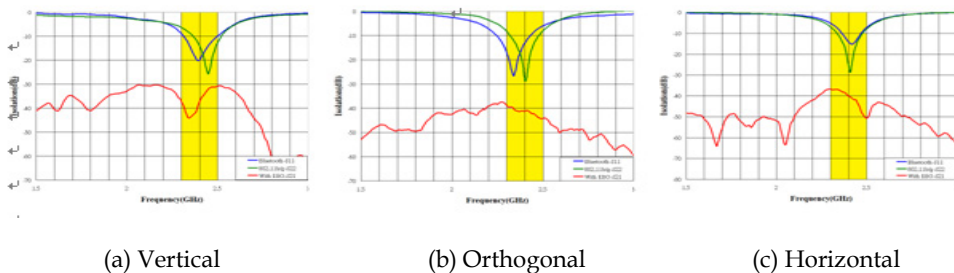


Fig. 43. Isolation and S11 of antennas for various orientation with antennas separation 20mm and EBG-antennas spacing 15mm.

Table 6 compares the isolation performance with and without EBG structure for various antenna orientations, and it shows that 37.7 ~ 48dB and 35.7 ~ 40.9dB isolation between antennas can be achieved with EBG structure placed beneath antennas 7.5mm and 15mm respectively. When EBG structure moves closer to antennas, we can obtain better isolation between antennas.

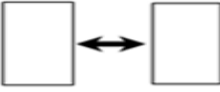

Orientation \ Separation Between Antennas and EBG	Vertical	Orthogonal	Horizontal
	Isolation		
Without EBG	30.8dB	36.2dB	36dB
7.5 mm	37.7dB	48dB	47.4dB
15 mm	35.7dB	40.9dB	40.5dB

Table 6. Comparison of isolation improvement from EBG structure with antennas separation 20mm.

6.4 Keyboard grounding requirements

The large metal plate of a keyboard can act as an integrated shield to prevent noise radiating from the base chassis. However, the mechanical grounding structure needs to be taken serious care of its EMI effect on radio performance, because it is not uncommon that EMI noise radiated from the ground-base to antennas mounted around LCD panel when the keyboard is removed as shown in Figure 44. This means that we are not getting any shield effect from the keyboard, and in fact the keyboard helps EMI ground noise radiating from the base like a large antenna. To make the keyboard a useful shield for all WWAN frequencies, it is necessary to conductively contact ground structure every 1/20th wavelength or so as shown in Figure 45.

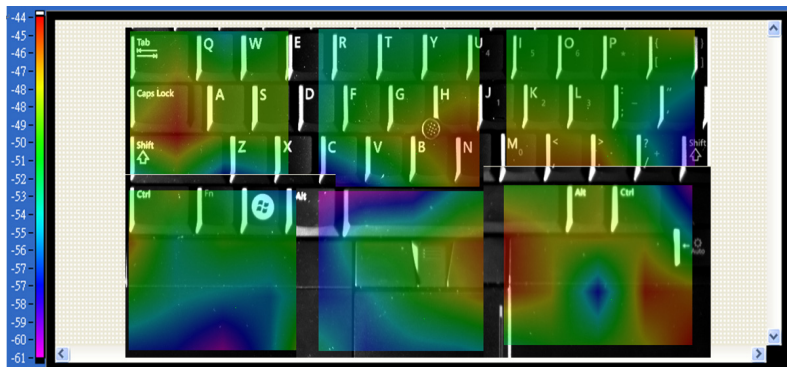


Fig. 44. Field distribution on metal shield of LCD panel and nearby antenna.

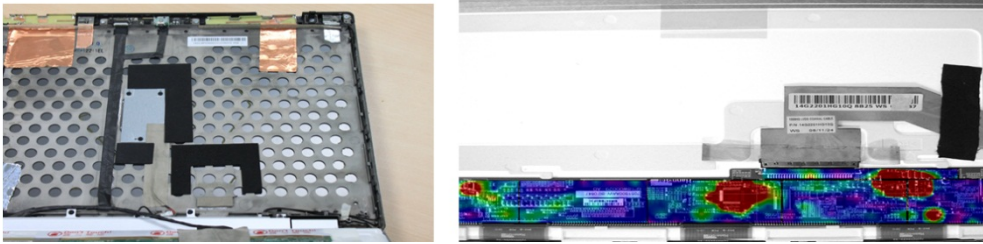


Fig. 45. Metal plate for LCD panel shielding purpose.

Grounding the motherboard to the chassis wherever possible (screw holes, connectors, etc.) will help reduce EMI radiation from the motherboard. The grounding contacts should follow EMC design guidelines for length to width ratio requirement to prevent from adding a radiation structure rather than providing a noise-reduction ground point. Every effort should be made for each screw and contact point of the motherboard to the chassis for a good chassis ground point.

The grounding of the heat sink also raises a problem for ESD, EMI, and platform noise of wireless devices, therefore improvement should be made in the grounding of the heat sink (cooler). The improvement should include: enabling better DC grounding at “spring contact” points, and no non-grounded arms longer than 7mm are allowed.

6.5 Component and signaling cable considerations

Two components (local oscillator and clock chip circuit) on motherboard probably will need an extra local shield placed on it to pass regulatory testing if the components radiate over-limit spurious noise. Hence the mechanical chassis design should also provide sufficient space with height clearance to accommodate this kind of SMT shield/can.

The camera module is also a potential RF noise source of the system, especially when it is located in the proximity of antennas. A mechanical shielding solution – sheet metal, EMI paint, foil, and/or magnesium walls should be considered to isolate the EMI noise from the antennas as shown in Figure 46. Although most of the camera modules equipped with metal shielded enclosure, but the glass which cover the CCD gate, transfer gate may leak or coupling the magnetic field to the nearby traces through the glass aperture.

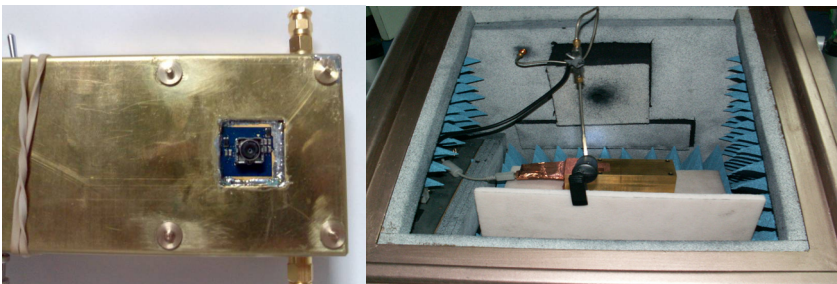


Fig. 46. Shielded enclosure and loop probe for noise measurement of camera module.

LVDS is now a popular signaling system that can deliver information at very high speed over twisted pair copper cables. LVDS technology uses the voltage difference between two wires to carry signal information for high-speed data transfer through the panel hinge to minimize EMI related problems. In order to minimize the aforementioned EMI problems, we can utilize LVDS cables with the following mechanical recommendations:

1. Make sure that the twisted pair cabling, twin-axial cabling, or flex circuit with closely coupled differential lines is used by grouping members of each pair together.
2. Differential impedance of the cables should be 100 ohms
3. Make sure that the cables used are well shielded
4. Place ground pins between pairs wherever it is possible
5. Connect shielding directly to the connectors of driver and receiver enclosure respectively.

The use of LVDS system must be cautious that most of the LVDS cables have good degree of balance for the fundamental frequency, but the balanced pair becomes unbalanced for those frequencies higher than 10 harmonics when the harmonic signal across the LVDS cable. Therefore it will usually generate the common mode voltage resulting in common mode radiation, and the shield of the LVDS cable become ground return.

The impedance mismatching at DDR2 or DDR3 memory socket usually causes higher noise current distribution around the socket area. These magnetic field may couple to the WWAN module when the shielding effectiveness of WWAN module's shielded enclosure is not adequate as shown in Figures 47-49. Lower impedance will make dI/dt increase and dramatically increase the current drawn from supply (not good for the design of power distribution system), while higher impedance will emit more EMI and also become more susceptible to external interference.



Fig. 47. RF module location and nearby field distribution.

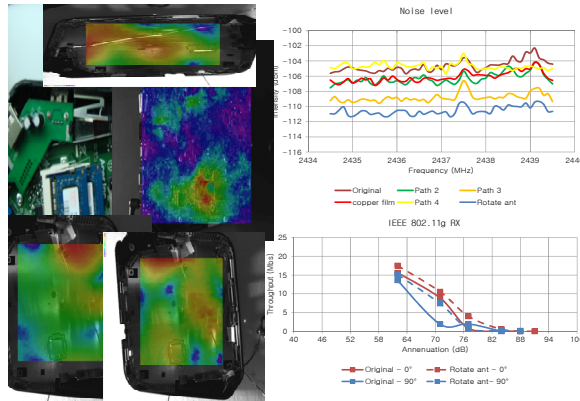


Fig. 48. Embedded antenna mini-coaxial cable and noise distribution.



Fig. 49. Circuit ground and chassis ground.

7. Application of noise budget concept

Finally, we will propose the Noise Budget concept for platform noise suppression. The noise budget for the wireless communication device can be considered as near-field EMC limit. Noise budget is a powerful tool to apply for the wireless product from initial design stage, QA, QC, and all the way to final production testing. For communications community and RF device manufactures, the link budget is established for system planning, therefore most of RF engineers understand that the characteristics of antenna, LNA, mixer play important role for range and coverage extending. However the parameters discussed in link budget are all about signal transmission alone, another important component parameter related to noise level, the Noise Budget, is the missing puzzle for the system integration.

7.1 Introduction to noise budget concept

Since the electronics of the notebook or laptop are the interference source for RF wireless device as discussed earlier. This final section covers some design guidelines and EMI measuring techniques of components, because EMI from internal ICs is also a major contribution to impact the RF performance. The primary purpose here is to address the idea of “Component Noise Budget for Wireless Integration”. The concept of noise budget for devices on wireless communication product stems from the link-budget for RF Tx/Rx performance. It also borrows the idea from EMC testing requirements for automobile industry to identify the potential interference sources that might cause safety problem and

thus to provide design guideline for compliance. The noise budget concept helps system designers to manage the EMC issues, as early as possible, such as coupling mechanisms, module placement, grounding, and routing for EMC test. The methodology is intended to develop modular architecture of analysis to accelerate system design and also provide the solutions for potential problems to improve performance in all aspects.

The preliminary goal of this research is to establish the noise budget for components and devices on laptop computer for further RF sensitivity analysis. To utilize the near-field EM scanner to detect the EMI sources on laptop, we can locate the major noise sources in 2D hot-spot distribution graph. From the emission levels and locations of the noisy components, we can figure out their impact on throughput and receiving sensitivity of wireless communication and find the solution to improve performance. The final goal of noise budget, however, is to establish the EMI limits for each digital components related to layout location, it would therefore help designers to choose the appropriate components for optimal placement and cost consideration to meet product requirement. The application of noise budget accompanied with near-field surface scanner not only can locate the EMI source and further to solve the problem, but also can utilize the EM analysis to improve the design efficiency and the performance of wireless communications.

The factors that would affect the receiving performance of a wireless receiver can be illustrate in Figure 50. Table 7 shows the relationship between link budget and noise budget for wireless communication system implementation.

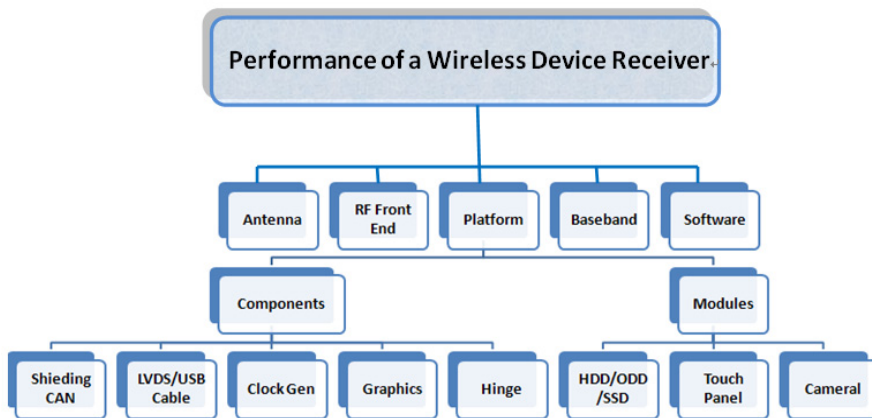


Fig. 50. Factors affecting the receiving performance of a wireless receiver.

PATH LOSS		NOISE LEVEL	
ANTENNA GAIN	-3dB	CPU	
DIPLEXER INSERTION LOSS	-2dB	NORTH BRIDGE(LVDS DRIVER...)	
MISMATCHING LOSS	-0.5dB	MEMORY	
LNA GAIN	+20dB	LVDS SOCKET	
LNA NOISE FIGURE	2dB	LVDS FLAT FLEX CABLE	
SAW FILTER LOSS	-2dB	LCD PANEL (T-CON, BACKLIGHT)	
MIXER CONVERSION LOSS	-0.5dB	TOUCH ME	
SAW FILTER LOSS	-2dB	WEB CAM	
LNA GAIN	+20dB	MINI COAXIAL CABLE FOR RF	
I/Q CONVERSION LOSS	-8dB	SOUTH BRIDGE	

Table 7. LINK budget (left) vs. NOISE budget (right).

The common interference noise sources on integrated high-speed digital wireless communication product nowadays include: CPU, LCD panel, Memory, digital components, high-speed I/O interconnect, wires and cables, etc. The modules' placement of the test setup is illustrated in Figure 51. The above mentioned noises are usually coupled to nearby sensitive devices through radiation, conduction, or crosstalk. The resulted EMI problem will further degrade the system sensitivity and performance for wireless communications.

After the emitted noise level and corresponding location of each component has been identified, we investigate the effect of component placement on in-band noise level at antenna port and thus the performance of wireless communications by changing distance between antenna and component under test. We can further find the optimal orientation and location of component to improve overall communications performance[9,10].

Because a variety of digital components exist inside laptop computer, we focus on LCD Panel that is equipped in all computers and usually placed in the proximity of antennas. To investigate the effect of various operation modes, (such as off, standby, and key-in alphan H pattern mode) on noise level at the antenna port, we first arranged the test setup as laptop normally working and scanned the ambient noise.

To clarify the influence of LCD panel noise on antenna port and thus receiving sensitivity, we first fixed the function setting on computer to avoid effect from software's inconsistent running mode. After activating the LCD Panel for various testing mode, we measured the noise spectrum at antenna port to find out the interference frequencies and then use near-field probes to scan the EMI noise from LVDS cables, connectors, and driver ICs of the LCD panel control circuits. Finally, we can investigate the impact of different LCD operation mode on the frequency bands of wireless communications by analyzing the measured throughput results.

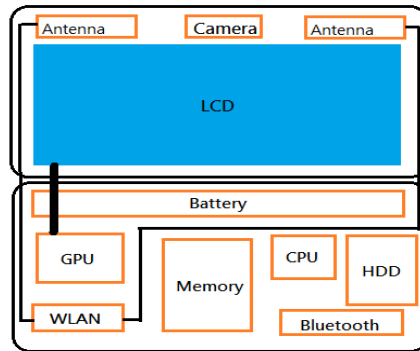


Fig. 51. Internal layout of laptop computer.

7.2 Analysis of platform noise effect from built-in CAMERA module

7.2.1 Test setup for noise level measurement

The system platform noise under investigation is first analyzed by noise floor measurement system. The complete PNS (platform noise measuring system) is composed of shielded box, pre-amplifier, spectrum analyzer, and EUT (Laptop computer). The noise level measuring system and setup for frequency domain is shown in Figure 52.

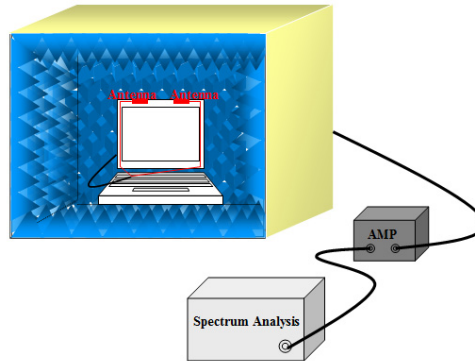


Fig. 52. Setup for antenna port noise level measurement.

Since the CAMERA or CMOS camera module is most adopted to the popular mobile devices like cellular phone or Netbook, we hence focus on EMI analysis of the built-in camera module by application of IEC 61967-2[1].

7.2.2 Test setup for TEM cell measurement[11]

The test setup for TEM cell method in this study is shown in Figure 53. One end of the TEM cell is terminated with a 50Ω resistance terminator, and the other end is connected to spectrum analyzer via pre-amplifier.

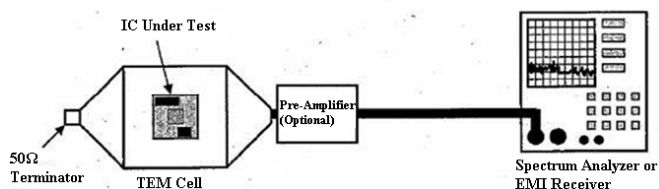


Fig. 53. TEM cell setup for IC or module EMI test.

There are two PCBs with module under test on it and are marked as 1 and 2 shown in Figure 54 and Figure 55. The tested boards for Webcam DUT are driven by USB with clock frequency 48 MHz, and the grounding connection between camera chip and PCB is utilized with wire bonding. The function of module and inner PCB routine for both boards under test are identical except for different number of bonding wires connecting to ground, there are more grounding wires for No. 2 PCB than No.1. The purpose of this measurement is to investigate the effect of multi-point grounding scheme to EMI level. Since the bonding wire is equivalent to inductance, we expect to reduce ground bounce and hence the EMI emission by parallel connection of multiple grounding wires. The connection between camera module and testing board is shown in Figure 56.

The experimental procedure for EMI test using TEM cell is following:

1. Connect the pre-amplifier (if needed) in front of spectrum analyzer at one end, and connect a 50Ω terminator at the other end.
2. Define or identify the four side of TEM cell to place the DUT oriented along all four directions and measure EMI one for each time as shown in Figure 57.
3. Set the measurement frequency range of spectrum analyzer from 150 kHz to 1 GHz.
4. Set the resolution bandwidth of spectrum analyzer around 9 to 10 kHz, and video bandwidth as more than three times of resolution bandwidth to meet the IEC standard specification[12].

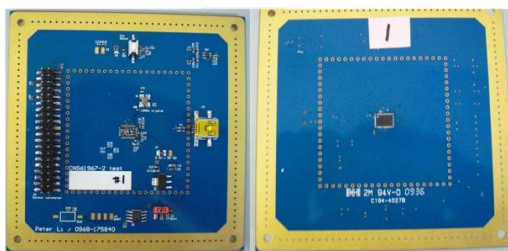


Fig. 54. Physical PCB with CAMERA module marked as NO.1.

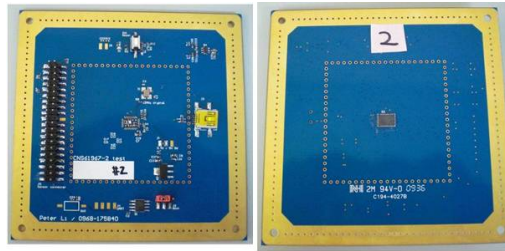


Fig. 55. Physical PCB with CAMERA module marked as NO.2.



Fig. 56. Physical camera connected to testing board.

There were two operation modes for camera module to be analyzed on EMI measurement. The first mode simulates the cellular phone activating the camera module for video communication. In the case of first mode, the camera is simply turned on for full function but does not execute the video file transferring from capturing camera to store on hard disc or memory card. However, the second mode simulates the cellular phone activating the camera module for video recording. In the case of the second mode, the camera is not only activated for full function but also execute the video file transferring from capturing camera to store on hard disc or memory card. The measured results for both operation modes are shown in Table 8 and 9 respectively. Compare the measured results for both operational mode, we can observe the occurring EMI phenomena during video file transferring from capturing camera to storage device. It can be used to find that if the more functions IC executes, would the severe EMI noise be generated or not.

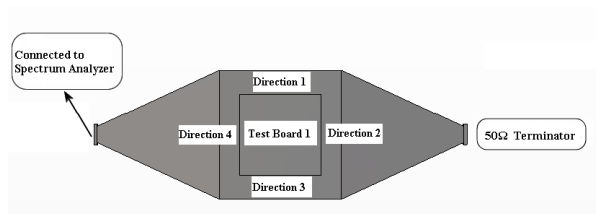


Fig. 57. Definition of 4 directions for TEM cell test orientation.

Direction Board	1	2	3	4
1	9.41 dBuV at 480 MHz	7.37 dBuV at 72 MHz	10.45 dBuV at 480 MHz	6.39 dBuV at 120 MHz
2	8.50 dBuV at 312 MHz	6.44 dBuV at 120 MHz	9.06 dBuV at 312 MHz	7.54 dBuV at 72 MHz

Table 8. Maximum EMI level with corresponding PCB orientation and frequency for video communications mode (mode 1).

Direction Board	1	2	3	4
1	10.63 dBuV at 480 MHz	6.61 dBuV at 720 MHz	11.15 dBuV at 480 MHz	6.39 dBuV at 120 MHz
2	11.14 dBuV at 72 MHz	7.74 dBuV at 960 MHz	9.25 dBuV at 480 MHz	7.32 dBuV at 815 MHz

Table 9. Maximum EMI level with corresponding PCB orientation and frequency for video file transfer mode (mode 2).

7.2.3 Test setup for near-field surface scanning[2,13]

The setup as shown in Figure 58 is to detect the EMI noise from LVDS cables, connectors, and driver ICs of the LCD panel control circuits. From the measured results, we can identify the locations of the significant EMI noise sources.

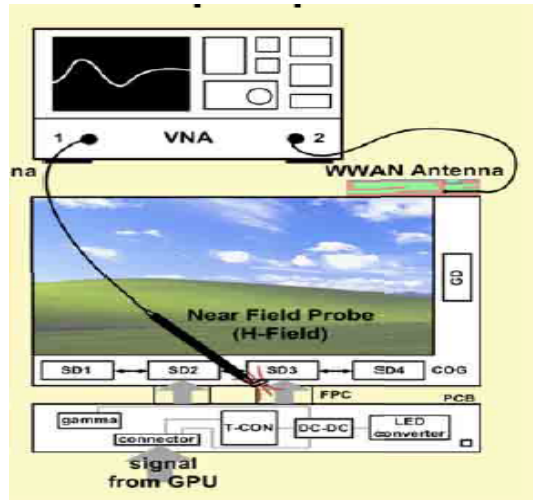


Fig. 58. Setup for LCD panel noise measurement.

7.2.4 Test setup for throughput

The setup of the throughput measurement for the analysis of communications performance in this study is shown in Figure 59. The throughput measurement system consists of WLAN AP (access point), device under test (DUT), attenuator, and Chariot software for data rate control.

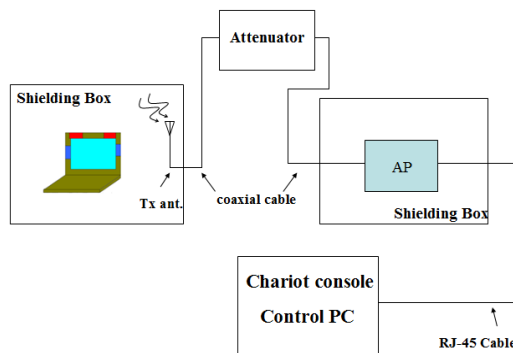


Fig. 59. Setup of throughput measurement system.

7.3 Analysis of measurement results[10-14]

7.3.1 Results of noise level measurement for different LCD panel

Variation of noise level for Camera module at different operation mode is shown in Figure 60. We can observe the significant variation of noise level in 2586~2600MHz frequency range when Camera is activated and operated at Record mode. Since the crystal oscillation

of Camera is 48 MHz, we can conclude that its 50th and 54th harmonics just fall at 2400MHz and 2592MHz, the most significant noise level frequencies, respectively. Therefore, the receiving sensitivity and thus the communications performance in 2.4 GHz band are degraded by the activation of Camera functions.

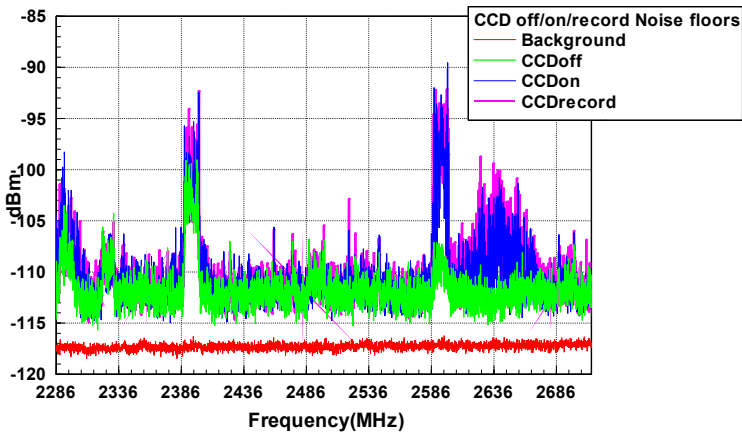
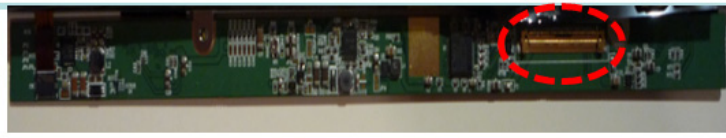


Fig. 60. Noise level for different Camera operation mode.

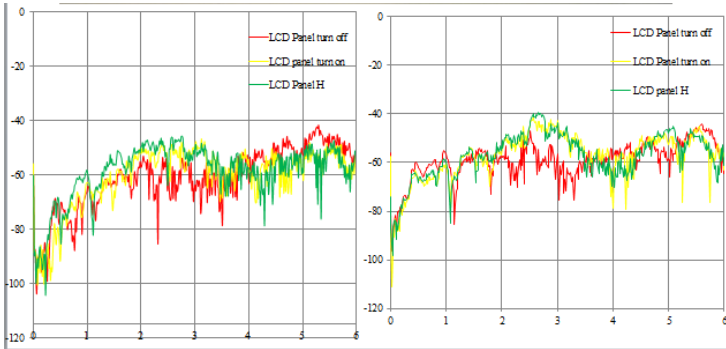
7.3.2 Results of surface-scanning measurement

From the results of noise level measurement, we first obtained the interference frequencies generated from LCD panel. We then used the magnetic near-field probe to observe the noise influence on wireless communication bands via antenna ports from LVDS cables, connectors, and driver ICs of the LCD panel control circuits when LCD panel was set to various operation modes, (such as off, standby, and key-in alphbat H pattern mode). Figures 61-63 show the change of transmission coefficients between antenna port and LVDS cables of the LCD panel control circuits. The measured result on left is for horizontal orientation of near-field probe placement, and the one on right is for vertical orientation. When magnetic probe is placed in vertical orientation, it is in parallel with routing traces of LVDS connector and thus results in higher sensitivity. We also observed that the coupling level are much higher for LCD panel operated in standby or key-in alphbat H pattern mode than shut off.

Figures 64 and 65 show the change of transmission coefficients between antenna port and driver IC of the LCD panel control circuits. The measured result on left is for horizontal orientation of near-field probe placement, and the one on right is for vertical orientation. We observed that the noises coupled to antenna port are much higher for LCD panel is turned on or displays H pattern mode than shut off, because the control IC is activated.



(a) Measurement position

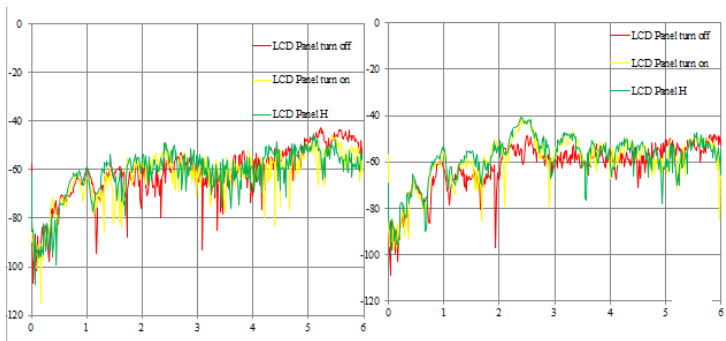


(b) Measured result of S21

Fig. 61. Measurement position (a) and (b) result of LCD Panel control circuit.



(a) Measurement position

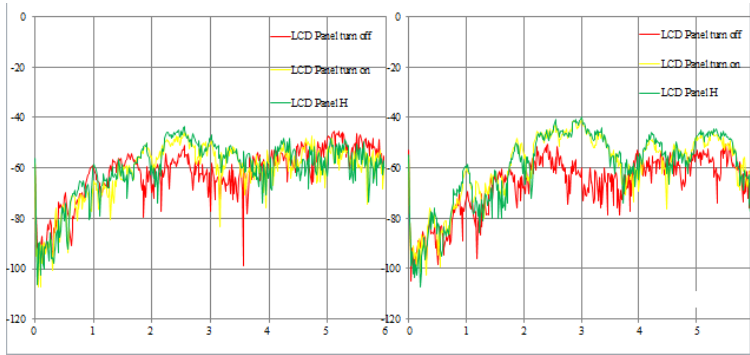


(b) Measured result of S21

Fig. 62. Measurement position (a) and (b) result of LCD Panel control IC.

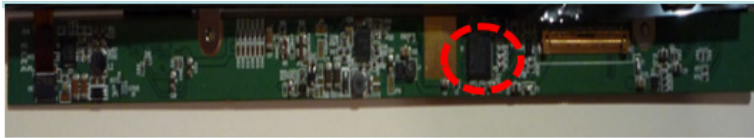


(a) Measurement position

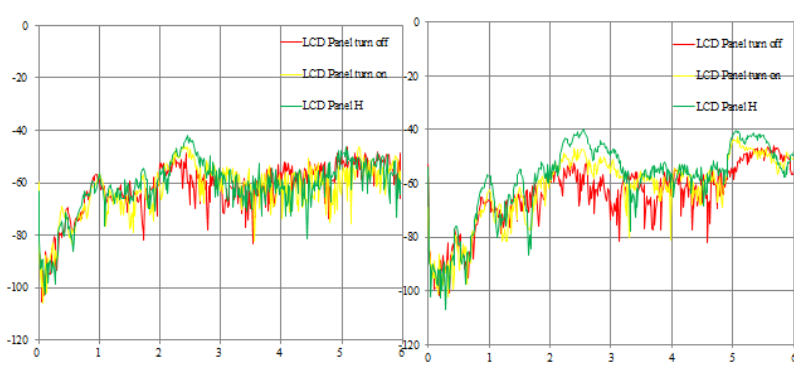


(b) Measured result of S21

Fig. 63. Measurement position (a) and (b) result of LCD Panel control circuit.

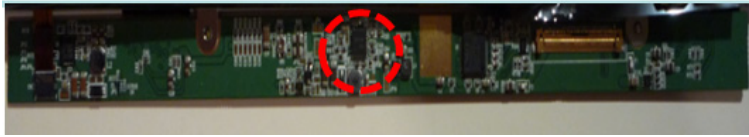


(a) Measurement position

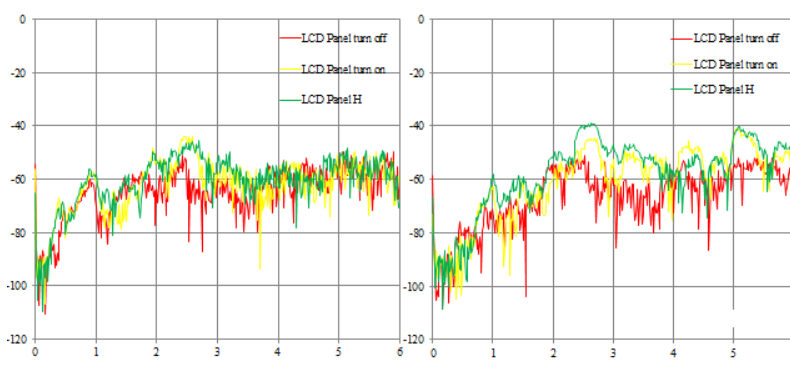


(b) Measured result of S21

Fig. 64. Measurement position (a) and (b) result of LCD Panel control circuit.



(a) Measurement position



(b) Measured result of S21

Fig. 65. Measurement position (a) and (b) result of LCD Panel control circuit.

7.4 Summary

Since the development of IC technologies advancing toward nm processing technology and higher operating frequencies in recent years, the systems of highly integrated high-speed digital circuits and multi-radio modules are now facing the challenge from performance degradation by more complicated electromagnetic noisy environment. With the development of the analyzing and measuring methodologies for this wireless platform noise problem and establishment of noise budget for digital component in the near future, we can provide the EMI coupling mechanism and noise level for each component to help system engineer analyze and design the EMC compliant wireless product in the first beginning as shown in Figure 66 and 67.

TP Noise Budget Fishbone Diagram

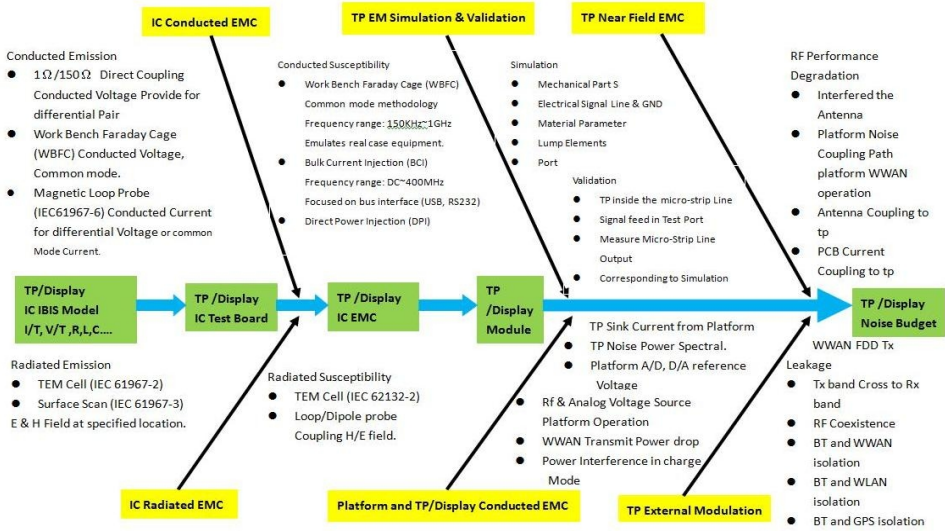


Fig. 66. Noise budget consideration for touch panel.

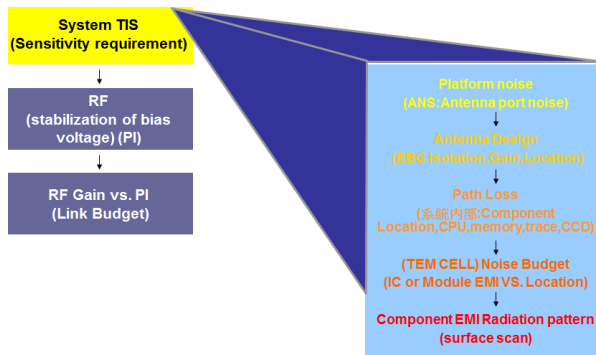


Fig. 67. System view for noise budget.

8. Acknowledgments

The author would like to thank the old friend and research partner Frank Tsai from TRC (Training and Research Company, Taiwan) for his inspiration, technical support and measurement assistance. The author would also like to thank the funding support from BSMI (Bureau of Standards, Metrology and Inspection) and NSC (National Science Council) Taiwan.

9. Reference

- IEC 61967-2: Integrated circuits - Measurement of electromagnetic emissions, 150 kHz to 1 GHz - Part 2: TEM cell method, International Electrotechnical Commission (IEC), Geneva, Switzerland Int. Std., July 2005.
- IEC 61967-3 Edition 1.0 (2005-06) Integrated circuits - Measurement of electromagnetic emissions, 150 kHz to 1 GHz - Part 3: Measurement of radiated emissions - Surface scan method.
- Test Plan for Mobile Station Over the Air Performance: Method of Measurement for Radiated RF Power and Receiver Performance. Ver. 3.1, CTIA - The Wireless Association, January 2011
- Han-Nien Lin, Ching-Hsien Lin, Tai-Jung Cheng, Min-Chih Liao, *Antenna Effect Analysis of Laptop Platform Noise on WLAN Performance*, PIERS 2009 in Beijing, Session 2A8, March 21-25, 2009
- Han-Nien Lin, Ming-Cheng Chang, Jia-Li Chang, Yung-Chi Tang, Jay-San Chen, *Influence Analysis of LCD Modules Noise on Performance of 802.11b*, 2011 APEMC in Korea, Poster I-6 P55, May 16-19, 2011
- Han-Nien Lin, Ching-Hsien Lin, Ming-Cheng Chang, Yu-Yang Shih, *Analysis of Platform Noise Effect on WWAN*, AMEMC 2010 in Beijing, TH-PM-A1-2: SS-13, April 12-16, 2010
- Nada Golmie, *Coexistence in Wireless Networks: Challenges and System-Level Solutions in the Unlicensed Bands*, Cambridge, 2006
- Han-Nien Lin, Ching-Hsien Lin, Chun-Chi Tang, and Ming-Cheng Chang, *Application of Periodic Structure on the Isolation and Suppression for Notebook Multi-antennas Coupling* PIERS 2010, Xi'an, Proceeding, p.160-164, March 22-26, 2010
- Han-Nien Lin, Chung-Wei Kuo, Jhih-Min Liao, Jian-Li Dong, *Design of TEM cell and high sensitive probe for EMI analysis of built-in Webcam module*, 2010 EDAPS in Singapore, THPMTS84, December 07-09, 2010
- Han-Nien Lin, Jing-Ting Cheng, Jian-Li Dong, Jay-San Chen, *Radiated EMI analysis for CMOS Camera module with TEM Cell and Far-field testing*, 2011 APEMC in Korea, Poster I-1 P47, May 16-19, 2011
- Han-Nien Lin, Ming-Feng Cheng, Han-Chang Hsieh, Jay-San Chen, *Design and characteristic analysis of TEM Cell for IC and module EMC testing*, 2011 APEMC in Korea, T-Tu3-5 P103, May 16-19, 2011
- IEC 61967-1: Integrated circuits - Measurement of electromagnetic emissions, 150 kHz to 1 GHz - Part 1: General conditions and definitions, International Electrotechnical Commission (IEC), Geneva, Switzerland Int. Std., March 2002.

Han-Nien Lin, Chung-Shun Chang, Gang-Wei Cao, Cheng-Chang Chen, Jay-San Chen, *Design of High Sensitivity Near-Field Probe and Application on IC EMI Detection*, 2011 APEMC in Korea, Poster I-4 P53, May 16-19, 2011

Han-Nien Lin, Tai-jung Cheng, Chih-Min Liao, *Radiated EMI Prediction and Mechanism Modeling from Measured Noise of Microcontroller*, AMEMC 2010 in Beijing, TH-PM-A1-4: SS-13, April 12-16, 2010

Part 3

Channel Estimation and Capacity

Indoor Channel Measurement for Wireless Communication

Hui Yu and Xi Chen
*Shanghai Jiao Tong University
China*

1. Introduction

In the past few years, Multiple Input Multiple Output (MIMO) system received lots of attentions, since it is capable in providing higher spectrum efficiency, as well as better transmission reliability. This improvement is brought by the multiple antennas at both sides of transmission. Since there are additional parallel sub-channels in spatial domain, system can not only increase the reliability by spatial diversity technology, but also provide higher data rate utilizing spatial multiplexing (see [1],[2], etc.). Orthogonal Frequency Division Multiplexing (OFDM) can also accommodate high data rate requirement by providing frequency multiplexing gain. To fully utilize the benefits of both technologies, the combination of the above two, MIMO-OFDM, has been employed in many wireless communication systems and protocols, such as WLAN [3] and LTE systems [4].

The introduction of MIMO-OFDM raises plenty challenges in channel estimation and measurement. Transmitted signals are reflected and scattered, resulting in a multipath spread in the received signals. Moreover, the transmitters, receivers, and reflecting or scattering objects are moving, which means that channels change rapidly over time [5]. Inter-Channel interference may also bring a destructive effect in transmission, which should be cancelled by accurate channel measurement or estimation.

As an important application of MIMO-OFDM technology, WLAN is suitable in providing high data rate service in hotspot area, such as office buildings, airports, libraries, stations, hospitals and restaurants. This means lots of MIMO-OFDM applications (such as WLAN) take place in indoor situations, where both transmitters and receivers are surrounded by mobile and static scatters. Different from outdoor scenario, there are some unique characteristics of indoor scenario. More scatters result in larger influence by multipath effect; higher density of users and overlap between different access points bring larger interference. Because of these differences in channel parameters, it is crucial to obtain a better understanding of indoor channels. Statistics such as delay spread, Doppler spread, angle spread and path loss must be estimated by detail channel measurement. This requirement rises the interests in indoor channel measurement in the past few years.

Channel measurement or estimation schemes can be divided into two major categories, blind and non-blind. Blind channel estimation method requires large data and can only exploit the statistical behavior of channel, hence, suffers a lot in fast fading channels.

On contrast, non-blind method utilizes pre-determined information in both transmitters and receivers, to measure the channel. One of the most frequently used methods is the pilot/data aided channel measurement/estimation. The method can be further divided into two sub-methods considering the resources occupied by pilots with each resource unit. In the first sub-method, pilots occupy the whole resource unit, for example, an OFDM training symbol. This type of pilot arrangement is largely used for pure purpose of channel measurement without the request of communication. In real-time communication, it is only suitable for slow channel variation and for burst type data transmission schemes. In the second sub-method, pilots only occupy part of the resource unit; the other part of the unit is allocated to data. This pilot arrangement is capable to provide real-time communication which takes place in time and frequency varying channels. However, a linear interpolation or higher order polynomial fitting should be applied to recover the whole channel, which will certainly cause some errors.

Recently, there are plenty works considering indoor channel measurement and estimations, with non-blind pilot/data aided method, each of which focuses on different aspects of the issue. In [6], authors introduced a detail design of a MIMO channel sounder. In their measurement process, they used a PN sequence as the probing signal (pilot), which occupy the whole frequency and time resources. Their measurement took place in Seoul railway station, and they used the results to illustrate the characteristics of delay and Doppler spread of indoor channel. In [7], authors provide a PC-FPGA design in solving a similar problem for WLAN system. Instead of occupying the whole channel resource, the PN sequences only insert in certain parts of the resource unit. In this way, channel measurement and transmission can be carried out simultaneously. The effect of polarized antennas has also been considered in [8]. Wireless situations include non-line-of-sight, propagation along the corridor and propagation over a metallic ceiling.

Other important applications include several scenarios such as: near-ground indoor channel aiming military or emergency usage [9], and indoor channel model for wireless sensor network and internet of things [10]. Pilot signal design is flexible. Instead of a PN sequence, other pseudolite signals are available, too. Also, the kinds of carrier signals are variable, such as an OFDM signal [7] or a GPS-based signal [11].

In addition, some rough estimations of channel parameters are provided. Most of these works based on the assumption that indoor channels follow the Ricean distribution. The most important parameter of Ricean distribution for indoor channels is the K-factor, which represents the ratio between the average power of deterministic and random components of the channel. In [12], a two-moment method of the Ricean K-factor is provided theoretically. Experimental results can be found in [13], which gives an application for the two-moment estimation of the Ricean K-factor in wideband indoor channels at 3.7 GHz. Although the Ricean distribution can only provide an unclear view of the channel, it is convenient and low-complexity estimation; thus can be applied in scenarios that require only partial channel state information.

2. Measurement schemes

As is introduced before, channel measurement schemes can be divided into two major categories, blind and non-blind. Since blind measurement and estimation is much less reliable, we only discuss non-blind pilot/data aided schemes. Two of the most frequently used schemes are discussed: measurement based on PN sequence which occupies a whole resource unit, and measurement based on OFDM pilot who occupies only part of an OFDM symbol.

2.1 Indoor channel model

The low-pass time-variant channel impulse response (CIR) is denoted as $h(\tau;t)$, which represents the response of the channel at time t due to an impulse applied at time $t - \tau$. Then transmission can be expressed as:

$$y(t, \tau) = \sum_{n=0}^{N_{\text{CIR}}-1} h(n, \tau)x(t - n, \tau) + w(t, \tau)_k \quad (1)$$

Where x is the transmit signal, and y is the transmit signal, N_{CIR} is the duration of the CIR, w is the noise sequence.

By taking the Fourier transform of $h(\tau;t)$, we can obtain the time-variant channel transfer function

$$H(f;t) = \int_{-\infty}^{+\infty} h(\tau;t)e^{-j2\pi f\tau} d\tau \quad (2)$$

On the assumption that the scattering of the channel at two different delays is uncorrelated, the autocorrelation function of $h(\tau;t)$ can be defined as:

$$\frac{1}{2}E\left[h^*(\tau_1;t)h(\tau_2;t + \Delta t)\right] = \varphi_h(\tau_1;\Delta t)\delta(\tau_1 - \tau_2) \quad (3)$$

If we let $\Delta t = 0$, the resulting autocorrelation function $\varphi_h(\tau) \equiv \varphi_h(\tau;0)$ is called delay power profile of the channel. The range of values of τ over which $\varphi_h(\tau)$ is essentially nonzero is called the multipath spread of the channel. Then the scattering function of the channel is defined as:

$$S(\tau;\lambda) = \int_{-\infty}^{+\infty} \varphi_h(\tau;\Delta t)e^{-j2\pi\lambda\Delta t} d\Delta t \quad (4)$$

By taking the integration of $S(\tau;\lambda)$, we obtain the Doppler power spectrum of the channel as:

$$S(\lambda) = \int_{-\infty}^{+\infty} S(\tau;\lambda)d\tau \quad (5)$$

The range of values of λ over which $S(\lambda)$ is essentially nonzero is called the Doppler spread of the channel.

Indoor channel conditions are much more complex than that of outdoor channel. There are plenty kinds of scattering figures, such as walls, tables, etc. People indoor can also act as scattering figures, and the movements of cell phones caused by this bring about worse channel conditions. As a result, angle of arrival, multipath spread and scattering factor of indoor channels are different from those of outdoor channels. Consequently, measurement schemes should be redesign carefully according to the above distinguish characteristics, so as to fulfill the needs of indoor channel measurements.

It is worth noticing that channels of phone calls made in indoor conditions are mixtures of both indoor and outdoor channels. Large scale fading, small scale fading and shadow fading should be equally considered in such channels. Each of these factors can cause a large channel capacity decrease.

2.2 Channel measurement using PN sequence

In this scheme, Pseudo-Noise (PN) Sequence is used as a probing signal. To authors' best knowledge, the most widely used binary PN sequence is the Maximum-Length-Shift-Register (MLSR) sequence. The length of MLSR sequence is $N_{PN} = 2^m - 1$ bits. And one of the possible generators of this sequence is an m-stage linear feedback shift register (see [14]). As a result, MLSR sequence is periodic with period n. Within each period, there are 2^{m-1} ones and 2^{m-1} zeros.

One of the most important characteristics of the periodic PN sequence is its sharp auto-correlation. Consider an PN sequence x_k , we have :

$$\sum_{k=0}^{N_{PN}-1} x_{k+m}x_k = \begin{cases} N_{PN}, & m = 0, \pm N_{PN}, \pm 2N_{PN}, \dots \\ 0, & \text{Others} \end{cases} \quad (6)$$

If $N_{PN} \gg 1$, it is approximate that:

$$\frac{1}{N_{PN}} \sum_{k=0}^{N_{PN}-1} x_{k+m}x_k \approx \sum_{i=-\infty}^{+\infty} \delta_{m-iN_{PN}} \quad (7)$$

We can represent the received signal y_k as the convolution of transmitted signal x_k and the CIR h_k as:

$$y_k = \sum_{n=0}^{N_{CIR}-1} h_n x_{k-n} + w_k \quad (8)$$

The transmitter use PN sequence as the transmitting data x_k . If the generators of PN sequence in both transmitter and receiver are synchronous, at each time slot, receiver is aware of the transmitting PN sequence. Hence, receiver can obtain the CIR by doing a cross-correlating between y_k and x_k :

$$\begin{aligned}
\hat{h}_k &= N_{\text{PN}} \sum_{m=0}^{N_{\text{PN}}-1} y_{m+k} x_m \\
&= \frac{1}{N_{\text{PN}}} \sum_{m=0}^{N_{\text{PN}}-1} \left(\sum_{n=0}^{N_{\text{CIR}}-1} h_n x_{m+k-n} + w_{m+k} \right) x_m \\
&= \sum_{n=0}^{N_{\text{CIR}}-1} h_n \frac{1}{N_{\text{PN}}} \sum_{m=0}^{N_{\text{PN}}-1} x_{m+k-n} x_m + \frac{1}{N_{\text{PN}}} \sum_{k=0}^{N_{\text{PN}}-1} w_{m+k} x_m \\
&= \sum_{n=0}^{N_{\text{CIR}}-1} h_n \sum_{i=-\infty}^{+\infty} \delta_{k-n-iN_{\text{PN}}} + \frac{1}{N_{\text{PN}}} \sum_{k=0}^{N_{\text{PN}}-1} w_{m+k} x_m \\
&= \sum_{i=-\infty}^{+\infty} h_{k-iN_{\text{PN}}} + w'_m
\end{aligned} \tag{9}$$

It can be seen that the result of cross-correlating is a summation of noise w' and the periodic extension of h_k . If $N_{\text{CIR}} \leq N_{\text{PN}}$, a period of \hat{h}_k can be used as the estimate of h_k [7].

There are two drawbacks in PN sequence measurement. First, the PN sequence takes up a great amount of time and frequency resources (most of the time, all the transmitting resources). This results in a great loss of throughput, as well as a channel mismatch caused by the delay between channel measurement and data transmission. Second, the accuracy of synchronizers in both sides should be in a high level, which raises the cost of equipments for channel measurement.

2.3 Channel measurement using OFDM pilot

Unlike measurement based on PN sequence, the pre-determined data of measurement based on OFDM pilot only occupied a relatively small percentage of time and frequency resources. Channel segments located on the pilots can be measured directly and correctly. However, other channel segments can only be estimated indirectly with some interpolations. Although the pilot-based measurement can only give an imperfect result, it provides a possibility of transmitting data and measuring channel simultaneously. This characteristic is crucial for real systems with limited feedback, such as WLAN, WiMax, LTE and LTE-A systems. There are two major problems in this scheme: how to design the pilot pattern and how to interpolate with discrete channel value on both time and frequency domain.

2.3.1 OFDM pilot pattern

OFDM pilots may be inserted in both time and frequency resources. A pilot pattern refers to the places where pilots are inserted in every OFDM symbol. An effective pilot pattern needs to be designed carefully in both frequency and time domains.

In frequency domain, according to the Nyquist sampling theorem, if we want to capture the variation of channel, the frequency space D_f between pilots should be small enough:

$$D_f \leq \frac{1}{\tau_{\text{max}} \Delta f} \tag{10}$$

where τ_{\max} represents the maximum delay of channel, and Δf denotes the frequency space between subcarriers.

If the above condition is not satisfied, the channel estimation cannot sample the accurate channel, since channel fading in frequency cannot be detected fast enough.

In time domain, spacing between pilots inserted in the same frequency is determined by the coherence time. In order to capture the variation of channel, the time space of pilots D_t must be correlated with coherence time, and must be small enough:

$$D_t \leq \frac{1}{2f_{\text{doppler}}T_f} \quad (11)$$

Where f_{doppler} represents the maximum Doppler spread of channel, and T_f denotes the duration of each OFDM symbol.

However, it is worth pointing out that pilot allocation is a tradeoff of many factors in real systems. These include channel estimation accuracy, spectral efficiency of the system, wasted energy in unnecessary pilot symbols, and fading process not being sampled sufficiently. As a result, there is no optimal pilot pattern for all the channels, as fading process are varied.

Another important element of pilot pattern is the power allocation. Power is equally allocated to pilots and data symbols in regular cases. However, the accuracy of channel estimation can increase greatly with the power allocated to pilots. Considering the total power constraint, this will result in a decline of data symbols' SNR. Hence, another tradeoff between channel estimation and transmission capacity has to be evaluated.

There is a lack of pilots at the edges of OFDM symbols, which leads to a much higher error rate in such places. One simple but less effective way is to place more pilots at the edge. The drawback of this scheme is obvious: it reduces the frequency efficiency of systems. Some other scheme utilizes periodic behavior of the Fourier Transform, and establishes certain correlations between the beginning and the end of OFDM symbols. Simulations are reported to verify the effectiveness of this scheme.

2.3.2 Measurements on pilots

Channel segments locating on the pilots can be measured directly by some well-known algorithm, such as Least Square (LS) and Linear Minimum Mean Square Error (LMMSE).

Both LS and LMMSE algorithm aim to minimize a parameter: $\min\{(y_k - x_k h_k)^H (y_k - x_k h_k)\}$.

Using LS algorithm, we have:

$$\tilde{h}_k^{\text{LS}} = x_k^{-1} y_k = h_k + x_k^{-1} w_k = \left[\frac{y_k^1}{x_k^1}, \frac{y_k^2}{x_k^2}, \dots, \frac{y_k^{M_k}}{x_k^{M_k}} \right]^T \quad (12)$$

where M_k is the length of transmit and receive signal.

Using LMMSE algorithm, we have:

$$\tilde{h}_k^{\text{LMMSE}} = \mathbf{R}_{H_k H_k} (\mathbf{R}_{H_k H_k} + \frac{\beta}{\text{SNR}} \mathbf{I})^{-1} \tilde{H}_k^{\text{LS}} \quad (13)$$

where $\mathbf{R}_{H_k H_k} = E\{H_k H_k^*\}$ autocorrelation of channel. and $\beta = E\{|x_k|^2\} E\{1/|x_k|^2\}$.

Comparing both LS algorithm and LMMSE algorithm, we can draw the following conclusion. LS algorithm is much easier to realize and apply. LS algorithm only needs one discrete divider to estimate the channels on all the pilots. Moreover, statistical information about channel and noise are not necessary while employing LS algorithm. However, the accuracy of LS algorithm is very sensitive to the noise and synchronization errors.

On the other hand, LMMSE algorithm can be seen as a filtering on the estimation result on LS algorithm. And this filtering is based on the MMSE criteria. It can be proven that, under the same MSE conditions, estimations results of LMMSE algorithm provide a larger gain than that of LS algorithm. Drawbacks of LMMSE algorithm are also obvious. Due to the inversion operation of matrices, complexity of LMMSE algorithm is relatively high. Furthermore, LMMSE algorithm requires knowing the statistical information of channel and noise in prior, which is unrealistic in applications.

While taking the errors of estimated $\mathbf{R}_{H_k H_k}$ into account, the MSE and SNR gains provided by LMMSE algorithm are marginally larger than that of LS algorithm. Considering the tradeoff between complexity and performance, LS algorithm may be a better solution than LMMSE algorithm.

2.3.3 Interpolations

By applying LS or LMMSE algorithm, one can easily obtain the CIR of piloted channel segments. However, we still have no idea of channel segments not occupied with pilots. In order to obtain CIR of these segments, interpolations should be used.

The best interpolation algorithm may be 2-D Wiener filtering, since it can cancel noise as much as possible, in both frequency and time domain. However, in order to decide the weights of Wiener filtering, channel statistics must be known. Moreover, the complexity brought by matrix inversion increases gigantically with data in pilots. All of the above prevent the usage of 2-D Wiener filtering in real systems.

Some achievable interpolations include: cascade 1-D Wiener filtering, Lagrange interpolation, and transform domain interpolation.

Cascade 1-D Wiener filtering tries to realize a 2-D Wiener filtering by cascading two 1-D Wiener filtering. The complexity of cascade 1-D Wiener filtering is much less than 2-D Wiener filtering, while the performance only decreases a little bit. There two kinds of cascade 1-D Wiener filtering, in respects of interpolation order of frequency and time domain.

In frequency domain, the major interpolations include Lagrange interpolation, LMMSE interpolation, transform domain interpolation, DFT based interpolation, and low-pass filtering interpolation. In time domain, the available schemes are LMMSE interpolation, Lagrange interpolation.

2.3.3.1 Two dimensions LMMSE interpolations

Assume that the estimated channel matrix of all the piloted subcarriers is $\tilde{H}_{n',i'}, \forall \{n',i'\} \in P$, where P is the set of positions of all the pilots, n' is the index in frequency domain, i' is the index in time domain. We have:

$$\tilde{H}_{n',i'} = \frac{Y_{n',i'}}{X_{n',i'}} = H_{n',i'} + \frac{N_{n',i'}}{X_{n',i'}}, \forall \{n',i'\} \in P \quad (14)$$

Then, estimate the channel parameters by two dimensions interpolation filtering:

$$\hat{H}_{n,i} = \sum_{\{n',i'\} \in \Gamma_{n,i}} w_{n',i',n,i} \tilde{H}_{n',i'}, \Gamma_{n,i} \subseteq P \quad (15)$$

Where $w_{n',i',n,i}$ are the weights of interpolation filter, $\hat{H}_{n,i}$ is the estimated channel, $\Gamma_{n,i}$ is the number of used pilots. The number of weights in the filter is $N_{tap} = \|\Gamma_{n,i}\|$.

Applying MSE criteria, the MSE $J_{n,i}$ in subcarrier (n,i) is:

$$\begin{aligned} \varepsilon_{n,i} &= H_{n,i} - \hat{H}_{n,i} \\ J_{n,i} &= E\{\|\varepsilon_{n,i}\|^2\} \end{aligned} \quad (16)$$

A filter following the MMSE criteria is a two dimensions Wiener filter. According to the orthogonality of such a filter, we have:

$$E\{\varepsilon_{n,i} \tilde{H}_{n'',i''}^*\} = 0, \forall \{n'',i''\} \in \Gamma_{n,i} \quad (17)$$

Where (n'',i'') represents the positions of pilots while channel estimation is conducted.

The Wiener-Hopf equation can be derived from (17), which follows:

$$E\{H_{n,i} \tilde{H}_{n'',i''}^*\} = \sum_{\{n',i'\} \in \Gamma_{n,i}} w_{n',i',n,i} E\{\tilde{H}_{n',i'} \tilde{H}_{n'',i''}^*\}, \forall \{n'',i''\} \in \Gamma_{n,i} \quad (18)$$

Define the correlation function as:

$$\theta_{n-n'',i-i''} = E\{H_{n,i} \tilde{H}_{n'',i''}^*\} \quad (19)$$

If the mean value of noise is zero, and is independent to the transmission data, the correlation can be further expressed as:

$$\theta_{n-n'',i-i''} = E\{\tilde{H}_{n,i} \tilde{H}_{n'',i''}^*\} \quad (20)$$

Define the right part of (18) as the autocorrelation of channels at pilots:

$$\varphi_{n'-n'',i'-i''} = E\{\tilde{H}_{n',i'} \tilde{H}_{n'',i''}^*\} \quad (21)$$

According to (14), it follows:

$$\begin{aligned}\varphi_{n'-n'',i'-i''} &= \theta_{n'-n'',i'-i''} + \frac{\sigma^2}{E\{|X_{n',i'}|^2\}} \delta_{n'-n'',i'-i''} \\ &= \theta_{n'-n'',i'-i''} + \frac{1}{\text{SNR}} \delta_{n'-n'',i'-i''}\end{aligned}\quad (22)$$

Where $E\{|X_{n',i'}|^2\}$ is the average power of pilot symbols.

Equation (22) shows that the correlation function depends on the distance between the position of channel being estimated (n,i) and the positions of pilots employed in the estimation process (n'',i'') . And the autocorrelation function depends on the distances between pilots.

Substituting (20) and (21) into (18), we have:

$$\boldsymbol{\theta}_{n,i}^T = \mathbf{w}_{n,i}^T \boldsymbol{\Phi} \quad (23)$$

Where $\boldsymbol{\Phi}$ is the $N_{\text{tap}} \times N_{\text{tap}}$ autocorrelation matrix, $\boldsymbol{\theta}_{n,i}$ is the correlation vector with length N_{tap} , and $\mathbf{w}_{n,i}$ is the parameter vector of filter with length N_{tap} . Therefore, the parameter of the 2-D Wiener filter is:

$$\mathbf{w}_{n,i}^T = \boldsymbol{\theta}_{n,i}^T \boldsymbol{\Phi}^{-1} \quad (24)$$

The full estimated channel matrix can be expressed as:

$$\hat{H}_{n,i} = \mathbf{w}_{n,i}^T \tilde{\mathbf{H}}_{n,i} \quad (25)$$

In conclusion, the design of such a filter is to decide its parameters $\mathbf{w}_{n,i}$, which can be derived by the correlation function $\theta_{n'-n'',i'-i''}$ and average SNR. Unfortunately, the correlation of channel cannot be achieved accurately in real systems. Hence, approximate models with typical multipath delay profile $\rho(\tau)$ and Doppler power spectrum $S_{f_D}(f_D)$ are employed.

2-D Wiener filtering is the optimal interpolation scheme in respect of MMSE; it can obtain optimal performance theoretically. However, its requirement of prior statistic knowledge of channel matrix, as well as the complexity of matrix inversion, makes it almost impossible to apply in real systems.

2.3.3.2 Interpolations in frequency domain

A 2-D interpolation can be form by a cascade of two 1-D interpolations. By appropriate designs, such a cascade can largely reduce the complexity while maintaining a good performance. The order of interpolation should be taken into consideration. We propose to interpolate firstly in frequency domain, then to conduct the time domain interpolation. The reason is that frequency-time scheme can start once the piloted OFDM symbols receive, while time-frequency scheme has to wait for the arrivals of all the symbols in one frame or

subframe before the interpolation begins. As a result, frequency-time scheme can decrease the delay of channel measurement, hence provides more effective interpolation.

Interpolations in frequency domain aim to obtain all the channel function respond (CFR) \hat{H}_C , according to the measured CFR \tilde{H}_p in each piloted subcarrier:

$$\hat{H}_C = \mathbf{w} \cdot \tilde{H}_p \quad (26)$$

where \mathbf{w} is the frequency domain interpolation matrix, and is the channel vector of piloted subcarriers obtained by (12) or (13). We have:

$$\tilde{H}_p = \mathbf{X}_{pp}^{-1} Y_p = H_p + \mathbf{X}_{pp}^{-1} N_p = H_p + \tilde{n} \quad (27)$$

2.3.3.2.1 LMMSE interpolation

Projecting equation (23) on frequency domain, we obtain the optimal interpolation parameter vector \mathbf{w} as:

$$\mathbf{R}_{\tilde{H}_p, \tilde{H}_p} \mathbf{w}^* = \mathbf{R}_{\tilde{H}_p, H_C} \quad (28)$$

where $\mathbf{R}_{\tilde{H}_p, \tilde{H}_p} = E\{\tilde{H}_p \tilde{H}_p^*\}$ represents the autocorrelation matrix of estimation channel segments with pilots \tilde{H}_p , $\mathbf{R}_{\tilde{H}_p, H_C} = E\{\tilde{H}_p H_C^*\}$ denotes the correlation matrix of estimation channel segments with pilots \tilde{H}_p and the real channel being interpolated H_C .

If $\mathbf{R}_{\tilde{H}_p, \tilde{H}_p}$ is invertible, \mathbf{w} can be expressed as:

$$\mathbf{w} = \mathbf{R}_{H_C, \tilde{H}_p} \mathbf{R}_{\tilde{H}_p, \tilde{H}_p}^{-1} \quad (29)$$

Combining (12) and (27), it follows:

$$\begin{aligned} \mathbf{R}_{\tilde{n}\tilde{n}} &= E\{\tilde{n}\tilde{n}^*\} = E\{\mathbf{X}_p^{-1} N_p \cdot N_p^* (\mathbf{X}_p^{-1})^*\} \\ &= \sigma_n^2 E\{\mathbf{X}_p^{-1} \cdot \mathbf{I} \cdot (\mathbf{X}_p^{-1})^*\} = \sigma_n^2 E\{(\mathbf{X}_p^* \mathbf{X}_p)^{-1}\} \\ &= \mathbf{I}_{N_p} E\{1 / |X_k^p|^2\} \sigma_n^2 \end{aligned} \quad (30)$$

where X_k^p is the constellation point of piloted channel, and σ_n^2 is the power of noise.

Substituting (12) into (29), we have:

$$\mathbf{w} = \mathbf{R}_{H_C, H_p} (\mathbf{R}_{H_p, H_p} + \mathbf{I}_{N_p} E\{1 / |X_k^p|^2\} \sigma_n^2)^{-1} \quad (31)$$

where \mathbf{R}_{H_C, H_p} and \mathbf{R}_{H_p, H_p} are the ideal correlation and autocorrelation matrices. Further representing the CFRs of channels with their CIRs, it is obvious that $H_C = \mathbf{F}_{CL} h_L$ and $H_p = \mathbf{F}_{PL} h_L$ (where h_L is the discrete CIR, \mathbf{F}_{CL} and \mathbf{F}_{PL} are the corresponding DFT transform matrices). This converts (31) as followed:

$$\mathbf{w} = \mathbf{F}_{\text{CL}} \mathbf{R}_{h_L, h_L} \mathbf{F}_{\text{PL}}^* (\mathbf{F}_{\text{PL}} \mathbf{R}_{h_L, h_L} \mathbf{F}_{\text{PL}}^* + \mathbf{I}_{N_p} E\{1/|X_k^p|^2\} \sigma_n^2)^{-1} \quad (32)$$

Applying Parseval Theorem, which certifies that the power in frequency domain equals that of time domain, we draw the following conclusion:

$$\begin{aligned} \mathbf{w} &= \mathbf{F}_{\text{CL}} \bar{\mathbf{R}}_{h_L, h_L} \mathbf{F}_{\text{PL}}^* (\mathbf{F}_{\text{PL}} \bar{\mathbf{R}}_{h_L, h_L} \mathbf{F}_{\text{PL}}^* + \mathbf{I}_{N_p} \frac{E\{1/|X_k^p|^2\} \sigma_n^2}{\text{trace}(\bar{\mathbf{R}}_{h_L, h_L})})^{-1} \\ &= \mathbf{F}_{\text{CL}} \bar{\mathbf{R}}_{h_L, h_L} \mathbf{F}_{\text{PL}}^* (\mathbf{F}_{\text{PL}} \bar{\mathbf{R}}_{h_L, h_L} \mathbf{F}_{\text{PL}}^* + \mathbf{I}_{N_p} \frac{E\{1/|X_k^p|^2\} \sigma_n^2}{E(\mathbf{H}_C^* \mathbf{H}_C)})^{-1} \\ &= \mathbf{F}_{\text{CL}} \bar{\mathbf{R}}_{h_L, h_L} \mathbf{F}_{\text{PL}}^* (\mathbf{F}_{\text{PL}} \bar{\mathbf{R}}_{h_L, h_L} \mathbf{F}_{\text{PL}}^* + \mathbf{I}_{N_p} \frac{E\{1/|X_k^p|^2\} E\{|X_k^p|^2\} \sigma_n^2}{E\{|X_k^p|^2\} E(\mathbf{H}_C^* \mathbf{H}_C)})^{-1} \end{aligned} \quad (33)$$

In special cases where QPSK modulation and equal power allocation are adopted, the interpolation is:

$$\mathbf{w} = \mathbf{F}_{\text{CL}} \bar{\mathbf{R}}_{h_L, h_L} \mathbf{F}_{\text{PL}}^* (\mathbf{F}_{\text{PL}} \bar{\mathbf{R}}_{h_L, h_L} \mathbf{F}_{\text{PL}}^* + \mathbf{I}_{N_p} \frac{1}{\text{SNR}})^{-1} \quad (34)$$

$$\text{where } \text{SNR} = \frac{P_r}{\sigma_n^2} = \frac{E\{|X_k|^2\} E\{\mathbf{H}_C^* \mathbf{H}_C\}}{\sigma_n^2} = \frac{E\{|X_k^p|^2\} E\{\mathbf{H}_C^* \mathbf{H}_C\}}{\sigma_n^2}.$$

2.3.3.2 Lagrange interpolation

Lagrange interpolation is widely used and easy to implement. It is a group of interpolation algorithms, including linear interpolation, Gaussian interpolation, cubic interpolation, etc. Lagrange interpolation is suitable for both frequency and time domain interpolation. However, the disadvantage of Lagrange interpolation is obvious. It is unable to cancel the noise.

Linear interpolation in frequency domain utilizes each pair of adjacent piloted channel segments to obtain the channel function within them. The interpolation process follows the following equation:

$$\hat{H}(l+d) = \left(1 - \frac{d}{D}\right) \tilde{H}_p(l) + \frac{d}{D} \tilde{H}_p(l+D), 1 \leq d \leq D-1 \quad (35)$$

where D is the interval between two adjacent pilots, $\tilde{H}_p(l)$ and $\tilde{H}_p(l+D)$ are the corresponding channel estimation results.

Gaussian interpolation in frequency domain employs the measured channels of three adjacent pilots, which can be represented as followed:

$$\hat{H}(x) = \begin{cases} \tilde{H}_p(l_{j-1}) \frac{x-l_j}{l_{j-1}-l_j} \frac{x-l_{j+1}}{l_{j-1}-l_{j+1}} + \tilde{H}_p(l_j) \frac{x-l_{j-1}}{l_j-l_{j-1}} \frac{x-l_{j+1}}{l_j-l_{j+1}} \\ + \tilde{H}_p(l_{j+1}) \frac{x-l_{j-1}}{l_{j+1}-l_{j-1}} \frac{x-l_j}{l_{j+1}-l_j}, (l_j \neq K_{\max}) \\ \tilde{H}_p(l_{j-2}) \frac{x-l_{j-1}}{l_{j-2}-l_{j-1}} \frac{x-l_j}{l_{j-2}-l_j} + \tilde{H}_p(l_{j-1}) \frac{x-l_{j-2}}{l_{j-1}-l_{j-2}} \frac{x-l_j}{l_{j-1}-l_j} \\ + \tilde{H}_p(l_j) \frac{x-l_{j-2}}{l_j-l_{j-2}} \frac{x-l_{j-1}}{l_j-l_{j-1}}, (l_j = K_{\max}) \end{cases} \quad (36)$$

Where K_{\max} denotes the maximum position of pilots, $\tilde{H}_p(l_{j-1})$, $\tilde{H}_p(l_j)$, and $\tilde{H}_p(l_{j+1})$ are the channel measurement results of three used pilots, and $l_{j-1} < x < l_j$.

Cubic interpolation further increase the number of used pilots onto four. The expression for interpolation is showed as followed:

$$\begin{aligned} \hat{H}(x) = & \tilde{H}_p(l_{j-2}) \frac{x-l_{j-1}}{l_{j-2}-l_{j-1}} \frac{x-l_j}{l_{j-2}-l_j} \frac{x-l_{j+1}}{l_{j-2}-l_{j+1}} + \tilde{H}_p(l_{j-1}) \frac{x-l_{j-2}}{l_{j-1}-l_{j-2}} \frac{x-l_j}{l_{j-1}-l_j} \frac{x-l_{j+1}}{l_{j-1}-l_{j+1}} \\ & + \tilde{H}_p(l_j) \frac{x-l_{j-2}}{l_j-l_{j-2}} \frac{x-l_{j-1}}{l_j-l_{j-1}} \frac{x-l_{j+1}}{l_j-l_{j+1}} + \tilde{H}_p(l_{j+1}) \frac{x-l_{j-2}}{l_{j+1}-l_{j-2}} \frac{x-l_{j-1}}{l_{j+1}-l_{j-1}} \frac{x-l_j}{l_{j+1}-l_j} \\ & (l_j \neq K_{\max}, l_j \neq d, l_{j-1} < x < l_j) \end{aligned} \quad (37)$$

All the above schemes are simple to apply in real systems. However, they all introduce certain level of noise, and yield effect of error floor. This can be eliminated by employing a low pass filter after interpolation.

2.3.3.2.3 Transform domain interpolation

The basic idea of transform domain interpolation is to reduce the complexity by conducting interpolation in various transform domains. The most widely used kind of transform domain interpolation is based on DFT.

The fundamental principle of DFT based interpolation is: in process of signal processing, zeroizing in time domain is equivalent to interpolating in frequency domain. If a sequence of N points has $N - N_p$ zeros in the end, its Fourier transform values at the positions of multiples of N_p are the same as the counterparts of Fourier transform of sequence formed by the former N_p points. On the other hand, the Fourier transform values not at the positions of multiples of N_p consist of linear combinations of the Fourier transform of truncated sequence.

After receiving the information of piloted channels, DFT based interpolation conducts IFFT of length N_p . Then the interpolation zeroizes the transformed sequence into a N pointed sequence. Finally, transform the sequence into frequency domain by a N points FFT.

The zeroizing can be conducted as followed:

$$\hat{h}_N(i) = \begin{cases} \tilde{h}_{N_p}(i) & 0 \leq i < N_p/2 \\ 0 & N_p/2 \leq i < N - N_p/2 \\ \tilde{h}_{N_p}(i - (N - N_p)) & N - N_p/2 \leq i \leq N - 1 \end{cases} \quad (38)$$

Considering the influences of noise and inter-channel interference, introducing a low pass filtering before zeroizing can effectively increase the accuracy of channel measurement. The block diagram of the above procedure is showed in Fig. 1.

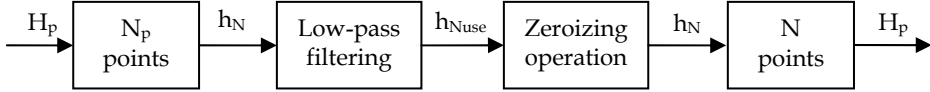


Fig. 1. DFT and low-pass filter based interpolation.

Since the most complex calculations in this kind of interpolation are FFT and IFFT, the complexity of DFT based interpolation is much lower than others. However, the performance will drop largely, if the multipath spread is not a multiple of sampling period.

2.3.3.3 Interpolations in time domain

After frequency interpolation is done, we can launch the interpolation in time domain. The interpolation can also be expressed as a interpolation matrix as followed:

$$\hat{H}_{Ct} = \mathbf{w}_t \cdot \tilde{H}_{pt} \quad (39)$$

Where \mathbf{w}_t is the time domain interpolation matrix, \tilde{H}_{pt} denotes the CFR of channel segments on the pilots, \hat{H}_{Ct} represents the CFR for all the channel segments. By assuming $\tilde{H}_{pt} = H_{pt} + \tilde{n}$, we consider the impact of AWGN in the interpolation.

2.3.3.3.1 LMMSE interpolation

LMMES interpolation in time domain is similar to that in frequency domain. The only difference is that we project equation (23) into time domain, so that it follows:

$$\mathbf{w}_t = \mathbf{R}_{H_{Ct}\tilde{H}_{pt}} \mathbf{R}_{\tilde{H}_{pt}\tilde{H}_{pt}}^{-1} \quad (40)$$

where $\mathbf{R}_{\tilde{H}_{pt}\tilde{H}_{pt}} = E\{\tilde{H}_{pt}\tilde{H}_{pt}^*\}$ represents the autocorrelation matrix of estimation channel segments with pilots \tilde{H}_{pt} , $\mathbf{R}_{H_{Ct}\tilde{H}_{pt}} = E\{H_{Ct}\tilde{H}_{pt}^*\}$ denotes the correlation matrix of estimation channel segments with pilots \tilde{H}_{pt} and the real channel being interpolated H_{Ct} .

Following the steps of derivation for frequency LMMSE interpolation, the interpolation matrix of time LMMSE interpolation can be simplified as below:

$$\mathbf{w}_t = \mathbf{R}_{H_{Ct}H_{pt}} \left(\mathbf{R}_{H_{pt}H_{pt}} + \frac{\beta}{SNR} \mathbf{I}_{N_{pt}} \right)^{-1} \quad (41)$$

We consider a special case where QPSK modulation and average power allocation are used. Then the correlation between adjacent pilots i and i'' is:

$$R_{H_{C,H_{pt}}}(i, i'') = J_0(2\pi f_{D_{\max}}(i - i'')T_s) \quad (42)$$

Where $f_{D_{\max}}$ is the maximum Doppler spread, T_s is the interval between two symbols, and $J_0(x)$ represents the first zero-order Bessel function.

However, the previous frequency interpolation and the corresponding filtering cause changes in the noise power of each subcarrier. Therefore, the signal-to-noise-ratio of each channel segment no longer equals the original value. An adjustment was proposed based on the MSE after the frequency interpolation.

Assume that channel frequency response after frequency LMMSE interpolation is:

$$H_f^n = [H_0^{(n)} + \tilde{w}_0^{(n)}, \dots, H_i^{(n)} + \tilde{w}_i^{(n)}, \dots, H_{L-1}^{(n)} + \tilde{w}_{L-1}^{(n)}]^T \quad (43)$$

Where L is the number of OFDM symbols in each frame, $H_i^{(n)}$ represents the channel frequency response of i th OFDM symbol in n th subcarrier, $\tilde{w}_i^{(n)}$ denotes the corresponding residual noise. Hence, variance of $\tilde{w}_i^{(n)}$ is equivalent to the MSE after interpolation.

$$MSE_{LMMSE, n} = [\mathbf{R}_{H_C H_C} - \mathbf{R}_{H_C H_{pt}} (\mathbf{R}_{H_{pt} H_{pt}} + \mathbf{I}_{N_p} E\{1/|X_k^p|^2\} \sigma_n^2)^{-1} \mathbf{R}_{H_C H_{pt}}^H]_{nn}, n = 0, 1, 2, \dots, N-1 \quad (44)$$

Where N is the number of subcarriers waiting for measured.

As a result, time domain LMMSE interpolation should be optimized according to the variance of noise. The interpolation matrix is then:

$$\mathbf{w}_t = \mathbf{R}_{H_C H_{pt}} (\mathbf{R}_{H_{pt} H_{pt}} + \text{diag}(\sigma_0^2, \dots, \sigma_i^2, \dots, \sigma_{L-1}^2))^{-1} \quad (45)$$

Where σ_i^2 represents the variance of residual noise $\tilde{w}_i^{(n)}$.

To reduce the complexity, channel segments in the same OFDM symbol can utilize their average noise variance in the interpolation, and an approximate interpolation matrix can be expressed as followed:

$$\mathbf{w}_t = \mathbf{R}_{H_C H_{pt}} (\mathbf{R}_{H_{pt} H_{pt}} + \frac{1}{N} \sum_{i=0}^{N-1} \sigma_0^2(i) \mathbf{I}_L)^{-1} \quad (46)$$

2.3.3.3.2 Lagrange interpolations

Lagrange interpolations in time domain are almost the same as those in frequency domain. The only difference is channel response of piloted channel segments. One can refer to previous sections for details.

3. Applications

Here we show some useful and easily implemented examples to illustrate the indoor channel measurement. Measurement based on PN sequence, as well as OFDM pilot, will be discussed.

3.1 Channel measurement system using PN sequence

In this section, we present an example of 2x2 MIMO channel measurement, utilizing a semi-sequential scheme. This semi-sequential scheme uses parallel receivers and a switch at the transmitter (Fig. 2). When a measurement process starts, the probing signal is firstly transmitted from the 1st transmit antenna (TX1) and the receive signal is sampled from 1st and 2nd receive antennas (RX1 and RX2) simultaneously. Thus the channel from TX1 to RX1 and RX2 can be measured at the same time. Then, a similar process is used to measure the channel from TX2 to RX1 and RX2. Strictly speaking, the semi-sequential MIMO channel sounder measures Single Input Multiple Output (SIMO) channels directly. The MIMO channel is obtained by combine the two SIMO channels, on the assumption that the MIMO channel doesn't change significantly in a single round of sequential measurement.

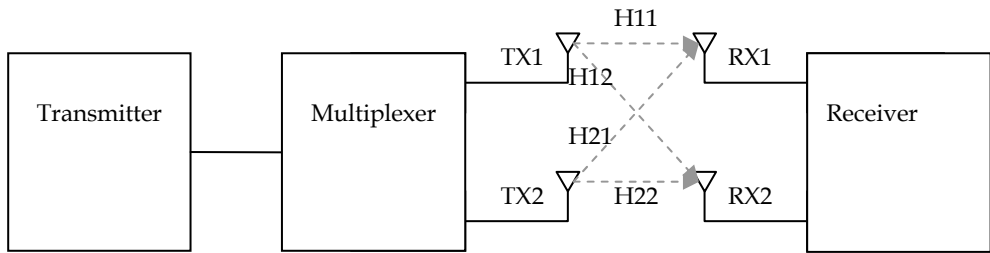


Fig. 2. Semi-sequential scheme of MIMO channel sounder [7].

Each SIMO channel can be measured by the algorithm introduced in Section 2.2 Then a combination of two SIMO channel construct the whole MIMO channel.

3.1.1 Baseband signal processing algorithm

3.1.1.1 System parameters

The link-level block diagram of the sliding correlation channel sounder for Single Input Single Output (SISO) channel is shown in Fig. 3. In the semi-sequential MIMO channel sounder (or SIMO channel sounder); there should be two parallel receivers.

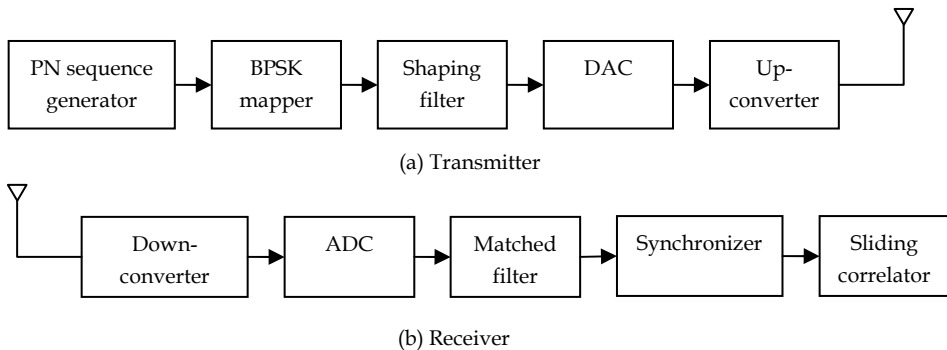


Fig. 3. Channel Sounder for SISO.

Here we only focus on the indoor wireless MIMO channel for WLAN like devices. The measurement system supports 20/40MHz bandwidth by suitable RFIC [16][17]. The system parameters of baseband are listed in Table 1.

Name	Symbol	Value
Sampling frequency	f_s	60MHz
Symbol rate of PN sequence	R_{symp}	20M Symbol/s
Period of PN sequence, express in units of T_{symp}	N_{PN}	127
Length of CIR, express in units of T_{symp}	N_{CIR}	127
Sampling interval	T_s	$1/f_s$
Interpolated sample interval	T_i	$T_{\text{symp}}/2$
Symbol interval of PN sequence	T_{symp}	$1/R_{\text{symp}}$

Table 1. System Parameters.

3.1.1.2 Symbol timing synchronizing algorithm

Symbol timing synchronizer is a critical module of the digital receiver design of the channel sounder based on sliding correlation channel measurement. Gardner’s symbol timing recovery method is used in this system [18][19]. The structure of the symbol timing synchronizer is shown in Fig. 4. All the processing of this synchronizer is done in digital domain. No interaction between analog and digital part of the system is needed. This synchronizer is capable of compensating sampling phase and frequency offset and is independent of carrier phase [20].

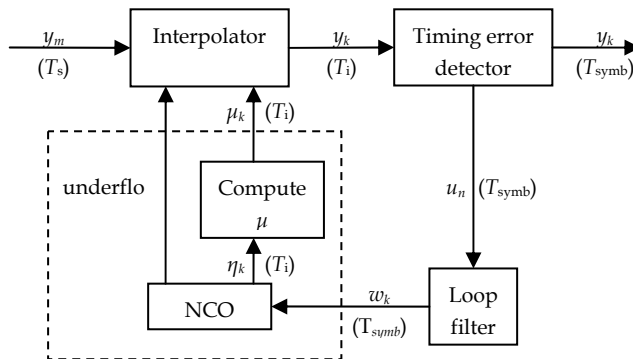


Fig. 4. Symbol timing synchronizer.

The sampled data y_m , which is filtered by matching filter, is then feed into the interpolator to compute the desired sampled strobe y_k . This is done by digital interpolation, controlled by NCO (Numerical Controlled Oscillator) and fraction interval μ_k . Ideally, the period of the NCO is $T_i = T_{\text{symp}} / K$, where K is an integer. The loop consisting of timing error detector, loop filter and NCO function just like a DPLL, where u_n , w_k and η_k represents timing error signal, NCO control word and NCO register content respectively.

In this design, the DTTL algorithm [20] is used to compute the timing error signal. This choice specifies $T_i = T_{\text{symp}} / 2$. In order to avoid up-sampling in the interpolator, T_s should be smaller than T_i . Thus, the sampling frequency f_s should be larger than two times the symbol rate R_{symp} . The interpolator performs linear interpolation, which is easy to implement. The loop filter is a proportional-plus-integral structure.

3.1.2 Hardware design

The whole measurement system hardware consists of several modules: antennas module, multi-channel AD/DA module, baseband processing FPGA board, USB access module and a computer Graphical User Interface (GUI) module. The architecture of hardware is showed in Fig. 5. The RF board is based on MAX2829, which can support MIMO operation. We choose the FFP board (IAF GmbH) as FPGA prototyping platform for baseband signal processing, RF control and interface to PC. The interface between the FFP board and the PC is an USB2.0 port. The GUI program runs on the computer for user to control the channel measurement functions and demonstrate the real time test results. Because the most effort is on the development of FPGA, here we focus on the design of baseband transceiver.

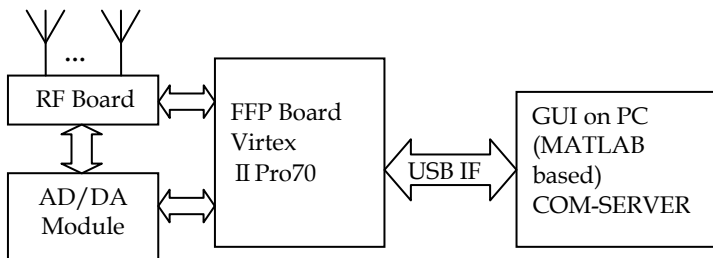


Fig. 5. Hardware architecture of the channel measurement system.

The baseband transceiver module performs the baseband signal processing of a 2x2 MIMO channel sounder. This module generates the baseband probing signal, i.e. a BPSK modulated PN sequence, and delivers the CIR extracted from the received signal to the upper-level module.

The block diagram of this module is shown in Fig. 6. The module can be divided into three parts. The first part is the transmitter, which includes signal generator and transmit multiplexer. The second part is the receiver, which includes receive buffers, signal processor, and data buffer. The last part is the control logic of the module.

The functions of the sub-modules are as follows:

- Signal generator generates the baseband probing signal, i.e. a BPSK modulated PN sequence.
- Transmit multiplexer distributes probing signal to different TX antennas.
- Receive buffers save the received signal from RX antennas.
- Receive multiplexer feed the signal stored in receive buffers into the signal processor in a sequential order.

- Signal processor performs the signal processing, i.e. filtering, symbol timing, and sliding correlation, to extract the CIR from received signal.

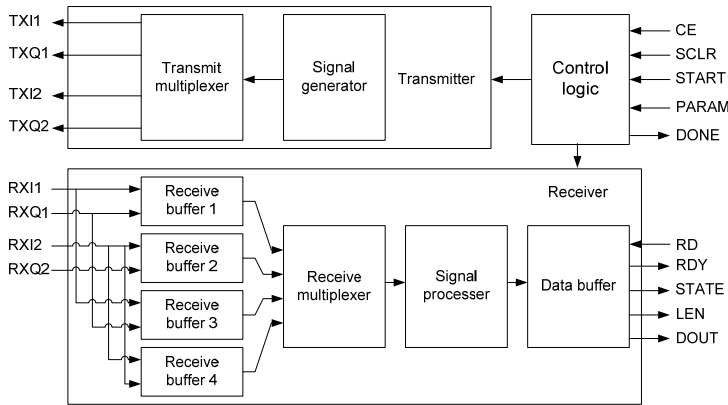


Fig. 6. Baseband transceivers.

3.1.3 GUI

To provide a user friendly human interface, we design a MATLAB based GUI. The real time data stream is accessible from the specific application software through a function call of the COM-Server from IAF. The software can provide several channel information from the original measured data. These channel information include channel impulse response, channel transfer function, delay power profile, scattering function and Doppler power spectrum. Fig. 7 is an indoor channel test result for example.

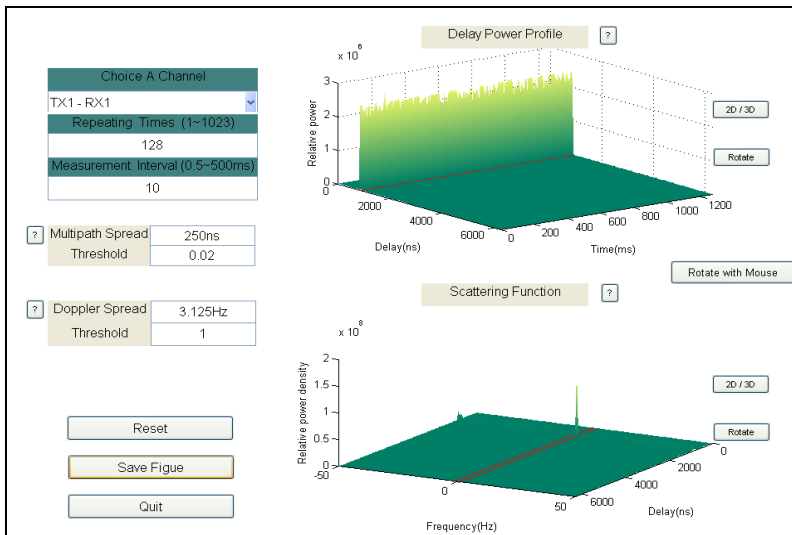


Fig. 7. Measurement result on GUI.

3.2 Channel measurement system using OFDM pilot

In the section, we present an example of OFDM-pilot-based MIMO channel measurement scheme. The measurement is conducted under LTE system. We utilize the reference signal (pilot) to carry out the 4x4 MIMO channel measurement. In this example, the measurement of channel occupied by pilots is LS algorithm, with the purpose of decreasing complexity.

One transmitter sends data according to the LTE agreement, so that each transmitted subframe consists of pilots and useful data. Receiver breaks down each subframe to obtain pilot segments and data segments, respectively. Such measurement equipment can implement the channel measurement without interrupting communications.

A cascade 1-D filtering is used for the 2-D interpolation. This cascade 1-D filtering firstly interpolates the channel in frequency domain with LMMSE interpolation, and then finishes the whole interpolation with a linear time domain interpolation.

There are several reasons why we choose a cascade of frequency LMMSE interpolation and time linear interpolation. LMMSE interpolation certainly has the best MSE performance among all the interpolation schemes. However, the complexity of LMMSE interpolation is much larger than that of linear interpolation. Thus, a tradeoff between performance and complexity has to be made. In frequency domain, LMMSE can provide a large performance increase. When achieving the same BLER or throughput performance, LMMSE interpolation can save about 2 dB SNR. On the other hand, the performance improvement in time domain by applying LMMSE interpolation is marginal, saving only 0.25 dB average. Considering the above, the usage of a cascade of frequency LMMSE interpolation and time linear interpolation is reasonable.

A block diagram of this example is showed in Fig. 8.

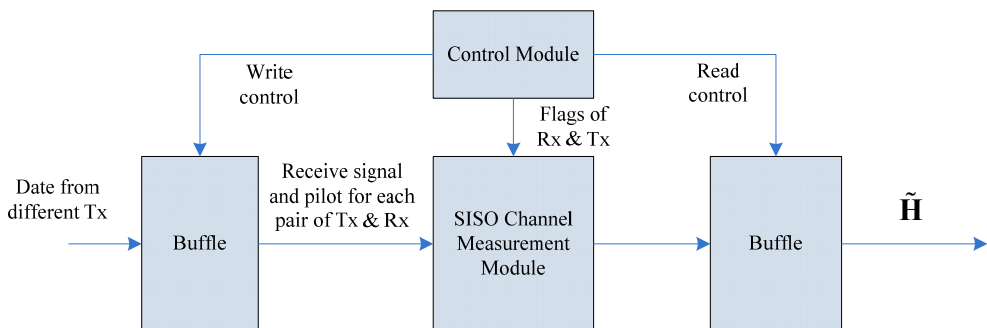


Fig. 8. A block diagram of LTE MIMO channel measurement system.

In addition, the pilot pattern of LTE system with 4 antennas can be seen in Fig. 9.

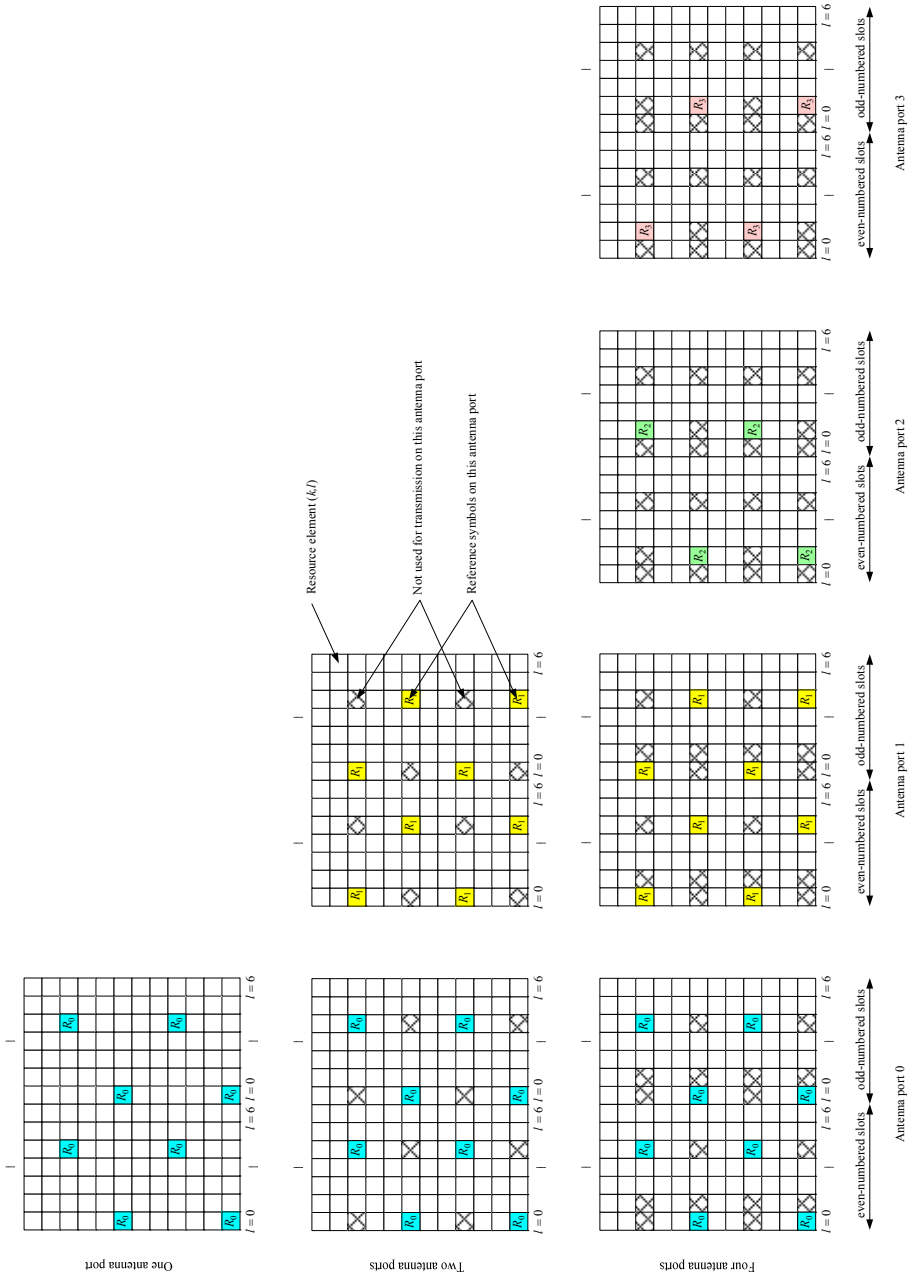


Fig. 9. Pilot pattern of LTE system [4].

3.2.1 Measurements on LTE pilots

Since LTE symbols are typical OFDM symbols, classic measurement schemes such as LS and LMMSE can be applied directly on LTE pilots. A matrix form of (12) for measured channel by LS algorithm is showed as followed:

$$\tilde{H}_p^{LS} = \mathbf{X}_{pp}^{-1} \mathbf{Y}_p = \mathbf{H}_p + \mathbf{X}_{pp}^{-1} \mathbf{N}_p = \left[\frac{Y_p^1}{X_p^1}, \frac{Y_p^2}{X_p^2}, \dots, \frac{Y_p^{M_p}}{X_p^{M_p}} \right]^T \quad (47)$$

Since LMMSE algorithm is very vulnerable to the speed of mobile stations, the benefit brought by LMMSE will be negligible while comparing to its large processing burden. Furthermore, LS algorithm can be helpful in cancelling the effect of noise brought by LMMSE interpolation, so that the overhead of LMMSE interpolation can be reduced.

3.2.2 Design of frequency domain interpolation

Considering equation (34), there are three major challenges in realizing frequency domain LMMSE interpolation: estimating autocorrelation matrix $\bar{\mathbf{R}}_{h_l, h_l}$, determining Signal-to-Noise-Ratio in receiver and obtaining the inversion of matrix.

3.2.2.1 Autocorrelation matrix $\bar{\mathbf{R}}_{h_l, h_l}$

Since the real channels are time-varying, it is impossible to obtain the accurate autocorrelation of channels. The most widely used scheme is to estimate the approximate autocorrelation through some known channel models. It is well-known that two of the most important factors in wireless channel models are multipath spread and Doppler spread. While in the frequency domain, we mainly consider the influence of multipath spread, and propose a simple but useful construction scheme for wireless channels as followed.

The CIR of such a multipath channel is showed as followed:

$$h(\tau) = \sum_{l=0}^{N_L-1} h_l \delta(\tau - \tau_l) \quad (48)$$

Where τ_l and h_l are the delay and amplitude of the l^{th} path. N_L denotes the max number of taps. δ represents the impulse function.

Define $L = \{0, 1, \dots, N_L - 1\}$. Define $\mathbf{h} \in \mathbb{C}^{N \times 1}$, $h_l = 0, \forall l \in \{N_L \dots N - 1\}$ as the multipath amplitude vector, N as subcarriers in each OFDM symbol.

Within digital baseband, we assume that the discrete delay as:

$$\tau_l = \frac{lT_s}{N}, l \in L \quad (49)$$

Where is the length of an OFDM symbol.

Further assume that power σ_l^2 of independent Rayleigh-distributed tap h_l is fading exponentially with time constant τ_d :

$$\sigma_l^2 \sim e^{-\frac{l}{\tau_d}}, l \in L \tag{50}$$

Then the normalized CIR autocorrelation can be expressed as:

$$\bar{\mathbf{R}}_{h_L, h_L} = \frac{\mathbf{R}_{h_L, h_L}}{\|\mathbf{R}_{h_L, h_L}\|} = \frac{\text{diag}(\sigma_0^2 \cdots \sigma_{N_L-1}^2)}{\|\mathbf{R}_{h_L, h_L}\|} = \frac{\text{diag}(\sigma_0^2 \cdots \sigma_{N_L-1}^2)}{\sum_i \sigma_i^2} \tag{51}$$

Base on the above derivation, we need to determine N_L and τ_d to obtain $\bar{\mathbf{R}}_{h_L, h_L}$. The number of available taps N_L can be same as the length of cyclic prefix (CP), with the purpose of simplification. Yet such a simplification is reasonable, since the multipath spread is less than the length of CP in most of the time. The multipath spread can be estimated with real-time scheme, so as to refine the channel model, as well as the autocorrelation $\bar{\mathbf{R}}_{h_L, h_L}$.

One of the possible schemes to estimate the multipath spread is provided as followed:

Step 1. Measure the channel matrix of piloted segments $\tilde{\mathbf{H}}_{p, \wedge}$ in a symbol, with LS algorithm. Take a N_p points IFFT to obtain the rough CIR \hat{h}_L , and set $N_L^{\max} = N_p$ as the max length of multipath spread.

Step 2. Define a parameter \hat{h}_{pow}^s as:

$$\hat{h}_{\text{pow}}^s = \left| \hat{h}_s \right|^2 \quad s = 1, 2, \dots, N_L^{\max} \tag{52}$$

Where \hat{h}_{pow}^s denotes the square of amplitude for the s -th element in \hat{h}_L .

Then obtain a decision object K_s as followed;

$$K_s = \frac{\left(\sum_{j=s-9}^s \hat{h}_{\text{pow}}^j \right) / (2 \times 10)}{\left(\sum_{k=s+1}^{N_L^{\max}} \hat{h}_{\text{pow}}^k \right) / (2 \times (N_L^{\max} - s))} \tag{53}$$

Step 3. Find a value of s by the following procedure:

Decrease the value of s from $N_L^{\max} - 15$ to 1 with a step of 5. Take the first value of s that satisfies $K_s > 2.55$ as the estimate multipath spread. One can refer to [15] for the reason of choosing 2.55 as the threshold.

After determining the multipath spread, one can obtain $\bar{\mathbf{R}}_{h_L, h_L}$ by following previous derivation.

3.2.2.2 Signal-to-noise-ratio

SNR value may be measured or estimated in other blocks of receiver. If it is not, the following estimation scheme can be applied.

Denote $\bar{\mathbf{R}}_{h_L} = \bar{\mathbf{R}}_{h_L, h_L}^{1/2}$, $\mathbb{F}_{PL} = \mathbf{F}_{PL} \bar{\mathbf{R}}_{h_L}$. Do a singular value decomposition on \mathbb{F}_{PL} , so that $\mathbb{F}_{PL} = \mathbf{USV}^*$. Project estimated channel matrix $\tilde{\mathbf{H}}_p$ and real channel matrix \mathbf{H}_p as

$U^* \tilde{H}_p$ and $U^* H_p$. The element in the project of real channel $U^* H_p$ tends to zero when the singular value of \mathbb{F}_{PL} is zero. But things are different in $U^* \tilde{H}_p$. Since we have $U^* \tilde{H}_p = U^* (H_p + X_{pp}^{-1} N_p) = U^* H_p + U^* X_{pp}^{-1} N_p$, it is clear that when the last elements of $U^* H_p$ are zeros, the corresponding elements of $U^* \tilde{H}_p$ reflect the impact of noise. As a result, we can estimate the noise power by these elements.

Let $s = \{N_p - N_s \cdots N_p - 1\}, 1 \leq N_s < N_p$ be the range of index, N_p denotes the number of pilots, N_s represents the number of zero singular value in \mathbb{F}_{PL} . Then the estimated noise power is $\tilde{p}_n = \frac{1}{N_s} \|U_{.s}^* \tilde{H}_p\|_2^2$, and the corresponding signal power is $\tilde{p}_s = \frac{1}{N_p} \|\tilde{H}_p\|_2^2 - \tilde{p}_n$.

Assume that the SNR is constant within adjacent k pilots, then an average SNR can be obtain as followed:

$$\overline{SNR} = \frac{\sum_{i=1}^k \tilde{p}_{n_i}}{\sum_{i=1}^k \tilde{p}_{s_i}} \quad (54)$$

Since the last element in $U_{.s}^* \tilde{H}_p$ rarely contains signal information, it is the most suitable one for SNR estimation. Therefore, we can simplify the process by setting $N_s = 1$.

3.2.2.3 Inversion of matrix

It is clear from equation (34) that in order to obtain the interpolation matrix \mathbf{w} , a N_p order matrix inversion operation must be conducted. The overhead will be very large. Fortunately, instead of the entire matrix, we only need several discrete $\bar{\mathbf{R}}_{h_l, h_l}$ matrices. Therefore, if we apply discrete average SNR in equation (34), the parameter of interpolation matrix \mathbf{w} will be discrete. We can pre-design the discrete range of \mathbf{w} , and save it in a table. Then the real-time calculation is simplified as a looking up in a table, according to the measured $\bar{\mathbf{R}}_{h_l, h_l}$ and SNR.

Specifically, we can adapt a look-up table which cuts the SNR range into several intervals. Each SNR interval combines with a corresponding multipath spread $\hat{\tau}$. Each of such pairs jointly determines a pre-designed \mathbf{w} . With this scheme, the complexity of matrix inversion in real-time process is converted to the design of look-up table. Since the look-up table is generated off-line, real-time calculation burden for LMMSE interpolation is largely reduced.

3.2.3 Design of time domain interpolation

According to LTE standardization, each transmission time interval (TTI) is of length 1ms, which is the exact length of a subframe. Consequently, mobile stations process date in units of subframe. When time domain interpolation is conducting, there are at most four pilots in each subframe. As a result, the reference of time domain interpolation of LTE system is at most four estimated channel segments. Two of the most widely used schemes in time domain interpolation are LMMSE interpolation and linear interpolation. The detailed procedures of these two interpolations are presented in previous sections, so we only provide some simulation results to illustrate the advantages and disadvantages of each scheme.

The following simulation considers a unban macro scenario, in which the bandwidth is 10MHz, center frequency is 2GHz and noise is AWGN. Fig. 10 shows the MSE performances of both LMMSE and linear interpolations under different MS speeds.

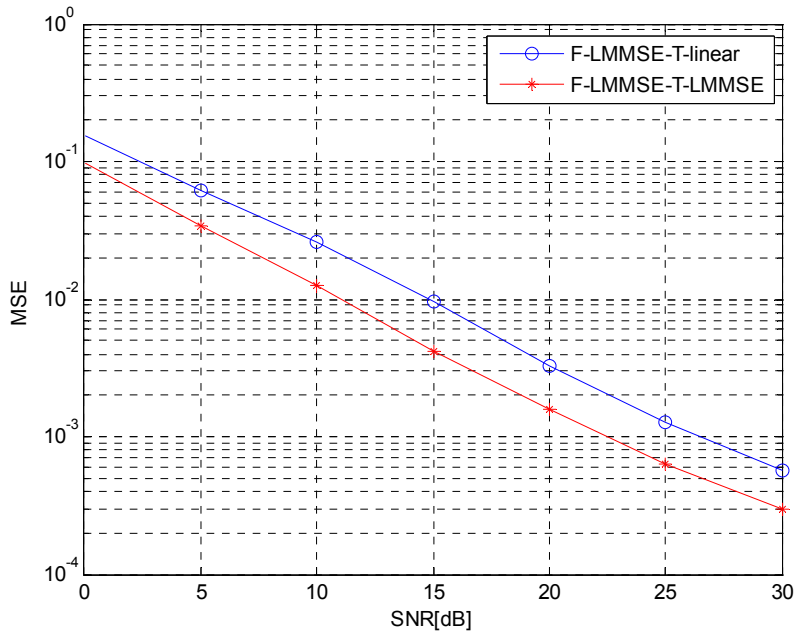


Fig. 10. (a) MSE performances of LMMSE and linear interpolations under MS speed 1m/s.

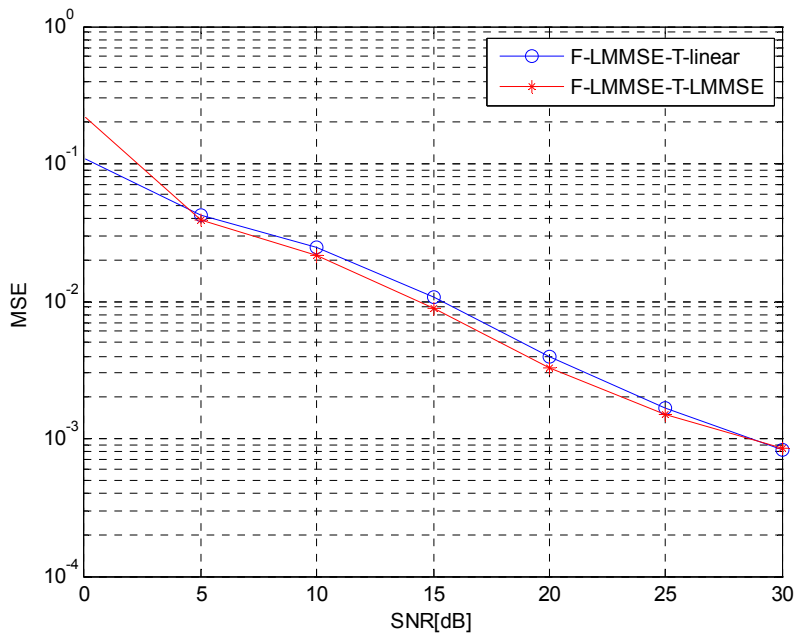


Fig. 10. (b) MSE performances of LMMSE and linear interpolations under MS speed 30m/s.

The following conclusions can be inferred from the simulation results.

When the speed of MS is small, correspondingly small Doppler spread, LMMSE interpolation can save 4 dB SNR while achieving the same MSE performance of linear interpolation. However, when the speed (as well as the Doppler spread) of MS increases to a relatively high level, performances of LMMSE and linear schemes become very close. This means that the large overhead spent on LMMSE outputs marginal gains on the performance. When the errors of Doppler spread estimations are taken into account, the MSE performance of LMMSE scheme may even be worse than that of linear interpolation. Consequently, after considering the tradeoff between performance and complexity, we propose to use a simple linear interpolation in time domain.

4. Reference

- [1] I. E. Telatar, "Capacity of Multi-Antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, Nov./Dec. 1999.
- [2] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, Achievable Rates, and Sum-Rate Capacity of Gaussian MIMO Broadcast Channels," *IEEE Transactions on Information Theory*, vol. 49, No. 10, pp. 2658–2668, Oct. 2003.
- [3] I. W. Group, "IEEE 802.11 Wireless Local Area Networks," May 2001
- [4] 3GPP TS 36.211: "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation".
- [5] M.K. Ozdemir and H. Arslan, "Channel Estimation for Wireless OFDM Systems," *IEEE Communications Surveys & Tutorials*, vol. 9, No. 2, pp. 18–48, 2007.
- [6] Myung-Don Kim, Heon Kook Kwon, Bum Soo Park, Jae Joon Park, and Hyun Kyu Chung, "Wideband MIMO Channel Measurements in Indoor Hotspot Scenario at 3.705GHz," *International Conference on Signal Processing and Communication Systems*, 2010.
- [7] Hui Yu, Ruikai Zhang, Xi Chen, Wentao Song, and Hailong Wang, "Design of an Indoor Channel Measurement System," *International Wireless Communications and Mobile Computing Conference*, 2010.
- [8] Jose-Maria, Molina-García-Pardo, José-Víctor Rodríguez, and Leandro Juan-Llácer, "Polarized Indoor MIMO Channel Measurements at 2.45 GHz," *IEEE Transactions on Antennas and Propagation*, vol.56, no. 12, Dec., 2008.
- [9] David W. Matolak, and Qian Zhang, "5 GHz Near-Ground Indoor Channel Measurements and Models," *IEEE Radio and Wireless Symposium*, 2009.
- [10] Ye Wang, Wenjun Lu and Hongbo Zhu, "Experimental Study on Indoor Channel Model for Wireless Sensor Networks and Internet of Things," *IEEE International Conference on Communication Technology*, 2010.
- [11] Alexandru Rusu- Casandra, Ion Marghescu and Elena Simona Lohan, "Estimators of the indoor channel for GPS-based pseudolite signal," *International Symposium on Electronics and Telecommunications*, 2009.
- [12] L. J. Greenstein, D. G. Michelson, and V. Erceg, "Moment-Method Estimation of the Ricean K-Factor," *IEEE Communications Letters*, vol. 3, No. 6, pp. 175-176, 1999.
- [13] Jae-Joon Park, Myung-Don Kim, and Hyun-Kyu Chung, "Characteristics of Ricean K-factor in Wideband Indoor Channels at 3.7 GHz," *International Conference on Signal Processing and Communication Systems*, 2010.

- [14] John G. Proakis, "Digital Communications", McGraw-Hill Companies, Inc and publishing house of electronics industry, China. ,fourth edition, pp.766 , 2001.
- [15] Guosong Li, "Research on channel estimation in wireless OFDM systems", Ph.D thesis, University of Electronic Science and Technology of China, 2005.
- [16] IEEE P802.11n/D1.0, March 2006.
- [17] MAXIM Integrated Products, Datasheet of MAX2828/2829, 19-3455, rev0, Oct. 2004
- [18] F. M. Gardner, "Interpolation in digital modems - Part I: Fundamentals," IEEE Trans. Commun., vol. 41, pp. 501-507, Mar. 1993.
- [19] F. M. Gardner, "Interpolation in digital modems - Part II: Implementation and performance," IEEE Trans. Commun., vol. COM-41, pp. 998-1008, Jun. 1993.
- [20] F. M. Gardner, "A BPSK/QPSK timing-error detector for detector for sampled receivers," IEEE Trans. Commun., vol. COM-34, pp. 423-429, May. 1986.

Superimposed Training-Aided Channel Estimation for Multiple Input Multiple Output-Orthogonal Frequency Division Multiplexing Systems over High-Mobility Environment

Han Zhang¹, Xianhua Dai², Daru Pan¹ and Shan Gao¹

¹*School of Physics and Telecommunications Engineering,
South China Normal University, Guangzhou*

²*School of Information Science and Technology,
SUN Yat-sen University, Guangzhou
China*

1. Introduction

The combination of multiple-input multiple-output (MIMO) antennas and orthogonal frequency-division multiplexing (OFDM) can achieve a lower error rate and/or enable high-capacity wireless communication systems by flexibly exploiting diversity gain and/or the spatial multiplexing gains. Such systems, however, rely upon the knowledge of propagation channels. In many mobile communication systems, transmission is impaired by both delay and Doppler spreads [1]-[7]. In such cases, explicit incorporation of the time-varying characteristics of mobile wireless channel is called for.

The coefficients of a linearly time-varying (LTV) channel can be usually modeled as uncorrelated stationary random processes which are assumed to be low-pass, Gaussian, with zero mean (Rayleigh fading) or non-zero mean (Rician fading) depending on whether line-of-sight propagation is absent or present [1][6]. Recently, the basis expansion models, i.e. the truncated discrete Fourier basis (DFT) models, polynomial models and discrete prolate Spheroidal sequence models, have gained special attentions, especially for the situation that channel is caused by a few strong reflectors and path delays exhibit variations due to the kinematics of the mobiles [1]-[2] [5]-[6] [16] [25]-[28].

In conventional pilot-aided channel estimation approaches, MIMO channels can be effectively estimated by utilizing the time-division multiplexed (TDM) and (or) frequency-division multiplexed (FDM) training sequences [5]-[7] [20]-[23] [25]. Although the channel estimates are in general reliable, extra bandwidth or time slot is required for transmitting known pilots. In recent years, an alternative approach, referred to as superimposed training (ST), has been extensively studied in [8]-[19] [26]-[28]. In the idea of ST, additional periodic training sequences are arithmetically added to information sequence in time- or frequency-

domain. The advantage of the scheme is that there is no loss in information rate, and thus enables higher bandwidth efficiency. However, some useful power must inevitably be allocated to the pilots, and thus resulting in information signal-to-noise ratio (SNR) reduction. Meanwhile, the information sequences are viewed as interference to channel estimation since pilot symbols are superimposed at a low power to the information sequences at the transmitter. The existing ST-based channel estimations are mainly restricted to the case where the channel is linearly time-invariant (LTI), where the channel transfer function can be estimated by using first-order statistics [8]-[13] [17]-[18]. In the latest contributions, J. K. Tugnait [16] extended the conventional ST to time-varying environment where the LTV channels are modeled by complex exponential bases. For the issue of training power allocation, the optimal pilot power has been investigated by [24] for different taps of low-pass filter, and then, the optimization of ST power allocation for LTI channel is mathematically analyzed based on equalizer design [15] [19].

In this paper, a new ST-based channel estimator is proposed for OFDM/MIMO systems over LTV multipath fading channels. The main contributions are twofold. First, the LTV channel coefficients modeled by the truncated discrete Fourier bases (DFB), unlike the existing approaches [1]-[2] [5]-[6] [16], cover multiple OFDM symbols. Then, a two-step channel estimation approach is adopted for LTV channel estimation. Furthermore, a closed-form expression of the estimation variance is derived, which provides a guideline for designing the superimposed pilot symbols. We demonstrate by analytical analysis that the estimation variance, unlike that of conventional ST-based schemes [8]-[19], approaches to a fixed lower-bound as the training length increases. Second, for wireless communication systems with a limited transmission power, unlike [10] where the issue of ST power allocation is derived by optimizing the SNR for equalizer design, we provide an optimal solution of ST power allocation with a different point of view by maximizing the lower-bound of channel capacity. Comparatively, the training power allocation scheme [10] can be otherwise considered as a special case compared with the proposed approach. In simulations presented in this paper, we compare the results of our approach with that of the FDM training approaches [5] as latter serves as a “benchmark” in related works. It is shown that the proposed algorithm outperforms that of FDM training, and yields higher transmission efficiency.

The rest of the paper is organized as follows. Section II presents the system and channel models. In Section III, we estimate the LTV channel coefficients with the proposed two-step channel estimation approach. In Section IV, we derive the closed-form expression of the channel estimation variances. Section V determines the optimal ratio of the ST power to the total transmission power by maximizing the lower-bound of channel capacity. Section VI reports on some simulation experiments in order to test the validity of theoretic results, and we conclude the paper with Section VII.

Notation: The letter t represents the time-domain variable and k is the frequency-domain variable. Bold letters denote the matrices and column-vectors, and the superscripts $[\bullet]^T$ and $[\bullet]^H$ represent the transpose and conjugate transpose operations, respectively. $[\bullet]_{k,t}$ denotes the (k, t) element of the specified matrix.

2. System and channel model

2.1 System model

Consider an MIMO/OFDM system of N transmitters or mobile users and a receive array of M receive antennas with perfect synchronization. At transmit terminals, an inverse fast Fourier transform (IFFT) is used as a modulator. The modulated outputs are given by

$$\mathbf{X}_n(i) = [x_n(i,0), \dots, x_n(i,t), \dots, x_n(i,B-1)]^T = \mathbf{F}^{-1} \mathbf{S}_n(i) \quad n = 1, \dots, N \quad (1)$$

where B is OFDM symbol-size, $\mathbf{S}_n(i) = [s_n(i,0), \dots, s_n(i,k), \dots, s_n(i,B-1)]^T$ is the i th transmitted symbol of the n th transmit antenna. \mathbf{F}^{-1} is the IFFT matrix with $[\mathbf{F}^{-1}]_{k,t} = e^{j2\pi kt/B}$ and $j^2 = -1$. Then, $\mathbf{X}_n(i)$ is concatenated by a cyclic-prefix (CP) of length \bar{L} , propagating through the respective channels. At receiver, the received signals of m th receive antenna, discarding CP and stacking the received signals $y^{(m)}(i,t)$ $t = 0, \dots, B-1$, can be written in a vector-form as

$$\mathbf{Y}^{(m)}(i) = [y^{(m)}(i,0), \dots, y^{(m)}(i,t), \dots, y^{(m)}(i,B-1)]^T \quad m = 1, \dots, M \quad (2)$$

and the received signals $y^{(m)}(i,t)$ in (2) is given by

$$\begin{aligned} y^{(m)}(i,t) &= \sum_{n=1}^N \mathbf{X}_n(i) \otimes \mathbf{h}_n^{(m)}(i) + v^{(m)}(i,t) \\ &= \sum_{n=1}^N \sum_{l=0}^{L-1} h_{n,l}^{(m)}(i,t) x_n(i,t-l) + v^{(m)}(i,t) \quad t = 0, \dots, B-1 \end{aligned} \quad (3)$$

where $\mathbf{h}_n^{(m)}(i) = [h_{n,0}^{(m)}(i,t), \dots, h_{n,L-1}^{(m)}(i,t), 0_{1 \times B-L}]^T$ is the impulse response vector of the propagating channel from the n th transmit to the m th receive antenna. The channel coefficients $h_{n,l}^{(m)}(i,t)$, $l = 0, \dots, L-1$ is the functions of time variable t which will be defined by (6). The notation \otimes represents the cyclic convolution and $v^{(m)}(i,t)$ is the additive Gaussian noise.

At receiver, an FFT operation is performed on the vector (2), and the demodulated outputs can be written as

$$\mathbf{U}^{(m)}(i) = [u^{(m)}(i,0), \dots, u^{(m)}(i,k), \dots, u^{(m)}(i,B-1)]^T = \mathbf{F} \mathbf{Y}^{(m)}(i) \quad m = 1, \dots, M. \quad (4)$$

From (3) and the duality of time and frequency, the FFT demodulated signals in (4) can be written as

$$\begin{aligned} u^{(m)}(i,k) &= FFT \left\{ \sum_{n=1}^N \sum_{l=0}^{L-1} h_{n,l}^{(m)}(i,t) x_n(i,t-l) + v^{(m)}(i,t) \right\} \\ &= \sum_{n=1}^N \sum_{l=0}^{L-1} FFT \left\{ h_{n,l}^{(m)}(i,t) \right\} \otimes FFT \left\{ x_n(i,t) \right\} + \bar{v}^{(m)}(i,k) \end{aligned} \quad (5)$$

where $\text{FFT}\{\bullet\}$ represents the FFT vector of the specified function and $\bar{v}^{(m)}(i, k)$ is the frequency-domain noise. Compared with the FFT demodulated signals of OFDM systems with LTI channels, the convolution in (5) between the information sequences and the FFT vectors of time-varying channel coefficients may introduce inter-carrier interference (ICI).

2.2 Channel model

As mentioned in [1], the coefficients of the time- and frequency-selective channel can be modeled as Fourier basis expansions. Thereafter, this model was intensively investigated and applied in block transmission, channel estimation and equalization (e.g. [2][5]-[6][16]). In this paper, we extend the block-by-block process [2][5]-[6][16] to the case where multiple OFDM symbols are utilized. Consider a time interval or segment $\{t: (\ell - 1)\Omega \leq t \leq \ell\Omega\}$, the channel coefficients in (3) can be approximated by truncated discrete Fourier bases (DFB) within the segment as

$$\begin{aligned} h_{n,l}^{(m)}(i, t) &\approx \sum_{q=0}^Q h_{n,l,q}^{(m)} e^{-j2\pi(q-Q/2)t/\Omega} \\ &= \sum_{q=0}^Q h_{n,l,q}^{(m)} \eta_q(t) \quad t = (\ell - 1)\Omega, \dots, \ell\Omega, \quad \ell = 1, 2, \dots \end{aligned} \quad (6)$$

where $h_{n,l,q}^{(m)}$ is a constant coefficient, Q represents the basis expansion order that is generally defined as $Q \geq 2f_d\Omega/f_s$ [1], $\Omega > B$ is the segment length and ℓ is the segment index. Unlike [1]-[2] [5]-[6] [16], the approximation frame Ω covers multiple OFDM symbols, denoted by $i = 1, \dots, I$, where $I = \Omega/B'$ and $B' = B + \bar{L}$. Since the proposed two-step channel estimation as will be shown in Section III is adopted within one frame, we omit the segment index ℓ for simplicity.

3. ST-based channel estimation

In this section, we propose a ST-based two-step approach for LTV channel estimation. In ST-based approaches [8]-[19], the pilot symbols are superimposed (arithmetically added) to the information sequences as

$$s_n(i, k) = c_n(i, k) + p_n(i, k) \quad k = 0, \dots, B - 1 \quad (7)$$

where $c_n(i, k)$ and $p_n(i, k)$ are the information and pilot sequence, respectively. Compared with the FDM/TDM training aided methods [20]-[22], ST requires no additional bandwidth (or time-slot) for transmitting known pilots, and thus offers a higher data rate.

3.1 ST-based channel estimation over one OFDM symbol

For LTV environment where the channel coefficient $h_{n,l}^{(m)}(t)$ is a function of time variable t , the vectors $\text{FFT}\{h_{n,l}^{(m)}(t)\}$ in (5) cannot be approximated as a δ -sequences and, the FFT demodulated signals at the sub-carrier k of the i th symbol is given by

$$\begin{aligned}
 u^{(m)}(i, k) &= \sum_{n=1}^N \sum_{l=0}^{L-1} \text{FFT} \left\{ h_{n,l}^{(m)}(i, t) \right\} \otimes P_n(i) + \bar{v}^{(m)}(i, k) \\
 &\approx \sum_{n=1}^N \sum_{l=0}^{L-1} \text{FFT} \left\{ \sum_{q=0}^Q h_{n,l,q}^{(m)} \eta_q(t) \right\} \otimes P_n(i) + \bar{v}^{(m)}(i, k) \\
 &= \sum_{n=1}^N \sum_{q=0}^Q H_{n,q}^{(m)}(i, k) \eta_q(t_i) \mathbf{W}_q(i, k) \otimes P_n(i) + \bar{v}^{(m)}(i, k)
 \end{aligned} \tag{8}$$

where $\bar{v}^{(m)}(i, k) = \sum_{n=1}^N \sum_{l=0}^{L-1} \text{FFT} \left\{ h_{n,l}^{(m)}(t) \right\} \otimes C_n(i) + \bar{v}^{(m)}(i, k)$ and $t_i = (i-1)B + B/2$. $\mathbf{W}_q(i, k)$ with $k = 0, \dots, B-1$ is the FFT vector of the complex exponential function (CEF) and, can be written as

$$\begin{aligned}
 \mathbf{W}_q(i, 0) &= [w_q(i, 0), \dots, w_q(i, k), \dots, w_q(i, B-1)]^T \\
 &= \mathbf{F} \left[\eta_q(t_i - B/2) / \eta_q(t_i), \dots, \eta_q(t_i + B/2 - 1) / \eta_q(t_i) \right]^T.
 \end{aligned} \tag{9}$$

Notice that $\mathbf{W}_q(i, k)$ is a cyclic-shifted vector of $\mathbf{W}_q(i, 0)$ with a shifting length k . On the other hand, ICI introduced by the cyclic convolution $\mathbf{W}_q(i, k) \otimes S_n(i)$ depends explicitly on $\eta_q(t)$, $t = 0, \dots, B-1$. When q is not large, the complex exponential functions in (9) are slowly time-varying over an OFDM symbol-duration and, thereby, the principal power or major-lobe of the FFT vector $\mathbf{W}_q(i, 0)$ may concentrate on its two ends (low frequency tones) with indexes $0, \dots, T$ and $B-T, \dots, B-1$. Using the major-lobe to approximate the CEF vectors $\mathbf{W}_q(i, 0)$ $q = 0, 1, \dots, Q$, we have

$$\mathbf{W}_q(i, 0) \approx [w_q(i, 0), \dots, w_q(i, T), 0, \dots, 0, w_q(i, B-T), \dots, w_q(i, B-1)]^T \quad q = 0, 1, \dots, Q \tag{10}$$

where T is a positive integer.

In general, the FFT vector of the function $\eta_q(t)$ in (10) may have a great side-lobe that results in a great error. For improving the approximation performance, an intuitional idea is to apply a window function to the received signals in order to reduce the side-lobe leakage. The windowed vector of received signals in (3) of the i th symbol is

$$\bar{y}^{(m)}(i, t) = \sum_{n=1}^N h_n^{(m)}(i, t) \psi_B(t) x_n(t-l) + v^{(m)}(t) \psi_B(t) \quad t = 0, \dots, B-1 \tag{11}$$

where $\psi_B(t)$ is a time-domain windowing function with a length B . Performing the FFT demodulated operation on the windowed sequences in (10), the demodulated signals, by (10), can be written by

$$\begin{aligned}
 u^{(m)}(i, k) &\approx \sum_{n=1}^N \sum_{l=0}^{L-1} \text{FFT} \left\{ \sum_{q=0}^Q h_{n,l,q}^{(m)} \eta_q(t) \psi_B(t) \right\} \otimes P_n(i) + \bar{v}^{(m)}(i, k) \\
 &= \sum_{n=1}^N \sum_{q=0}^Q H_{n,q}^{(m)}(i, k) \eta_q(t_i) \bar{\mathbf{W}}_q(i, k) \otimes P_n(i) + \bar{v}^{(m)}(i, k)
 \end{aligned} \tag{12}$$

where $\bar{W}_q(i, k)$ is the CEF vector with the windowing function $\psi_B(t)$ as

$$\begin{aligned} \bar{W}_q(i, k) &= [\bar{w}_q(i, 0), \dots, \bar{w}_q(i, k), \dots, \bar{w}_q(i, B-1)]^T \\ &= \mathbf{F} \left[\psi_B(t) \eta_q(t_i - B/2) / \eta_q(t_i), \dots, \psi_B(t) \eta_q(t_i + B/2 - 1) / \eta_q(t_i) \right]^T \\ &\approx [\bar{w}_q(i, 0), \dots, \bar{w}_q(i, T), 0, \dots, 0, \bar{w}_q(i, B-T), \dots, \bar{w}_q(i, B-1)]^T. \end{aligned} \tag{13}$$

Compared with (10), the approximation of windowing based vector has a much smaller side-lobe with the same index T . The experiment studies show that by using a *Kaiser* function [5], the approximation in (13) of $T = 2$ may capture almost 99% power of $FFT[\psi_B(t) \eta_q(t_i - B/2 : t_i + B/2 - 1) / \eta_q(t_i)]^T$ for truncated DFBs when $q < B/10$. Substituting (13) into (12), the FFT demodulated outputs can be approximated by

$$\begin{aligned} u^{(m)}(i, k) &= \sum_{n=1}^N \sum_{q=0}^Q H_{n,q}^{(m)}(i, k) \eta_q(t_i) \left[\sum_{k'=0}^T \bar{w}_q(i, k') p_n(i, k - k') + \right. \\ &\quad \left. \sum_{k'=T}^1 \bar{w}_q(i, B - k') p_n(i, k + k') \right] + \bar{v}^{(m)}(i, k). \end{aligned} \tag{14}$$

The first term of (14) illustrates that $2T + 1$ tones, i.e. $p_n(i, k - k'), \dots, p_n(i, k + k')$ should be jointly designed for estimating $H_{n,q}^{(m)}(i, k)$. We refer to such $2T + 1$ consecutive pilot tones as a pilot cluster for differentiating from the isolated tones utilized in the LTI channel estimation [19] [22]-[23]. Denote k_1, \dots, k_Γ as the pilot cluster indexes located at the τ th pilot symbol and, $p_n(k_\tau - T), \dots, p_n(k_\tau + T)$ as the pilot sequences at the pilot cluster k_τ . Since the ST does not entail additional bandwidth, two adjacent pilot-clusters, i.e. k_τ and $k_{\tau+1}$ can be placed closed together. The pilot tone distribution is shown in Fig. 1.

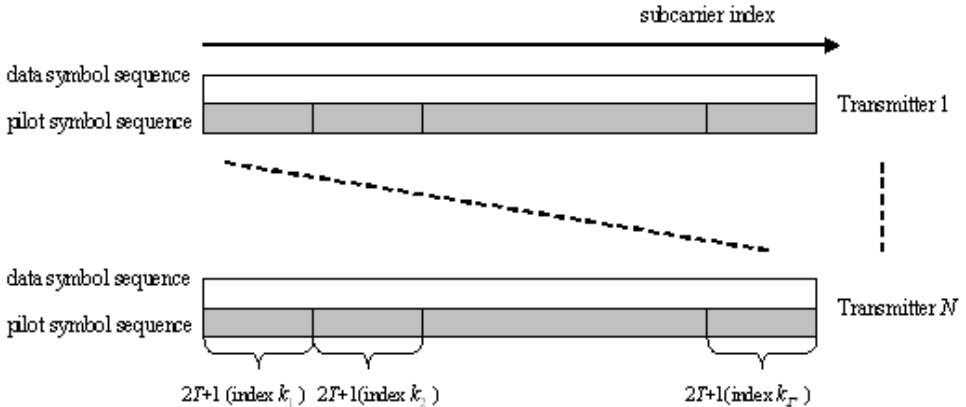


Fig. 1. A typical pilot tone distribution. $2T + 1$ consecutive tones are grouped together as one pilot cluster. All pilot clusters are uniformly distributed in frequency domain with each adjacent pilot cluster being closed together.

Then, we focus on ST design. From (14), when the training sequence at each pilot cluster is designed as either a constant modulus sequence, i.e.

$$p_n(i, k_\tau) = p_n(i, k_\tau \pm k') \quad \tau = 1, \dots, \Gamma, k' = 1, \dots, T \tag{15}$$

or a δ sequence, i.e.

$$p_n(i, k_\tau \pm k') = \begin{cases} p_n(i, k_\tau), k' = 0 \\ 0, & \text{otherwise} \end{cases} \quad \tau = 1, \dots, \Gamma, k' = 1, \dots, T. \tag{16}$$

Accordingly, the FFT demodulated outputs at pilot cluster k_τ can be approximated as

$$\begin{aligned} u^{(m)}(i, k_\tau) &\approx \sum_{n=1}^N \sum_{q=0}^Q H_{n,q}^{(m)}(i, k_\tau) \eta_q(t_i) g_q(i, k_\tau) p_n(i, k_\tau) + \bar{v}^{(m)}(i, k_\tau) \\ &= \sum_{n=1}^N H_n^{(m)}(i, k_\tau) p_n(i, k_\tau) + \bar{v}^{(m)}(i, k_\tau) \end{aligned} \tag{17}$$

where $g_q(i, k_\tau) = \sum_{k=0}^T \bar{w}_q(i, k_\tau + k') + \sum_{k=-T}^1 \bar{w}_q(i, k_\tau + B - k')$ if the training sequence takes value from (15) and $g_q(i, k_\tau) = \bar{w}_q(i, k_\tau)$ for (16), which are all known, respectively. The channel transfer functions $H_n^{(m)}(i, k_\tau)$ is given by

$$\begin{aligned} H_n^{(m)}(i, k_\tau) &= \sum_{q=0}^Q H_{n,l,q}^{(m)}(i, k_\tau) \eta_q(t_i) g_q(i, k_\tau) \\ &= \sum_{l=0}^{L-1} \sum_{q=0}^Q h_{n,l,q}^{(m)} \eta_q(t_i) g_q(i, k_\tau) e^{-j2\pi k_\tau l/B} \approx \sum_{l=0}^{L-1} \hat{h}_{n,l}^{(m)}(t_i) e^{-j2\pi k_\tau l/B}. \end{aligned} \tag{18}$$

From (17)-(18), we note that $H_n^{(m)}(i, k_\tau), \tau = 1, \dots, \Gamma$ is in fact a LTI system transfer function of which the coefficients are the mid-values of the LTV channel at the i th OFDM symbol interval. As a result, the LTV channel estimation can be approximately reduced into that of the LTI channel [22] and [23] by simply designing the ST sequences as (15) or (16).

Let $\mathbf{H}^{(m)}(i) = [\hat{h}_{1,0}^{(m)}(t_i), \dots, \hat{h}_{1,L-1}^{(m)}(t_i), \dots, \hat{h}_{N,0}^{(m)}(t_i), \dots, \hat{h}_{N,L-1}^{(m)}(t_i)]^T$ be the channel coefficient vector associated with the i th OFDM symbol and stack the FFT demodulated signals at pilot clusters of the i th OFDM symbol to form a vector

$$\mathbf{U}^{(m)}(i, k_1 : k_\Gamma) = [u^{(m)}(i, k_1), \dots, u^{(m)}(i, k_\tau), \dots, u^{(m)}(i, k_\Gamma)]^T. \tag{19}$$

The received signals at pilot clusters can be thus written as

$$\begin{aligned} \mathbf{U}^{(m)}(i, k_1 : k_\Gamma) &= \underbrace{\mathbf{A}(i) \mathbf{H}^{(m)}(i)}_{\text{desired signal for channel estimation}} + \underbrace{\mathbf{\Xi}^{(m)}(i, k_1 : k_\Gamma)}_{\text{information interference on channel estimation}} \\ &\quad + \mathbf{\bar{V}}^{(m)}(i, k_1 : k_\Gamma) \end{aligned} \tag{20}$$

where $\bar{\mathbf{V}}^{(m)}(i, k_1 : k_r)$ is the noise vector in frequency-domain, $\Xi^{(m)}(i, k_1 : k_r) = [\Xi^{(m)}(i, k_1), \dots, \Xi^{(m)}(i, k_r)]^T$ is the interference vector produced by the information sequences with $\Xi^{(m)}(i, k_\tau) = \sum_{n=1}^N H_n^{(m)}(i, k_\tau) c_n(i, k_\tau)$, $\mathbf{A}(i) = [\mathbf{A}(1, 0), \dots, \mathbf{A}(1, L-1), \dots, \mathbf{A}(N, l), \dots, \mathbf{A}(N, L-1)]$ is a $\Gamma \times NL$ matrix with the column-vectors

$$\mathbf{A}(n, l) = [p_n(i, k_1) e^{-j2\pi k_1 l/B}, \dots, p_n(i, k_\tau) e^{-j2\pi k_\tau l/B}, \dots, p_n(i, k_r) e^{-j2\pi k_r l/B}]^T. \quad (21)$$

Since the matrix $\mathbf{A}(i)$ is known, when $\Gamma \geq NL$, the matrix $\mathbf{A}(i)$ is of full column rank, and the channel coefficient vectors can be thus estimated by

$$\begin{aligned} \hat{\mathbf{H}}^{(m)}(i) &= \mathbf{A}^\dagger \mathbf{U}^{(m)}(i, k_1 : k_r) \\ &= \mathbf{H}^{(m)}(i) + \mathbf{A}(i)^\dagger \Xi^{(m)}(i, k_1 : k_r) + \mathbf{A}(i)^\dagger \bar{\mathbf{V}}^{(m)}(i, k_1 : k_r) \quad m = 1, \dots, M, i = 1, \dots, I \end{aligned} \quad (22)$$

where the superscript ‘ \dagger ’ is the pseudo-inverse operation, and the hat ‘ $\hat{\cdot}$ ’ indicates the estimation. From (22), the mainly computational effort is directly proportional to the unknown parameter number NL .

Using the specifically designed ST sequences in (15) and (or) (16), the problem of LTV channel estimation for MIMO/OFDM systems can be reduced into that of LTI channel. From (20) and (22), however, we notice that the interference vector due to information sequence can hardly be neglected since the power of data symbol is much larger than the pilot power. For conventional ST based schemes stated in [8]-[13] [17]-[18], first-order statistics are employed to suppress the information sequence interference over multiple training periods in the case that the channel is LTI during the record length. Such arithmetical average process, however, is no longer feasible to the channel assumed in this paper where the channel coefficients are linearly time-variant between consecutive OFDM symbols.

3.2 Channel estimation over multiple OFDM symbols

In this sub-section, a weighted average approach is developed to suppress the abovementioned information sequence interference over multiple OFDM symbols, and thus overcoming the shortcoming of the existing ST-based approach in estimating the time-variant channels.

By (22), the LTV channel coefficients can be obtained following the relationship $h_{n,l}^{(m)}(t_i) = \sum_{q=0}^Q h_{n,l,q}^{(m)} n_q(t_i)$. Taking the LTV channel coefficient estimation of each OFDM symbol $\hat{h}_{n,l}^{(m)}(t_i)$ $i = 1, \dots, I$ by (22) as a temporal result, and form a vector as $\hat{\mathbf{h}}_{n,l}^{(m)} = [\hat{h}_{n,l}^{(m)}(t_1), \dots, \hat{h}_{n,l}^{(m)}(t_I)]^T$, we thus have

$$\hat{\mathbf{h}}_{n,l}^{(m)} = \boldsymbol{\eta} \hat{\mathbf{h}}_{n,l,q}^{(m)}$$

$$= \begin{bmatrix} e^{j2\pi(0-Q/2)t_l/\Omega} & \dots & e^{j2\pi(Q-Q/2)t_l/\Omega} \\ \vdots & \ddots & \vdots \\ e^{j2\pi(0-Q/2)t_l/\Omega} & \dots & e^{j2\pi(Q-Q/2)t_l/\Omega} \end{bmatrix} \begin{bmatrix} \hat{h}_{n,l,0}^{(m)} \\ \vdots \\ \hat{h}_{n,l,Q}^{(m)} \end{bmatrix} \quad n = 1, \dots, N, l = 0, \dots, L-1 \quad (23)$$

where $\hat{\mathbf{h}}_{n,l,q}^{(m)} = [\hat{h}_{n,l,0}^{(m)}, \dots, \hat{h}_{n,l,q}^{(m)}, \dots, \hat{h}_{n,l,Q}^{(m)}]^T$ is estimation of the complex exponential coefficients vector modeling the LTV channel, $\boldsymbol{\eta}$ is a $I \times (Q + 1)$ matrix with $[\boldsymbol{\eta}]_{q,i} = e^{j2\pi(q-Q/2)t_i/\Omega}$. Thus, when $I \geq Q + 1$, the matrix $\boldsymbol{\eta}$ is of full column rank, and the basis expansion model coefficients can be computed by

$$\hat{\mathbf{h}}_{n,l,q}^{(m)} = \boldsymbol{\eta}^\dagger \hat{\mathbf{h}}_{n,l}^{(m)} \quad n = 1, \dots, N, l = 0, \dots, L-1 \quad (24)$$

Substituting $t_i = (i - 1)B + B/2$ into the matrix $\boldsymbol{\eta}$, we obtain the pseudo-inverse matrix as

$$[\boldsymbol{\eta}^\dagger]_{i,q} = e^{-j2\pi(q-Q/2)((i-1)B+B/2)/\Omega} / I \quad (25)$$

By (23)-(25), the modeling coefficients (6) can be computed by

$$\hat{h}_{n,l,q}^{(m)} = \sum_{i=1}^I e^{-j2\pi(q-Q/2)((i-1)B+B/2)/\Omega} \hat{h}_{n,l}^{(m)}(i) / I \quad (26)$$

In fact, (26) is estimated over multiple OFDM symbols with a weighted average function of $e^{-j2\pi(q-Q/2)t_i/\Omega} / I$.

Compared with the conventional ST strategies, the proposed channel estimation is composed of two steps: First, with specially designed ST signals in (15) and (16), channel estimation can be reduced into that of LTI channel, and we are allowed to estimate the channel coefficients during each OFDM symbol as temporal results. Second, the temporal channel estimates are further enhanced over multiple OFDM symbols by using a weighted average procedure. That is, not only the target OFDM symbol, but also the OFDM symbols over the whole frame are invoked for channel estimation. Similar to the first-order statistics of LTI case [8]-[13] [17]-[18], it is thus anticipated that the weighted average estimation may also exhibit a considerable performance improvement for the LTV channels over a long frame Ω .

4. Channel estimation analysis

In this section, we analyze the performance of the channel estimator proposed in Section III and derive a closed-form expression of the channel estimation variance which can be, in turn, used for ST power allocation. Before going further, we make the following assumptions:

(H1) The information sequence $\{c_n(i, k)\}$ is zero-mean, finite-alphabet, i.i.d., and equi-powered with the power σ_c^2 .

(H2) The additive noise $\{v^{(m)}(i, t)\}$ is white, uncorrelated with $\{c_n(i, k)\}$, with $E\left[\left|v^{(m)}(i, t)\right|^2\right] = \sigma_v^2$.

(H3) The LTV channel coefficients $\mathbf{h}_{n,l}^{(m)}$ are complex Gaussian variables, and statistically independent for different values of n and l .

From (22)-(26), the mean square error (MSE) of channel estimation is given by

$$\begin{aligned} \text{MSE}^{(m)} &= E\left\{\sum_{n=1}^N \sum_{l=0}^{L-1} \left\|h_{n,l}^{(m)}(i, t) - \sum_{q=0}^Q \hat{h}_{n,l,q}^{(m)}\eta_q(t)\right\|^2\right\} \\ &= E\left\{\sum_{n=1}^N \sum_{l=0}^{L-1} \left\|h_{n,l}^{(m)}(i, t) - \sum_{q=0}^Q h_{n,l,q}^{(m)}\eta_q(t) + \sum_{q=0}^Q h_{n,l,q}^{(m)}\eta_q(t) - \sum_{q=0}^Q \hat{h}_{n,l,q}^{(m)}\eta_q(t)\right\|^2\right\} \end{aligned} \quad (27)$$

where $\|\cdot\|$ is the Euclidean norm. In (27), the first error term $\sum_{n=1}^N \sum_{l=0}^{L-1} \left[h_{n,l}^{(m)}(i, t) - \sum_{q=0}^Q h_{n,l,q}^{(m)}\eta_q(t)\right]$ is caused by the orthonormal basis expansion model in (6), which is referred to as the channel modeling error. The second error term $\sum_{n=1}^N \sum_{l=0}^{L-1} \sum_{q=0}^Q \left[h_{n,l,q}^{(m)}\eta_q(t) - \hat{h}_{n,l,q}^{(m)}\eta_q(t)\right]$ is due to the information interference to channel estimation (22) and additive noise. Explicitly, two error signals are mutually independent. Herein, we do not elaborate the topic of channel modeling error and focus on channel estimation error, which is mainly produced by the interference of information sequence. By (H2), the MSE of the estimation in one OFDM symbol can be written as

$$\begin{aligned} \text{MSE}^{(m)}(i) &\stackrel{\text{def}}{=} \frac{1}{(Q+1)NL} E\left\{\sum_{n=1}^N \sum_{l=0}^{L-1} \sum_{q=0}^Q \left\|h_{n,l,q}^{(m)}\eta_q(t) - \hat{h}_{n,l,q}^{(m)}\eta_q(t)\right\|^2\right\} \\ &= \frac{1}{(Q+1)NL} \text{tr}\left\{\mathbf{A}(i)^\dagger E\left\{\underbrace{\Xi^{(m)}(i, k_1 : k_r)(\Xi^{(m)}(i, k_1 : k_r))^H}_{\text{estimation variance due to information sequence interference}}\right\}(\mathbf{A}(i)^\dagger)^H\right\} \\ &+ \frac{1}{(Q+1)NL} \text{tr}\left\{\mathbf{A}(i)^\dagger E\left\{\underbrace{\bar{\mathbf{V}}^{(m)}(i, k_1 : k_r)(\bar{\mathbf{V}}^{(m)}(i, k_1 : k_r))^H}_{\text{estimation variance due to additive noise}}\right\}(\mathbf{A}(i)^\dagger)^H\right\}. \end{aligned} \quad (28)$$

For zero-mean white noise, we have

$$E\left\{\bar{\mathbf{V}}^{(m)}(i, k_1 : k_r)(\bar{\mathbf{V}}^{(m)}(i, k_1 : k_r))^H\right\} = \sigma_v^2 \mathbf{I}_r. \quad (29)$$

Invoking the assumption (H1), information sequence interference $\Xi^{(m)}(i, k_\tau) = \sum_{n=1}^N H_n^{(m)}(i, k_\tau)c_n(i, k_\tau)$, $\tau = 1, \dots, r$ is approximately Gaussian distributed for a

large Γ . Therefore, the channel estimation variance due to information sequence interference can be obtained as

$$E\left\{\Xi^{(m)}(i, k_1 : k_{\Gamma})\left(\Xi^{(m)}(i, k_1 : k_{\Gamma})\right)^H\right\} = \frac{\sigma_c^2}{\Gamma^2} \sum_{\tau=0}^{\Gamma-1} \sum_{n=1}^N \sum_{l=0}^{L-1} \left|H_n^{(m)}(i, k_{\tau})\right|^2 = \frac{\rho_c}{\Gamma} \sum_{\tau=0}^{\Gamma-1} \sum_{n=1}^N \sum_{l=0}^{L-1} \left|\bar{h}_{n,l}^{(m)} e^{-2\pi k_{\tau} l/B}\right|^2. \quad (30)$$

Substituting (29) and (30) into (28), we have

$$MSE^{(m)}(i) \stackrel{def}{=} \frac{1}{(Q+1)NL} \left(\sigma_v^2 + \frac{\sigma_c^2}{\Gamma} \sum_{\tau=0}^{\Gamma-1} \sum_{n=1}^N \sum_{l=0}^{L-1} \left|\bar{h}_{n,l}^{(m)} e^{-2\pi k_{\tau} l/B}\right|^2 \right) \text{tr}\left[\left(\mathbf{A}(i)\right)^H \mathbf{A}(i)\right]^{-1}. \quad (31)$$

Apparently, the channel estimation performance depends crucially on the matrix $\mathbf{A}(i)$. The optimal estimation or minimum MSE (MMSE) estimation may require $\mathbf{A}(i)^H \mathbf{A}(i) = \Phi \mathbf{I}$ where Φ is a constant. From (21), we adopt the training sequence as $p_n(i, k) = \sigma_p, k = 0, \dots, B-1$ as (15), the above MMSE condition can be well satisfied. We thus have

$$\text{tr}\left[\left(\mathbf{A}(i)\right)^H \mathbf{A}(i)\right] = \Gamma \sigma_p^2 \mathbf{I}_{NL(Q+1)}. \quad (32)$$

Substituting (32) into (31), the MSE of channel estimation over one OFDM symbol can be derived as

$$MSE^{(m)}(i) = \frac{\sigma_c^2}{\Gamma^2 \sigma_p^2} \sum_{\tau=0}^{\Gamma-1} \sum_{n=1}^N \sum_{l=0}^{L-1} \left|\bar{h}_{n,l}^{(m)} e^{-2\pi k_{\tau} l/B}\right|^2 + \frac{\sigma_v^2}{\Gamma \sigma_p^2}. \quad (33)$$

It is seen that the first term of (33) is the estimation variance due to information interference, and depends upon the channel transfer functions. We thus define the normalized variance as

$$NMSE^{(m)}(i) = \frac{\sigma_c^2}{\Gamma^2 \sigma_p^2} \sum_{\tau=0}^{\Gamma-1} \sum_{n=1}^N \sum_{l=0}^{L-1} \left|\bar{h}_{n,l}^{(m)} e^{-2\pi k_{\tau} l/B}\right|^2 \bigg/ \left|\bar{h}^{(m)}(i)\right|^2 \quad (34)$$

where $\left|\bar{h}^{(m)}(i)\right|^2 = \sum_{\tau=0}^{\Gamma-1} \sum_{n=1}^N \sum_{l=0}^{L-1} \left|\bar{h}_{n,l}^{(m)} e^{-2\pi k_{\tau} l/B}\right|^2 / NL\Gamma$. Following the definition of (34), we obtain the normalized variance as

$$NMSE^{(m)}(i) = \frac{\sigma_c^2}{\Gamma^2 \sigma_p^2} \sum_{\tau=0}^{\Gamma-1} \sum_{n=1}^N \sum_{l=0}^{L-1} \left|\bar{h}_{n,l}^{(m)} e^{-2\pi k_{\tau} l/B}\right|^2 \bigg/ \left|\bar{h}^{(m)}(i)\right|^2 = \frac{NL}{\Gamma} \frac{\sigma_c^2}{\sigma_p^2}. \quad (35)$$

From (35), we can find that the estimation variance due to the information interference is directly proportional to the information-to-pilot power ratio σ_c^2 / σ_p^2 , thereby resulting in an inaccurate solution for the general case that $\sigma_c^2 \gg \sigma_p^2$.

Then, we analyze the channel estimation performance of the weighted average approach over multiple OFDM symbols (the whole frame Ω). Define the vectors $\mathbf{A} = [\mathbf{A}(1), \dots, \mathbf{A}(I)]^T$,

$\Xi^{(m)} = [\Xi^{(m)}(1, k_1 : k_r), \dots, \Xi^{(m)}(I, k_1 : k_r)]^T$ and $V^{(m)} = [\bar{V}^{(m)}(1, k_1 : k_r), \dots, \bar{V}^{(m)}(I, k_1 : k_r)]^T$, the MSE of the weighted average channel estimator over multiple OFDM symbols is given by

$$MSE^{(m)} = \frac{1}{(Q+1)NL} \text{tr} \left\{ \boldsymbol{\eta}^\dagger E \left\{ \left\| \mathbf{A}^\dagger \Xi^{(m)} + \mathbf{A}^\dagger V^{(m)} \right\|^2 \right\} (\boldsymbol{\eta}^\dagger)^H \right\} = \frac{1}{I} \sum_{i=1}^I MSE^{(m)}(i) \text{tr} \left\{ \boldsymbol{\eta}^\dagger (\boldsymbol{\eta}^\dagger)^H \right\}. \quad (36)$$

Note that the column vectors of the matrix $\boldsymbol{\eta}$ in (23) are in fact the FFT vectors of a $l \times l$ matrix, we thus have $\boldsymbol{\eta}^H \boldsymbol{\eta} = \mathbf{I}_{(Q+1)}$ and $\text{tr} [\boldsymbol{\eta}^H \boldsymbol{\eta}]^{-1} = (Q+1)/I$. Substituting (33) into (36), the MSE of channel estimation over multiple OFDM symbols is given by

$$MSE^{(m)} = \frac{(Q+1)\sigma_c^2}{I^2\sigma_p^2} \sum_{\tau=0}^{\Gamma-1} \sum_{n=1}^N \sum_{l=0}^{L-1} \left| \bar{h}_{n,l}^{(m)} e^{-2\pi k_l l/B} \right|^2 + \frac{(Q+1)\sigma_v^2}{I\sigma_p^2} \quad (37)$$

In (37), the second term is caused by information sequence interference, which may become the dominant component of the channel estimation variance for the general case of $\sigma_c^2 \gg \sigma_p^2$, especially for large SNRs. Therefore, we solely consider information sequence effect. Similar to (34)-(35), we derive the normalized variance due to information interference by removing the channel gain as

$$NMSE^{(m)} = \frac{(Q+1)\sigma_c^2}{I^2\sigma_p^2} \sum_{\tau=0}^{\Gamma-1} \sum_{n=1}^N \sum_{l=0}^{L-1} \left| \bar{h}_{n,l}^{(m)} e^{-2\pi k_l l/B} \right|^2 / \left| \bar{h}^{(m)} \right|^2 \quad (38)$$

where $\left| \bar{h}^{(m)} \right|^2 = \sum_{i=1}^I \left| \bar{h}^{(m)}(i) \right|^2 / I$. It follows that

$$NMSE^{(m)} = \frac{\sigma_c^2}{\sigma_p^2} \frac{NL(Q+1)}{I\Gamma} \approx \frac{\sigma_c^2}{\sigma_p^2} \frac{NL(Q+1)}{\Omega} \frac{B}{\Gamma} = \frac{\sigma_c^2}{\sigma_p^2} \frac{NL(Q+1)}{\theta\Omega} \quad (39)$$

where $\theta = \Gamma/B$ is the training ratio of one OFDM symbol. For conventional ST-based LTI schemes where isolated pilots are exploited for channel estimation [8]-[13] [17]-[18], we have $\theta = 1$. However, for estimating the LTV channels addressed in this paper, Γ pilot clusters, instead of isolated pilot tones, are exploited. Thus, the corresponding training ratio yields $\theta \leq 1/(2T+1)$. From (39), the normalized variance is directly proportional to the information-pilot power ratio σ_c^2/σ_p^2 , the training ratio θ and the ratio of unknown parameter number $NL(Q+1)$ over the frame length Ω .

Compared with the variances of channel estimation over one OFDM symbol as in (33)-(35), the estimation variances of the weighted average estimator(37)-(39) is significantly reduced owing to the fact that $I/(Q+1) \gg 1$. Theoretically, the weighted average operation can be considered as an effective approach in estimating LTV channel, where the information sequence interference can be effectively suppressed over multiple OFDM symbols. As stated

in conventional ST-based LTI schemes [8]-[13] [17]-[18], channel estimation performance can be improved along with the increment of the recorded frame length Ω , i.e. the estimation variance approaches to zero as $\Omega \rightarrow \infty$. This can be easily comprehended that larger frame length Ω means more observation samples, and hence lowers the MSE level. From the LTV channel model (6), however, we note that as the frame length Ω is increased, the corresponding truncated DFB requires a larger order Q to model the LTV channel (maintain a tight channel model), and the least order should be satisfied $Q/2 \geq f_d \Omega / f_s$, where f_d and f_s are the Doppler frequency and sampling rate, respectively. Consequently, as the frame length Ω increases, the LTV channel estimation variance (39) approaches to a fixed lower-bound associate with the system Doppler frequency as well as the information to pilot power ratio. This is quite different from the existing ST-based channel estimation approaches [8]-[19].

According to the theoretic analysis in (37)-(39), the proposed two-step LTV channel estimator achieves a significant improvement over multiple OFDM symbols compared with that of block-by-block process (33)-(35). However, as the frame length Ω is increased, the estimation variances approach to a fixed lower-bound. Further enhancement of the channel estimation should resort to increasing the ST power σ_p^2 . For wireless communication systems with a limited transmission power, however, an increased ST power allocation reduces the data power σ_c^2 , leading to SER degradation. Accordingly, in the analysis presented in the next section, the ratio of ST power allocation is determined by maximizing the lower bound of the average channel capacity.

5. Analysis of ST power allocation and system capacity

In this section, we consider the issue of ST power allocation where the lower bound of the average channel capacity is maximized and then mathematically derived for the proposed two-step channel estimator.

Define the ST power allocation factor

$$\beta = \frac{E\left[|p_n(k)|^2\right]}{E\left[|p_n(k)|^2\right] + E\left[|c_n(k)|^2\right]} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_c^2}. \quad (40)$$

For a fixed SNR or transmitted power budget, higher β implies smaller effective SNR at the receiver due to decreased power in the information sequence but higher channel estimation accuracy. Having removed ST sequence, we obtain the received signals in a vector-form as

$$\begin{aligned} \bar{\mathbf{U}}^{(m)}(i) &= \left[\bar{u}^{(m)}(i,0), \dots, \bar{u}^{(m)}(i,k), \dots, \bar{u}^{(m)}(i,B-1) \right]^T \\ &= \underbrace{\sum_{n=1}^N \hat{\mathbf{H}}_n^{(m)}(i) \mathbf{C}_n(i)}_{\text{desired signals: } = \boldsymbol{\lambda}^{(m)}(i)} + \underbrace{\sum_{n=1}^N \Delta \hat{\mathbf{H}}_n^{(m)}(i) [\mathbf{P}_n(i) + \mathbf{C}_n(i)]}_{\text{interference to information signal recovery: } = \boldsymbol{\mu}^{(m)}(i)} + \bar{\mathbf{V}}^{(m)}(i) \end{aligned} \quad (41)$$

with the received signals $\bar{u}^{(m)}(i,k), k = 0, \dots, B-1$ in (41) as

$$\begin{aligned} \bar{u}^{(m)}(i, k) &= \lambda^{(m)}(i, k) + \mu^{(m)}(i, k) + \bar{v}^{(m)}(i, k) \\ &= \sum_{n=1}^N \hat{H}_n^{(m)}(i, k) c_n(i, k) + \sum_{n=1}^N \Delta \hat{H}_n^{(m)}(i, k) [p_n(i, k) + c_n(i, k)] + \bar{v}^{(m)}(i, k) \end{aligned} \quad (42)$$

where $\Delta \hat{H}_n^{(m)}(i, k) = \sum_{n=1}^N [H_n^{(m)}(i, k) - \hat{H}_n^{(m)}(i, k)]$ is the estimation error due to information interference as well as additive noise. Using the proposed two-step estimator (23)-(26), the channel estimation variance can be smoothed over multiple OFDM symbols, and approaches to a small fixed lower bound. The estimated vector $\hat{H}_n^{(m)}(i)$ as well as the error vector $\Delta \hat{H}_n^{(m)}(i)$, therefore, can be thus approximated to be of the similar characteristics of distribution as that of $H_n^{(m)}(i)$. Consequently, following the assumption (H1)-(H3), the interference vector $\mu^{(m)}(i)$ is approximately white for a large symbol-size B , and independent of the noise vector $\bar{V}^{(m)}(i)$. Similar to the procedure of (29)-(30), the covariance matrix of $\mu^{(m)}(i)$ and $\lambda^{(m)}(i)$ can be obtained as

$$\text{var}(\lambda^{(m)}(i)) = E\left\{\left(\lambda^{(m)}(i)\right)^H \lambda^{(m)}(i)\right\} = \sigma_{\hat{H}}^2 \sigma_p^2 \mathbf{I} \quad (43)$$

$$\text{var}(\mu^{(m)}(i)) = E\left\{\left(\mu^{(m)}(i)\right)^H \mu^{(m)}(i)\right\} = \sigma_{\Delta \hat{H}}^2 (\sigma_p^2 + \sigma_c^2) \mathbf{I} \quad (44)$$

where $\sigma_{\hat{H}}^2 = \left|\bar{h}^{(m)}(i)\right|^2 = \sum_{k=0}^{B-1} \sum_{n=1}^N \sum_{l=0}^{L-1} \left|\hat{h}_{n,l}^{(m)} e^{-2\pi k l / B}\right|^2 / NLB$,

and $\sigma_{\Delta \hat{H}}^2 = \sum_{k=0}^{B-1} \sum_{n=1}^N \sum_{l=0}^{L-1} \left|\Delta \hat{h}_{n,l}^{(m)} e^{-2\pi k l / B}\right|^2 / NLB$. Since the ST power allocation factor is derived within each isolated OFDM symbol, we neglect the symbol-index i for simplicity. A lower bound on the OFDM channel capacity with channel estimation error has been derived in [20]-[21] for uniform pilot distribution. Such expression can readily be extended to issue of ST where the pilots are spread over the whole frequency band. Therefore, the lower bound of the average channel capacity for an ST-based OFDM system can be obtained by summing over all the subcarriers, i.e.,

$$C^{(m)} \geq \bar{C}^{(m)} = \frac{1}{B} \sum_{k=0}^{B-1} E \left\{ \log \left[1 + \frac{\sigma_c^2}{(\sigma_p^2 + \sigma_c^2) \sigma_{\Delta \hat{H}}^2 / \sigma_{\hat{H}}^2 + \sigma_v^2 / \sigma_{\hat{H}}^2} \right] \right\} \quad (45)$$

For the sake of simplicity, we assume the transmission power satisfies that $\sigma_p^2 + \sigma_c^2 = 1$. By (40), we thus have $\sigma_p^2 = \beta$ and $\sigma_c^2 = 1 - \beta$. Considering that the normalized MSE of the proposed two-step channel estimator is sufficiently small and approaches to a fixed lower bound (37)-(39), it allows us to make the approximation of $\sigma_{\Delta \hat{H}}^2 / \sigma_{\hat{H}}^2 \approx \sigma_{\Delta \hat{H}}^2 / \sigma_{\hat{H}}^2 = NMSE^{(m)}$. As a result, $C^{(m)}$ in (45) can be approximated as

$$\begin{aligned}
 C^{(m)} &\approx \frac{1}{B} \sum_{k=0}^{B-1} E \left\{ \log \left[1 + \frac{\sigma_c^2}{\sigma_{\Delta\hat{H}}^2 / \sigma_H^2 + \sigma_v^2 / \sigma_H^2} \right] \right\} \\
 &= \frac{1}{B} \sum_{k=0}^{B-1} E \left\{ \log \left[1 + \frac{1 - \beta}{(1 - \beta)(Q + 1)NL / \beta I\Gamma + (Q + 1)\sigma_v^2 / \beta I\Gamma \sigma_H^2 + \sigma_v^2 / \sigma_H^2} \right] \right\} \quad (46) \\
 &= \log \left(1 + \frac{(1 - \beta)\beta I\Gamma}{\beta [I\Gamma / \mathfrak{R}_{\text{SNR}} - (Q + 1)NL] + (Q + 1)NL(1 / \mathfrak{R}_{\text{SNR}} + 1)} \right)
 \end{aligned}$$

where $\mathfrak{R}_{\text{SNR}} = \sigma_H^2 (\sigma_p^2 + \sigma_c^2) / \sigma_v^2 = \sigma_H^2 / \sigma_v^2$. In fact, the averaged channel capacity of (46) is a log-function of β , which is a monotonically increasing function. Therefore, the lower-bound of $C^{(m)}$ with respect to β can be achieved by maximizing the following function

$$\Upsilon^{(m)}(\beta) = \frac{\beta(1 - \beta)}{\alpha_1\beta + \alpha_2} = \frac{(1 - \beta)\beta}{\beta [1 / \mathfrak{R}_{\text{SNR}} - (Q + 1)NL / I\Gamma] + (Q + 1)NL(1 / \mathfrak{R}_{\text{SNR}} + 1) / I\Gamma}. \quad (47)$$

where

$$\alpha_1 = 1 / \mathfrak{R}_{\text{SNR}} - (Q + 1)NL / I\Gamma, \quad \alpha_2 = (Q + 1)NL(1 / \mathfrak{R}_{\text{SNR}} + 1) / I\Gamma. \quad (48)$$

Setting the first derivation of $\Upsilon^{(m)}(\beta)$ with respect to β to zero, we obtain (after some manipulations) a quadratic equation in β , i.e.

$$\beta^2 + \frac{2\alpha_2}{\alpha_1}\beta - \frac{\alpha_2}{\alpha_1} = 0. \quad (49)$$

Consequently, the global maximum of $\Upsilon^{(m)}(\beta)$ can be obtained when

$$\beta = \frac{\sqrt{(1 / \mathfrak{R}_{\text{SNR}} + 1)(I\Gamma / NL(Q + 1)\mathfrak{R}_{\text{SNR}} + 1 / \mathfrak{R}_{\text{SNR}})} - (Q + 1)NL(1 / \mathfrak{R}_{\text{SNR}} + 1)}{I\Gamma / \mathfrak{R}_{\text{SNR}} - (Q + 1)NL}. \quad (50)$$

As will be shown in simulations, an increase in the training power allocation factor β does not necessarily improve the overall system performance since a larger β implies a better channel estimation while substantially scarifying the effective received signal SNR at the same time.

6. Simulations

We assume the MIMO/OFDM system with $N = 2$ and $M = 4$. The symbol-size is $B = 1024$ and the transmitted data $s_n(i, k)$ is 8-PSK signals with symbol rate $f_s = 10^7$ /second. Before transmission, the transmitted data are coded by 1/2 convolutional coding and block interleaving over one OFDM symbol. The channel is assumed to be $L = 10$ taps and, the

coefficients $h_{n,l}^{(m)}(t)$ are generated as low-pass, Gaussian and zero mean random processes and uncorrelated for different values of n and l . The multi-path intensity profile is chosen to be $\phi(l) = \exp(-l/10)$ $l=0, \dots, L-1$. The Doppler spectra are $\Psi(f) = (\pi\sqrt{(f_n)^2 - f^2})$ for $f \leq f_n$, where f_n is the Doppler frequency of the n th user, otherwise, $\Psi(f) = 0$. CP-length is chosen to be 32 to avoid inter-symbol interferences. The additive noise is a Gaussian and white random process with a zero mean.

Test Case 1. Channel Estimation

We run simulations with the Doppler frequency $f_n = 300\text{Hz}$ that corresponds to the maximum mobility speed of 162 km/h as the users operate at carrier frequency of 2GHz. In order to model the LTV channel, the frame is designed as $\Omega = B' \times 128 = (B + \text{CP-length}) \times 128 = 135168$, i.e. each frame consists of 128 OFDM symbols. During the frame, the channel variation is $f_n \Omega / f_s = 4.1$. Over the frame Ω , we utilize truncated DFB of order $Q = 10 > 2f_n \Omega / f_s$ to model the LTV channel coefficients. In order to estimate the MIMO/OFDM channels, the superimposed pilots are designed according to (15) with the pilot power $\sigma_p^2 = 0.2\sigma_c^2$. Fig.2 depicts the LTV channel coefficient estimation over the frame Ω . It is clearly observed that although the channel coefficient is accurately estimated during the centre part of the frame, the outmost samples over the whole frame still exhibit errors. A possible explanation is that as the Fourier basis expansions in (6) are truncated, and an effect similar to the Gibbs phenomenon, together with spectral leakages, will lead to some errors at the beginning and the end of the frame. This may be a common problem for the proceeding literature [1]-[2] [5]-[6] [16] that employing basis expansions to model the LTV channels. To solve the problem, the frames are designed to be partially overlapped, e.g. the frames are designed as $(\ell - 1)\Omega - \Psi B' \leq t \leq \ell\Omega$, $\ell = 2, 3, \dots$, where Ψ is a positive integer. By the frame-overlap, the channel at the beginning and the end of one frame can be modeled and estimated from the neighboring frames.

To further evaluate the new channel estimator, we use the mean square errors to measure the channel estimation performance by

$$MSE_n^{(m)} = \sum_{i=1}^{\Omega/B'} MSE_n^{(m)}(i) / (\Omega/B') = \frac{B' \Omega / B'}{\Omega} \sum_{i=1}^{\Omega/B'} E \left\{ \left| \sum_{t=0}^{B-1} \sum_{l=0}^{L-1} \hat{h}_{n,l}^{(m)}(i, t) - \sum_{q=0}^Q \hat{h}_{n,l,q}^{(m)} e^{j2\pi(q-Q/2)t/\Omega} \right|^2 / BL \left| h_{n,l}^{(m)}(i, t) \right|^2 \right\} \quad (51)$$

where $\hat{h}_{n,l,q}^{(m)}$ is the channel coefficient estimation.

We firstly test the two-step channel estimator under the different pilot powers and different channel coefficient numbers to verify the channel estimation variance analysis. The LTV channel is the same as that in Fig.2. As shown in Fig.3, the MSE of the channel estimation approach are almost independent of the additive noises, especially as $\text{SNR} > 5\text{dB}$. This is consistent with the channel estimation analysis (38)-(39) where the additive noise has been

greatly suppressed by the weighted average procedure. Thus, the estimation errors depend mainly on the information- pilot power ratio as well as the system unknowns NL . This is rather different from the FDM training based schemes [20]-[23].

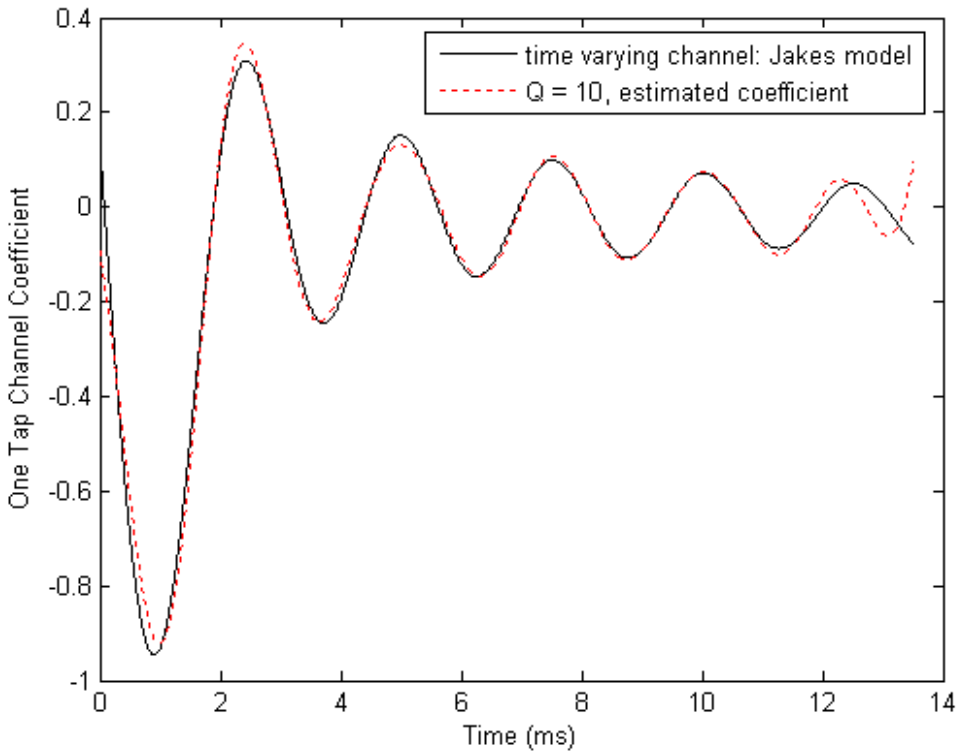


Fig. 2. One tap coefficient of the LTV channel and the estimation over the frame $\Omega = 135168/10^7 \approx 13.52$ ms.

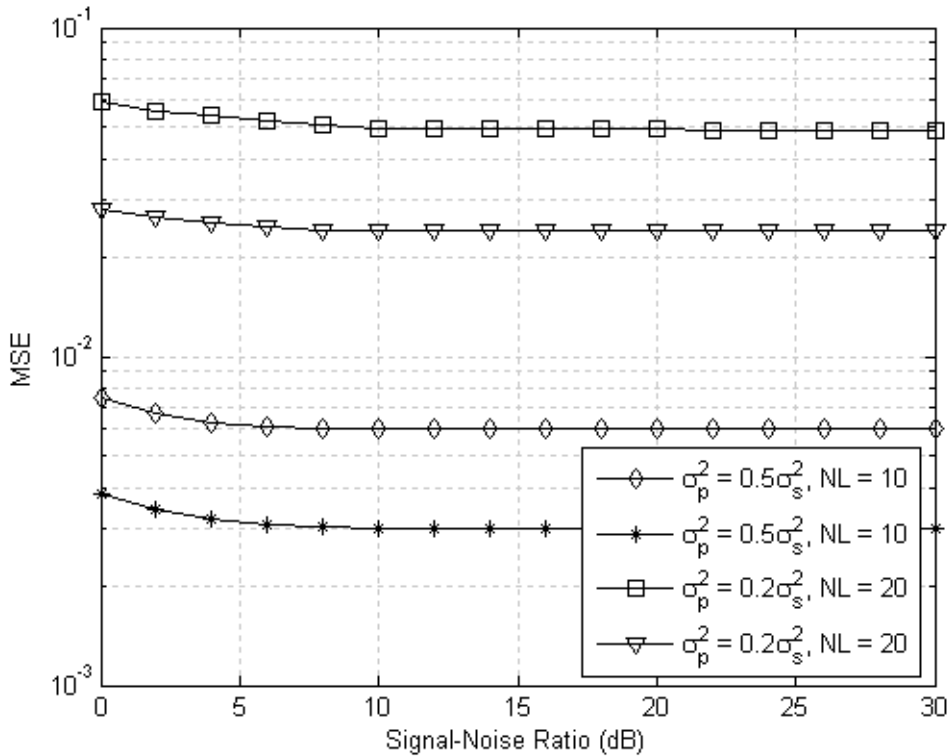


Fig. 3. MSE of the weighted average estimation versus SNR for the LTV channel of $f_n = 300\text{Hz}$ and $\Omega = 13.52\text{ ms}$ under the different pilot powers and different unknown parameters.

We then compare the proposed two-step channel estimation scheme with the conventional ST-based methods [8]-[13] [17]-[18] under different Doppler frequencies. In the conventional ST scheme, the LTV channel is firstly estimated from the LTI assumption at each OFDM symbol, and then all the estimations from the frame Ω are averaged to confront the information sequence interferences. It shows clearly in Fig. 4 that for the LTI channel of $f_n = 0\text{Hz}$, both the conventional ST and the weighted average estimator exhibit the similar performance. In addition, the estimation performance can be improved with the increment of the frame or average length. However, when the channel involved in simulations is time-varying, the channel estimation performance of the conventional ST-based schemes is degraded extensively. The simulation reveals the shortcoming of the conventional ST in estimating the LTV channels. On the contrary, the MSE level is reduced by the weighted average process (23)-(26) for the LTV channels of $f_n = 100\text{Hz}$, 300Hz with $T = 2$ (one pilot cluster is composed of $2T + 1 = 5$ pilots). We also observe that the MSE approaches to a constant as the increment of the frame length, i.e. the lower-bound that associated with the given Doppler frequency.

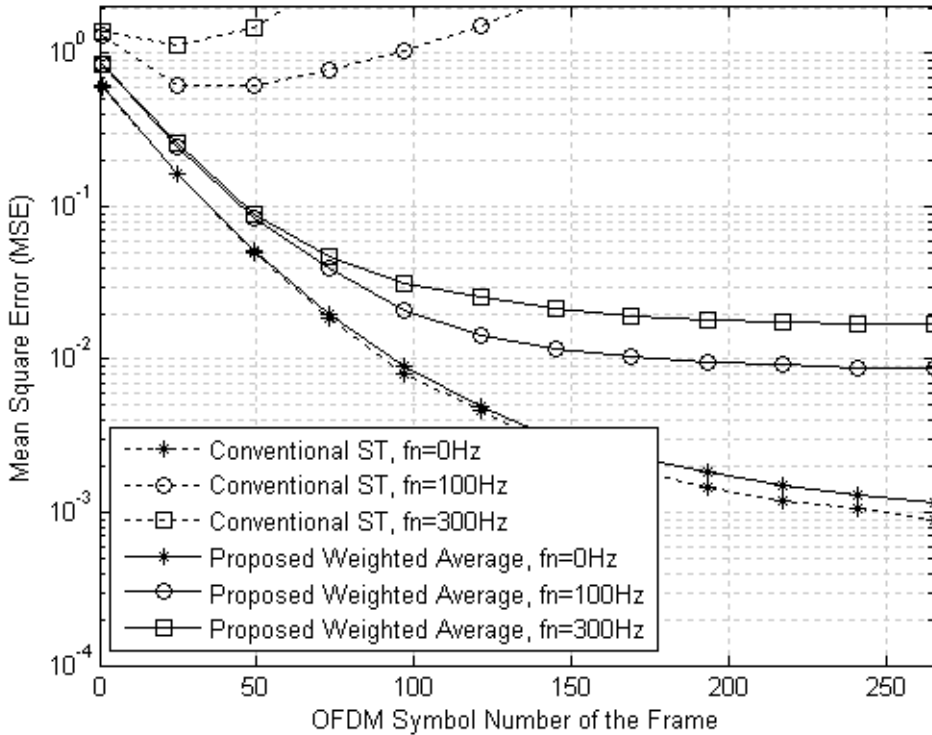


Fig. 4. MSE versus frame or average length under the different Doppler frequencies of the LTV channel with $\sigma_p^2 = 0.25\sigma_c^2$, SNR = 20dB.

From Fig. 4, we observe that channel estimation performance would be degraded as the increment of mobile users' speed (or corresponding system Doppler shift). To further enhance the channel estimation performance of the systems with a limited pilot power while suffering from a high Doppler shift, an iterative decision feedback (DF) approach can be adopted at the receiver. Explicitly, the iterative method can be considered as a twofold process. First, the information sequences are recovered by a hard detector [5] based on the LTV channel estimation in Section III. Second, the recovered data symbols are removed from the received signals to cancel the information sequence interference and, thus to enhance the channel estimation performance. Fig. 5 depicts the performance between the weighted average scheme and the iterative DF estimator in terms of channel MSE. For a fairness of comparison, we also simulate the MSE of the FDM training-based channel estimator [5] as latter serves as a "benchmark" in related works. For estimating the MIMO/OFDM channels, $L = 40$ pilot clusters with $L(2T + 1) = 200$ known pilot symbols which are subject to the

proposed pilot specifications in (15) are used in one OFDM symbol. That is, approximately 10% total bandwidth is assigned for pilot tones. Comparatively, as shown in Fig.5, the iterative DF estimation exhibits a more significant improvement than that of weighted average estimation, and outperforms the FDM channel estimator [5] by using a small pilot power of $\sigma_p^2 = 0.25\sigma_c^2$, which conforms that the information sequence interferences can be effectively cancelled by iterative DF procedure. Moreover, it should be noted that since the superimposed pilots are spread over the entire band, the proposed ST-based channel estimator is also feasible to estimate the channel with a very long delay spread, i.e. cluster-based channel.

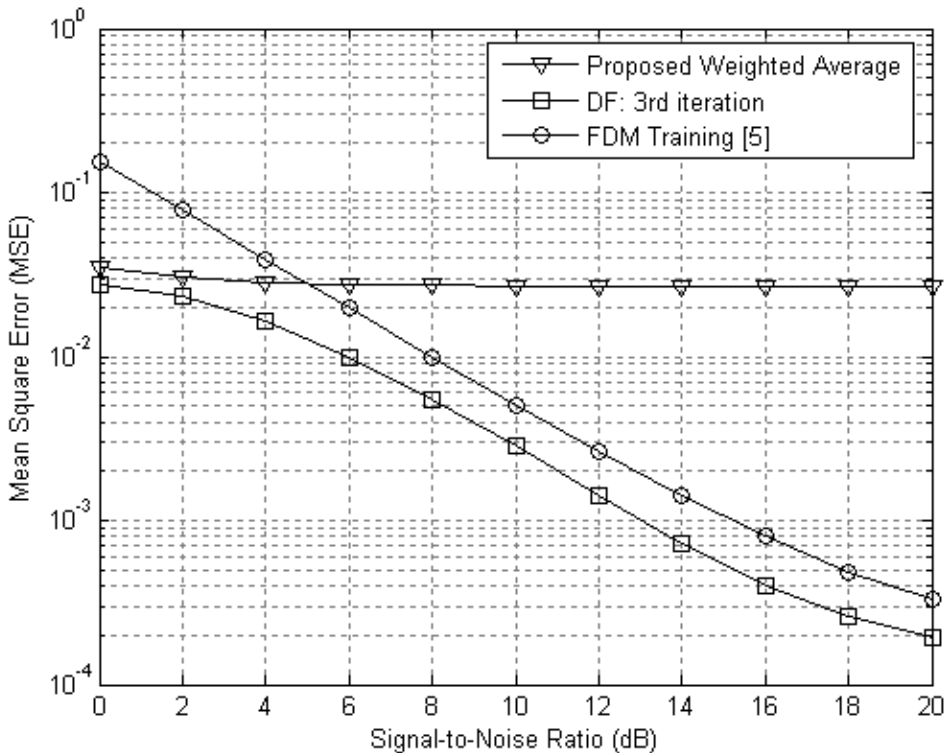


Fig. 5. MSE versus SNR for different estimators for the LTV channel of $f_n = 300$ Hz, $NL = 20$.

To further validate the effectiveness of the DF scheme, we also provide the channel estimation MSE of the DF method versus the iteration numbers under SNR = 15dB. Fig. 6 shows that the iterative DF method is feasible for a wide range of system Doppler spreads. Obviously, the enhancement of the iterative DF is at the cost of an increment in computational complexity that is directly proportional to the iteration number. However, as is shown in Fig. 6 that the iterative DF approach converges to the steady-state performance by only a few iterations, the overall computational complexity will be acceptable for many wireless communication systems.

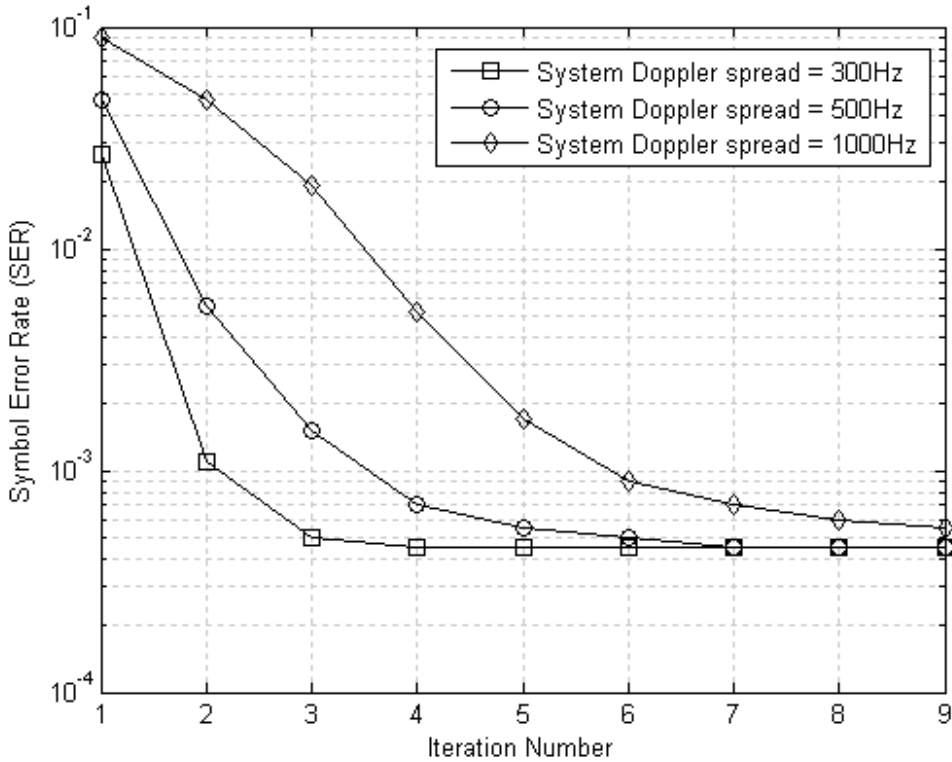


Fig. 6. MSE of the iterative DF channel estimation versus iterations for the LTV channel of different system Doppler spreads when SNR=15dB.

Test Case 2. Training Power Allocation

As aforementioned, for wireless communication systems with a limited transmission power, some useful power must inevitably be allocated to the superimposed pilots, and thus resulting in the received signal SNR reduction. Herein, we carry out several experiments to assess the effect of ST power allocation factor on the lower-bound of the average channel capacity for different SNRs.

Fig. 7 shows the effect of different value of training power allocation factor β on the lower bound of the average channel capacity for received signal SNR = 10 and 20 dB, respectively. It is seen that the average channel capacity decreases with the increment of β . It reveals that although higher β implies that higher fraction of transmitted power is allocated to training leading to more accurate channel estimates, the received signal SNR is substantially decreased, resulting in potential decrement of the average channel capacity. In addition, we further simulate the approximated β in order to test the validity of theoretic results in (50). It can be seen that the approximation of β is almost consistent with that of the actual results.

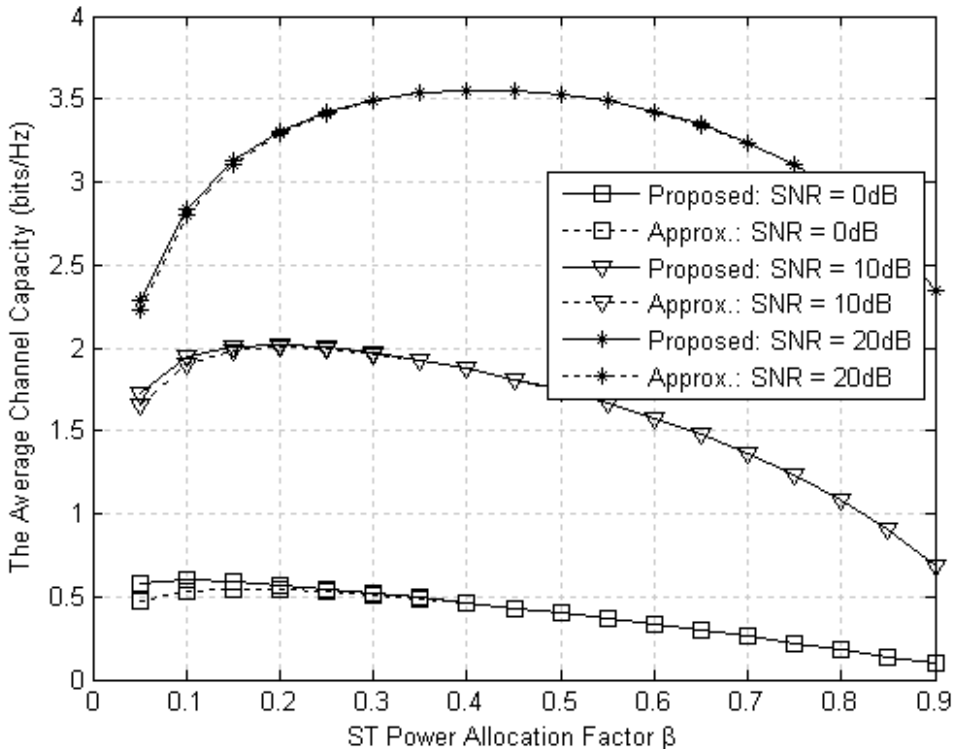


Fig. 7. Lower bound of the average channel capacity versus different values of ST power allocation factor β under SNR = 0dB, 10dB and 20dB, respectively.

Fig. 8 shows the plots of the optimal value of training power allocation factor β versus received SNR for different frame length. It is observed that the increment of SNR leading to a corresponding increase in the optimal ST power allocation factor. This can be easily comprehended that according to (41), the effective interference is composed of two factors, i.e. the bias of channel estimation and the additive noise. That is, for large SNRs, higher β is required to improve the channel estimation performance, thus leading to a reduction of the effective interference. Conversely, when SNR is small, improving the channel estimation accuracy has a small effect in reducing the effective interference. On the other hand, we notice that β decreases as the frame length increases but approximately unvaried when Ω is sufficiently large, i.e. β is almost unchanged when $l \geq 192$. This result arises because we have theoretically analyzed in Section III that the estimation variance approaches to a fixed lower bound that can be only improved by increasing ST power allocation when the frame length is large enough. Therefore, the power allocated to the training sequence can be reduced with no loss in channel estimation performance when the frame length is increased, but finally approaches to a fixed lower bound associate with the channel estimation variance when Ω is sufficiently large. This is somewhat different from those presented in [10].

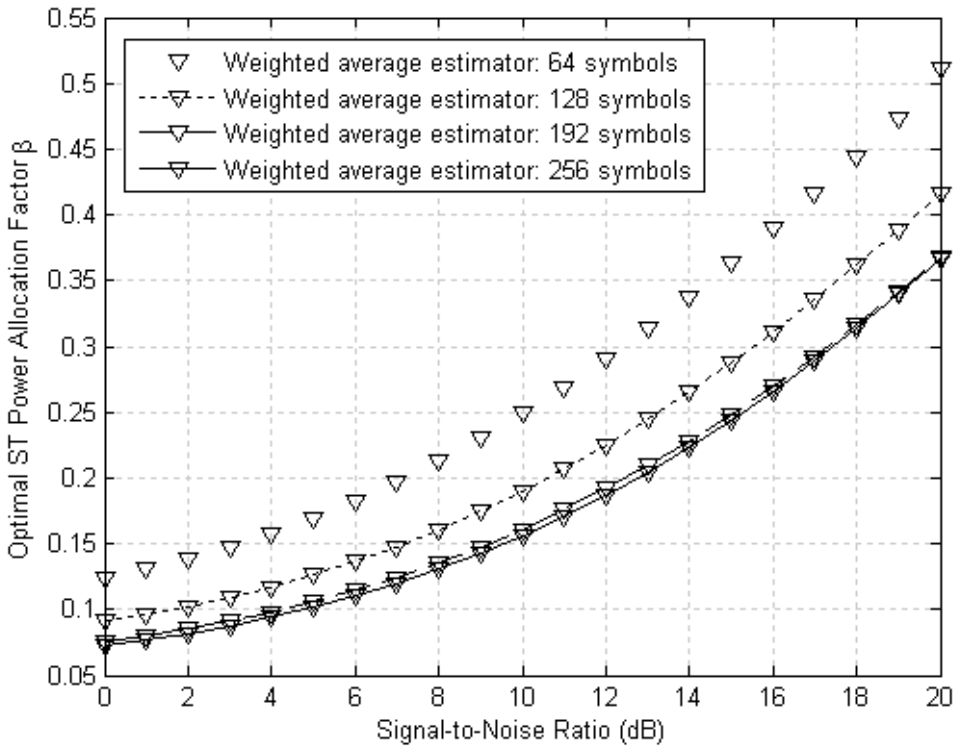


Fig. 8. Optimal ST power allocation factor of the proposed weighted average channel estimator versus SNR for different frame lengths.

7. Conclusions

In this paper, we have developed superimposed training-aided LTV channel estimation approach for MIMO/OFDM systems. The LTV channel coefficients were firstly modeled by truncated DFB, and then estimated by using a two-step approach over multiple OFDM symbols. We also present a performance analysis of the proposed estimation approach and derive closed-form expressions for the channel estimation variances. It is shown that the estimation variances, unlike the conventional ST, approach to a fixed lower-bound that can only be reduced by increasing the pilot power. Using the developed channel estimation variance expression, we analyzed the system capacity and optimize the training power allocation by maximizing the lower bound of the average channel capacity for systems with a limited power. Compared with the existing FDM training based schemes, the new estimator does not entail a loss of rate while yields a better estimation performance, and thus enables a higher efficiency.

8. Acknowledgment

This work is supported by the national Natural Science Foundation (NSF) of China, (Grant No. 61002012), the Project of NSF of Guangdong Province, (Grant No. 10451063101006074), and also supported by national innovation experiment program for university students, Grant No.C1025338, and Key project of college students' extracurricular science and technology, Grant No. 10WDKB05

9. Reference

- [1] G. B. Giannakis, C. Tepedelenlioglu "Basis expansion models and diversity techniques for blind identification and equalization of time-varying channels," *Proc. of the IEEE*, vol.86, pp. 1969-1986, Oct, 1998.
- [2] X. L. Ma, G. B. Giannakis and B. Lu "Block differential encoding for rapidly fading channels," *IEEE Trans. Commun.*, vol. 52, no. 3, pp.416-425, March 2004.
- [3] H.-C. Wu, "Analysis and characterization of intercarrier and interblock interferences for wireless mobile OFDM systems," *IEEE Trans. Broadcasting*, vol. 52, no. 2, pp. 203-210, Jun. 2006.
- [4] X. Dai "Adaptive blind source separation of multiple-input multiple-output linearly time-varying FIR system," *IEE Proc.-Vis. Image Signal Process.*, Vol. 151, No. 4, pp.279-286, August 2004.
- [5] Z. Tang, R. C. Cannizzaro, G. Leus and P. Banelli, "Pilot-assisted time-varying channel estimation for OFDM systems," *IEEE Trans. Signal Process.* Vol. 55, no. 5, pp. 2226-2238, May 2007.
- [6] X. Dai, "Optimal estimation of linearly time-varying MIMO/OFDM channels modeled by a complex exponential basis expansion," *IET Commun.*, vol. 1, no. 5, pp. 945-953, 2007.
- [7] H. C. Wu and Y. Wu, "Distributive pilot arrangement based on modified M-sequences for OFDM intercarrier interference estimation," *IEEE Trans. Wireless Commun.*, vol. 5, no. 6, pp. 1605-1609, May 2007.
- [8] G. T. Zhou, M. Viberg, and T. McKelvey, "A first-order statistical method for channel estimation," *IEEE Signal Processing Letters*, vol. 10, no. 3, pp.57-60, March 2003.
- [9] J. K. Tugnait and W. Luo "On channel estimation using superimposed training and first-order statistics," *IEEE Comm. Letters*, vol. 7, no. 9, pp. 413-416, Sept. 2003.
- [10] J. K. Tugnait and X.H. Meng "On superimposed training for channel Estimation: performance analysis, training power allocation, and frame synchronization," *IEEE Trans. Signal Process.*, vol. 54, no.2, pp.752-765, Feb. 2006.
- [11] M. Ghogho, D. McLernon, E. A. Hernandez, and A. Swami "Channel estimation and symbol detection for block transmission using data-dependent superimposed training," *IEEE Signal Process. Letters*, vol. 12, no. 3, pp. 226-229, March 2005.
- [12] W. C. Huang, C. P. Li and H. J. Li, "On the power allocation and system capacity of OFDM systems using superimposed training schemes," *IEEE Trans. Veh. Technol.*, vo. 58, no. 4, pp. 1731-1740, May, 2009.

- [13] Q. Y. Zhu and Z.Q. Liu, "Optimal pilot superimposition for zero-padded block transmissions," *IEEE Trans. Wireless Commun.*, vol. 5, no. 8, pp. 2194-2201, Aug. 2006.
- [14] H. Zhang, X. Dai, "Time-varying Channel Estimation and Symbol Detection for OFDM Systems using Superimposed Training," *IEE Electronics Letter*, vol. 43, no. 22, Oct. 2007.
- [15] S. M. A. Moosvi, D. C. McLernon, A. G. Orozco-Lugo, M. M. Lara, and M. Ghogho, "Carrier frequency offset estimation using data dependent superimposed training," *IEEE Commun. Letter*, vol. 12, no. 3, pp. 179-181, Mar. 2008.
- [16] R. C.-Alvarez, R. P. -Michel, A. G. O. -Lugo, and J. K. Tugnait, "Enhanced channel estimation using superimposed training based on universal basis expansion," *IEEE Trans. Signal Process.* vol. 57, no. 3, Mar. 2009.
- [17] F. Mazzenga, "Channel estimation and equalization for M-QAM transmission with a hidden pilot Sequence," *IEEE Trans. Broadcasting*, vol. 46, no. 2, pp. 170-176, June 2000.
- [18] A. Goljahani, N. Benvenuto, S. Tomasin and L. Vangelista, "Superimposed sequence versus pilot aided channel estimations for next generation DVB-T systems," *IEEE Trans. Broadcasting*, vol. 55, no. 1, pp. 140-144, Mar. 2009.
- [19] C. -P. Li and W. Hu, "Super-imposed training scheme for timing and frequency synchronization for OFDM systems," *IEEE trans. Broadcasting*, vol. 53, no. 2, pp. 574-583, Jun. 2007.
- [20] A. Y. Panah, R. G. Vaughan and R. W. Heath, Jr. , "Optimizing pilot locations using feedback in OFDM systems," *IEEE trans. Vehicular Tech.* , vol. 58, no. 6, pp. 2803-2814, July 2009.
- [21] S. Ohno and G.. B. Giannakis, "Capacity maximizing MMSE-optimal pilots for wireless OFDM over frequency- selective block Rayleigh-fading channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 2138-2145, Sep. 2004.
- [22] I. Barhumi, G. Leus and M. Moonen "Optimal training design for MIMO OFDM systems in mobile wireless channels," *IEEE Trans. Signal Process.*, vol. 51, no. 6, pp.1615-1624, June 2003.
- [23] Y. Xiaoyan, W. Jiaqing, Y. Luxi and H. Zhenya, "Doubly selective fading channel estimation in MIMO OFDM systems," *SCIENCE CHINA Information Sciences*, vol. 48, no. 6, pp. 795-807, 2005.
- [24] J. Z. Wang and J. Chen, "Performance of wideband CDMA with complex spreading and imperfect channel estimation," *IEEE Journal on Selected Areas in Commun.*, vol. 19, no. 1, pp. 152-163, Jan. 2001.
- [25] J. Z. Wang and L. B. Milstein, "CDMA overlay situations for microcellular mobile communications," *IEEE Trans. on Commun.*, vol. 43, no. 2/3/4, pp. 603-614, Feb/March/April 1995.
- [26] X. Dai, H. Zhang and D. Li, 'Linearly time-varying channel estimation for MIMO-OFDM systems using superimposed training,' *IEEE trans. Commun.*, vol. 58, no. 2, pp. 681-693, Feb. 2011.

- [27] H. Zhang, et. al., 'Linearly time-varying channel estimation and symbol detection for OFDMA uplink using superimposed training,' *EURASIP J. Wireless Commun. Networking*, vol. 2009.
- [28] H. Zhang, X. Dai and D. Li, 'Time-varying channel estimation for MIMO/OFDM systems using superimposed training,' in *Proc. IEEE WiCOM 08*, pp. 1-6, 2008.

Channel Capacity Analysis Under Various Adaptation Policies and Diversity Techniques over Fading Channels

Mihajlo Stefanović¹, Jelena Anastasov¹, Stefan Panić²,
Petar Spalević³ and Ćemal Dolićanin³

¹*Faculty of Electronic Engineering, University of Niš,*

²*Faculty of Natural Science and Mathematics, University of Priština,*

³*State University of Novi Pazar
Serbia*

1. Introduction

The lack of available spectrum for expansion of wireless services requires more spectrally efficient communication in order to meet the consumer demand. Since the demand for wireless communication services have been growing in recent years at a rapid pace, conserving, sharing and using bandwidth efficiently is of primary concern in future wireless communications systems. Therefore, channel capacity is one of the most important concerns in the design of wireless systems, as it determines the maximum attainable throughput of the system [1]. It can be defined as the average transmitted data rate per unit bandwidth, for a specified average transmit power, and specified level of received outage or bit-error rate [2]. Skilful combination of bandwidth efficient coding and modulation schemes can be used to achieve higher channel capacities per unit bandwidth. However, mobile radio links are, due to the combination of randomly delayed reflected, scattered, and diffracted signal components, subjected to severe multipath fading, which leads to serious degradation in the link signal-to-noise ratio (SNR). An effective scheme that can be used to overcome fading influence is adaptive transmission. The performance of adaptation schemes is further improved by combining them with space diversity, since diversity combining is a powerful technique that can be used to combat fading in wireless systems resulting in improving link performance [3].

1.1 Channel and system model

Diversity combining is a powerful technique that can be used to combat fading in wireless systems [4]. The optimal diversity combining technique is maximum ratio combining (MRC). This combining technique involves co-phasing of the useful signal in all branches, multiplication of the received signal in each branch by a weight factor that is proportional to the estimated ratio of the envelope and the power of that particular signal and summing of the received signals from all antennas. By co-phasing, all the random phase fluctuations of the signal that emerged during the transmission are eliminated. For this

process it is necessary to estimate the phase of the received signal, so this technique requires the entire amount of the channel state information (CSI) of the received signal, and separate receiver chain for each branch of the diversity system, which increases the complexity of the system [5].

One of the least complicated combining methods is selection combining (SC). While other combining techniques require all or some of the amount of the CSI of received signal and separate receiver chain for each branch of the diversity system which increase its complexity, selection combining (SC) receiver process only one of the diversity branches, and is much simpler for practical realization, in opposition to these combining techniques [4-7]. Generally, SC selects the branch with the highest SNR, that is the branch with the strongest signal, assuming that noise power is equally distributed over branches. Since receiver diversity mitigates the impact of fading, the question is whether it also increases the capacity of a fading channel.

Another effective scheme that can be used to overcome fading influence is adaptive transmission. Adaptive transmission is based on the receiver's estimation of the channel and feedback of the CSI to the transmitter. The transmitter then adapts the transmit power level, symbol/bit rate, constellation size, coding rate/scheme or any combination of these parameters in response to the changing channel conditions [8]. Adapting certain parameters of the transmitted signal to the fading channel can help better utilization of the channel capacity. These transmissions provide a much higher channel capacities per unit bandwidth by taking advantage of favorable propagation conditions: transmitting at high speeds under favorable channel conditions and responding to channel degradation through a smooth reduction of their data throughput. The source may transmit faster and/or at a higher power under good channel conditions and slower and/or at a reduced power under poor conditions. A reliable feedback path between that estimator and the transmitter and accurate channel estimation at the receiver is required for achieving good performances of adaptive transmission. Widely accepted adaptation policies include optimal power and rate adaptation (OPRA), constant power with optimal rate adaptation (ORA), channel inversion with fixed rate (CIFR), and truncated CIFR (TIFR). Results obtained for this protocols show the trade-off between capacity and complexity. The adaptive policy with transmitter and receiver side information requires more complexity in the transmitter (and it typically also requires a feedback path between the receiver and transmitter to obtain the side information). However, the decoder in the receiver is relatively simple. The non-adaptive policy has a relatively simple transmission scheme, but its code design must use the channel correlation statistics (often unknown), and the decoder complexity is proportional to the channel decorrelation time. The channel inversion and truncated inversion policies use codes designed for additive white Gaussian noise (AWGN) channels, and are therefore the least complex to implement, but in severe fading conditions they exhibit large capacity losses relative to the other techniques.

The performance of adaptation schemes is further improved by combining them with space diversity. The hypothesis that the variation of the combiner output SNR is tracked perfectly by the receiver and that the variation in SNR is sent back to the transmitter via an error-free feedback path will be assumed in the ongoing analysis [8]. Also, it is assumed that time delay in this feedback path is negligible compared to the rate of the channel variation.

Following these assumptions, transmitter could adapt its power and/or rate relative to the actual channel state.

There are numerous published papers based on study of channel capacity evaluation. In [9], the capacity of Rayleigh fading channels under four adaptation policies and multibranch system with variable correlation is investigated. The capacity of Rayleigh fading channels under different adaptive transmission and different diversity combining techniques is also studied in [7], [10]. In [11], channel capacity of MRC over exponentially correlated Nakagami- m fading channels under adaptive transmission is analyzed. Channel capacity of adaptive transmission schemes using equal gain combining (EGC) receiver over Hoyt fading channels is presented in [12]. In [13], dual-branch SC receivers operating over correlative Weibull fading under three adaptation policies are analyzed.

In this chapter we will focus on more general and nonlinear fading distributions. We will perform an analytical study of the κ - μ fading channel capacity, e.g., under the OPRA, ORA, CIFR and TIFR adaptation policies and MRC and SC diversity techniques. To the best of authors' knowledge, such a study has not been previously considered in the open technical literature. The expressions for the proposed adaptation policies and diversity techniques will be derived. Capitalizing on them, numerically obtained results will be graphically presented, in order to show the effects of various system parameters, such as diversity order and fading severity on observed performances. In the similar manner an analytical study of the Weibull fading channel capacity, under the OPRA, ORA, CIFR and TIFR adaptation policies and MRC diversity technique will be performed.

1.1.1 κ - μ channel and system model

The multipath fading in wireless communications is modelled by several distributions including Nakagami- m , Hoyt, Rayleigh, and Rice. By considering important phenomena inherent to radio propagation, κ - μ fading model was recently proposed in [14] as a fading model which describes the short-term signal variation in the presence of line-of-sight (LOS) components. This distribution is more realistic than other special distributions, since its derivation is completely based on a non-homogeneous scattering environment. Also κ - μ as general physical fading model includes Rayleigh, Rician, and Nakagami- m fading models, as special cases [14]. It is written in terms of two physical parameters, κ and μ . The parameter κ is related to the multipath clustering and the parameter μ is the ratio between the total power of the dominant components and the total power of the scattered waves. In the case of $\kappa=0$, the κ - μ distribution is equivalent to the Nakagami- m distribution. When $\mu=1$, the κ - μ distribution becomes the Rician distribution with κ as the Rice factor. Moreover, the κ - μ distribution fully describes the characteristics of the fading signal in terms of measurable physical parameters.

The SNR in a κ - μ fading channel follows the probability density function (pdf) given by [15]:

$$P_{\gamma}(\gamma) = \frac{\mu}{k^{(\mu-1)/2} e^{\mu k}} \left(\frac{1+k}{\gamma} \right)^{(\mu+1)/2} \gamma^{(\mu-1)/2} e^{-\mu(1+k)\gamma/\bar{\gamma}} I_{\mu-1} \left(2\mu \sqrt{\frac{(1+k)k\gamma}{\bar{\gamma}}} \right). \quad (1.1)$$

In the previous equation, $\bar{\gamma}$ is the corresponding average SNR, while $I_n(x)$ denotes the n -th order modified Bessel function of first kind [16], and κ and μ are well-known κ - μ fading parameters. Using the series representation of Bessel function [16, eq. 8.445]:

$$I_n(x) = \sum_{k=0}^{+\infty} \frac{x^{2k+n}}{2^{2k+n} \Gamma(k+n+1) k!}, \quad (1.2)$$

the cumulative distribution function (cdf) of γ can be written in the form of:

$$F_\gamma(\gamma) = \sum_{p=0}^{+\infty} \frac{\mu^p \kappa^p}{e^{\mu\kappa} \Gamma(p+\mu)} \Lambda \left(p + \mu, \frac{\mu(1+\kappa)\gamma}{\bar{\gamma}} \right) \quad (1.3)$$

with $\Gamma(x)$ and $\Lambda(a,x)$ denoting Gamma and lower incomplete Gamma function, respectively [16, eqs. 8.310.1, 8.350.1].

It is shown in [15], that the sum of κ - μ squares is κ - μ square as well (but with different parameters), which is an ideal choice for MRC analysis. Then the expression for the pdf of the outputs of MRC diversity systems follows [15, eq.11]:

$$p_\gamma^{MRC}(\gamma) = \frac{L\mu}{k^{(L\mu-1)/2} e^{L\mu k}} \left(\frac{1+k}{L\bar{\gamma}} \right)^{(L\mu+1)/2} \gamma^{(L\mu-1)/2} e^{-\mu(1+k)\gamma/\Omega} I_{L\mu-1} \left(2\mu L \sqrt{\frac{(1+k)k\gamma}{L\bar{\gamma}}} \right) \quad (1.4)$$

with L denoting the number of diversity branches.

The expression for the pdf of the outputs of SC diversity systems can be obtained by substituting expressions (1.1) and (1.3) into:

$$p_\gamma^{SC}(\gamma) = \sum_{i=1}^L p_{\gamma_i}(\gamma) \prod_{\substack{j=1 \\ j \neq i}}^L F_{\gamma_j}(\gamma) \quad (1.5)$$

where $p_{\gamma_i}(\gamma)$ and $F_{\gamma_i}(\gamma)$ define pdf and cdf of SNR at input branches respectively and L denotes the number of diversity branches.

1.1.2 Weibull channel and system model

The above mentioned well-known fading distributions are derived assuming a homogeneous diffuse scattering field, resulting from randomly distributed point scatterers. The assumption of a homogeneous diffuse scattering field is certainly an approximation, because the surfaces are spatially correlated characterizing a nonlinear environment. With the aim to explore the nonlinearity of the propagation medium, a general fading distribution, the Weibull distribution, was proposed. The nonlinearity is manifested in terms of a power parameter $\beta > 0$, such that the resulting signal intensity is obtained not simply as the modulus of the multipath component, but as the modulus to a certain given power. As β increases, the fading severity decreases, while for the special case of $\beta = 2$ reduces to the

well-known Rayleigh distribution. Weibull distribution seems to exhibit good fit to experimental fading channel measurements, for both indoor and outdoor environments.

The SNR in a Weibull fading channel follows the pdf given by [17, eq.14]:

$$p(\gamma) = \frac{\beta}{2a\bar{\gamma}} \left(\frac{\gamma}{a\bar{\gamma}} \right)^{\frac{\beta}{2}-1} e^{-\left(\frac{\gamma}{a\bar{\gamma}}\right)^{\beta/2}} \tag{1.6}$$

In the previous equation, $\bar{\gamma}$ is the corresponding average SNR, β is well-known Weibull fading parameter, and $a = 1/\Gamma(1+2/\beta)$.

It is shown in [18,19], that the expression for the pdf of the outputs of MRC diversity systems follows [19, eq.1]:

$$p_{\gamma}^{MRC}(\gamma) = \frac{\beta\gamma^{L\beta/2-1}}{2\Gamma(L)(\Xi\bar{\gamma})^{L\beta/2}} e^{-\left(\frac{\gamma}{\Xi\bar{\gamma}}\right)^{\beta/2}}; \quad \Xi = \frac{\Gamma(L)}{\Gamma(L+2/\beta)} \tag{1.7}$$

with L denoting the number of diversity branches.

Similarly, expression for the pdf of the outputs of SC diversity systems can be obtained as (1.5)

2. Optimal power and rate adaptation

In the OPRA protocol the power level and rate parameters vary in response to the changing channel conditions. It achieves the ergodic capacity of the system, i. e. the maximum achievable average rate by use of adaptive transmission. However, OPRA is not suitable for all applications because for some of them it requires fixed rate.

During our analysis it is assumed that the variation in the combined output SNR over κ - μ fading channels γ is tracked perfectly by the receiver and that variation of γ is sent back to the transmitter via an error-free feedback path. Comparing to the rate of channel variation, the time delay in this feedback is negligible. These assumptions allow the transmitter to adopt its power and rate correspondingly to the actual channel state. Channel capacity of the fading channel with received SNR distribution, $p_{\gamma}(\gamma)$, under optimal power and rate adaptation policy, for the case of constant average transmit power is given by [8]:

$$\langle C \rangle_{pra} = B \int_{\gamma_0}^{\infty} \log_2 \left(\frac{\gamma}{\gamma_0} \right) p_{\gamma}(\gamma) d\gamma, \tag{1.8}$$

where B (Hz) denotes the channel bandwidth and γ_0 is the SNR cut-off level below which transmission of data is suspended. This cut-off level must satisfy the following equation:

$$\int_{\gamma_0}^{\infty} \left(\frac{1}{\gamma_0} - \frac{1}{\gamma} \right) p_{\gamma}(\gamma) d\gamma = 1, \tag{1.9}$$

Since no data is sent when $\gamma < \gamma_0$, the optimal policy suffers a probability of outage P_{out} equal to the probability of no transmission, given by:

$$P_{out} = \int_0^{\gamma_0} p_\gamma(\gamma) d\gamma = 1 - \int_{\gamma_0}^{\infty} p_\gamma(\gamma) d\gamma \tag{1.10}$$

2.1 κ - μ fading channels

To achieve the capacity in (1.8), the channel fading level must be attended at the receiver as well as at the transmitter. The transmitter has to adapt its power and rate to the actual channel state; when γ is large, high power levels and rates are allocated for good channel conditions and lower power levels and rates for unfavourable channel conditions when γ is small. Substituting (1.1) into (1.9), we found that the cut-off level must satisfy:

$$\sum_{i=0}^{\infty} \frac{(kL\mu)^i}{e^{L\mu k} \Gamma(i+L\mu) i!} \left(\frac{1}{\gamma_0} \Lambda \left(L\mu + i, \frac{\mu(1+k)\gamma_0}{\bar{\gamma}} \right) - \frac{\mu(1+k)}{\bar{\gamma}} \Lambda \left(L\mu + i - 1, \frac{\mu(1+k)\gamma_0}{\bar{\gamma}} \right) \right) - 1 = 0 \tag{1.11}$$

Substituting (1.1) into (1.8), we obtain the capacity per unit bandwidth, $\langle C \rangle_{opra}/B$, as:

$$\frac{\langle C \rangle_{opra}^{MRC}}{B} = \sum_{i=0}^{\infty} \frac{L\mu}{k^{(L\mu-1)/2} e^{L\mu k}} \left(\frac{1+k}{L\bar{\gamma}} \right)^{(L\mu+1)/2} \int_{\gamma_0}^{\infty} \log_2 \left(\frac{\gamma}{\gamma_0} \right) \gamma^{L\mu+i-1} e^{-\mu(1+k)\gamma/\bar{\gamma}} d\gamma \tag{1.12}$$

Now, by making change of variables, $\langle C \rangle_{opra}/B$ can be obtained as:

$$\frac{\langle C \rangle_{opra}^{MRC}}{B} = \sum_{i=0}^{\infty} \frac{(L\mu k)^i}{\Gamma(i+L\mu) i! e^{L\mu k}} \left(\int_0^{\infty} \log_2 \left(\frac{t\bar{\gamma}}{\mu(1+k)\gamma_0} \right) t^{L\mu+i-1} e^{-t} dt - \int_0^{\gamma_0 \mu(1+k)/\bar{\gamma}} \log_2 \left(\frac{t\bar{\gamma}}{\mu(1+k)\gamma_0} \right) t^{L\mu+i-1} e^{-t} dt \right) = \sum_{i=0}^{\infty} \frac{(L\mu k)^i}{\Gamma(i+L\mu) i! e^{L\mu k}} (I_1 - I_2) \tag{1.13}$$

Integral I_1 can be solved by applying Gauss-Laguerre quadrature formulae:

$$I_1 = \int_0^{\infty} f_1(t) e^{-t} dt \cong \sum_{k=1}^R A_k f_1(t_k); \quad f_1(t) = \log_2 \left(\frac{t\bar{\gamma}}{\mu(1+k)\gamma_0} \right) t^{L\mu+i-1} \tag{1.14}$$

In the previous equation A_k and $t_k, k=1,2,\dots,R$, are respectively weights and nodes of Laguerre polynomials [20, pp. 875-924].

Similarly, integral I_2 can be solved by applying Gauss-Legendre quadrature formulae:

$$I_2 = \left(\frac{\gamma_0 \mu (1+k)}{2\gamma} \right)^{L\mu+i} \int_{-1}^1 f_2(u) du \cong \left(\frac{\gamma_0 \mu (1+k)}{2\gamma} \right)^{L\mu+i} \sum_{k=1}^R B_k f_2(u_k) \tag{1.15}$$

where B_k and $u_k, k=1,2,\dots,R$, are respectively weights and nodes of Legendre polynomials.

Convergence of infinite series expressions in (1.13) is rapid since we need about 10 terms to be summed in order to achieve accuracy at the 5th significant digit for corresponding values of system parameters.

2.2 Weibull fading channels

Substituting (1.7) in (1.8) integral of the following form need to be solved

$$I = \frac{1}{\ln 2} \int_{\gamma_0}^{\infty} \gamma^{L\beta/2-1} \ln \left(\frac{\gamma}{\gamma_0} \right) e^{-\left(\frac{\gamma}{\Xi \bar{\gamma}} \right)^{\beta/2}} d\gamma. \tag{1.16}$$

After making a change of variables $t = (\gamma / \gamma_0)^{\beta/2}$ and some simple mathematical manipulations, we get:

$$I = \frac{4\gamma_0^{L\beta/2}}{\beta^2 \ln 2} \int_1^{\infty} t^{L-1} \ln(t) e^{-\left(\frac{\gamma_0}{\Xi \bar{\gamma}} \right)^{\beta/2} t} dt \tag{1.17}$$

Furthermore, this integral can be evaluated using partial integration:

$$\int_1^{\infty} u dv = \lim_{t \rightarrow \infty} (uv) - \lim_{t \rightarrow 1} (uv) - \int_1^{\infty} v du \tag{1.18}$$

with respect to:

$$u = \ln t; \quad dv = t^{L-1} e^{-\left(\frac{\gamma_0}{\Xi \bar{\gamma}} \right)^{\beta/2} t} dt. \tag{1.19}$$

Performing $L-1$ successive integration by parts [16, eq. 2.321.2], we get

$$v = -e^{-mt} \sum_{p=1}^L \frac{(L-1)!}{(L-p)!} \frac{t^{L-p}}{m^p} \tag{1.20}$$

denoting $m = (\gamma_0 / \Xi \bar{\gamma})^{\beta/2}$. Substituting (1.20) in (1.18), we see that first two terms tend to zero. Hence, the integral in (1.17) can be solved in closed form using [16, eq 3.381.3]

$$I = \frac{(L-1)!}{m^L} \sum_{p=0}^{L-1} \frac{\Gamma(p, m)}{p!} \tag{1.21}$$

with $\Gamma(a, x)$ higher incomplete Gamma function [16]. Finally, $\langle C \rangle_{opra}/B$ using L -branch MRC diversity receiver over Weibull fading channels has this form

$$\frac{\langle C \rangle_{opra}^{MRC}}{B} = \frac{2}{\beta \ln 2} \sum_{p=0}^{L-1} \frac{\Gamma(p, m)}{p!} . \tag{1.22}$$

3. Constant power with optimal rate adaptation

With ORA protocol, the transmitter adapts its rate only while maintaining a fixed power level. Thus, this protocol can be implemented at reduced complexity and is more practical than that of optimal simultaneous power and rate adaptation.

The channel capacity, $\langle C \rangle_{ora}$ (bits/s) with constant transmit power policy is given by [1]:

$$\langle C \rangle_{ora} = B \int_0^\infty \log_2(1 + \gamma) p_\gamma(\gamma) d\gamma \tag{1.22}$$

3.1 κ - μ fading channels

To achieve the capacity in (1.22), the channel fading level must be attended at the receiver as well as at the transmitter.

After substituting (1.1) into (1.22), by using partial integration:

$$\int_0^\infty u dv = \lim_{\gamma \rightarrow \infty} (uv) - \lim_{\gamma \rightarrow 0} (uv) - \int_0^\infty v du \tag{1.23}$$

with respect to:

$$u = \ln(1 + \gamma); \quad du = \frac{d\gamma}{1 + \gamma}; \quad dv = \gamma^{p+\mu-1} e^{-\frac{\mu(1+k)\gamma}{\gamma}} ; \tag{1.24}$$

and performing successive integration by parts [16, eq. 2.321.2], we get

$$v = e^{-\frac{\mu(1+k)\gamma}{\gamma}} \sum_{q=1}^{p+\mu} \frac{(p + \mu - 1)! \gamma^{p+\mu-k}}{(p + \mu - q)!} \left(\frac{\gamma}{\mu(1+k)} \right)^q \tag{1.25}$$

By substituting (1.25) in (1.23), we see that first two terms tend to zero. Hence, the integral in (1.23) can be solved in closed form using [16, eq. 3.381.3]. Finally, $\langle C \rangle_{ora}/B$ over κ - μ fading channels has this form:

$$\langle C \rangle_{ora} = \frac{B}{\ln 2} \sum_{p=0}^{\infty} \sum_{q=1}^{p+\mu} \frac{\mu^{2p+\mu-q} \kappa^p (1+\kappa)^{p+\mu-q} (n-1)!}{e^{\mu\kappa} \gamma^{-p+\mu-q} \Gamma(p+\mu)p!} e^{\frac{\mu(1+\kappa)}{\gamma}} \Gamma\left(-n+p+\mu, \frac{\mu(1+\kappa)}{\gamma}\right) \tag{1.26}$$

On the other hand, substituting (1.4) into (1.22) and applying similar procedure, the expression for the $\langle C \rangle_{ora}/B$ with MRC diversity receiver is derived as:

$$\langle C \rangle_{ora}^{MRC} = \frac{B}{\ln 2} \sum_{p=0}^{\infty} \sum_{q=1}^{p+\mu} \frac{\mu^{2p+\mu-q} \kappa^p (1+\kappa)^{p+\mu-q} (n-1)! L^p}{e^{L\mu\kappa} \gamma^{-p+\mu-q} \Gamma(p+\mu)p!} e^{\frac{\mu(1+\kappa)}{\gamma}} \Gamma\left(-n+p+\mu L, \frac{\mu(1+\kappa)}{\gamma}\right) \tag{1.27}$$

Convergence of infinite series expressions in (1.26) and (1.27) is rapid, since we need 5-10 terms to be summed in order to achieve accuracy at the 5th significant digit for corresponding values of system parameters.

3.2 Weibull fading channels

After substituting (1.6) into (1.22), when MRC reception is applied over Weibull fading channel, we can obtain expression for the ORA channel capacity, in the form of:

$$\frac{\langle C \rangle_{ora}}{B} = \frac{\beta}{2\Gamma(L)(\Xi\bar{\gamma})^{L\beta/2} \ln 2} \int_0^{\infty} \gamma^{L\beta/2} \ln(1+\gamma) e^{-\left(\frac{\gamma}{\Xi\bar{\gamma}}\right)^{\beta/2}} d\gamma. \tag{1.28}$$

By expressing the logarithmic and exponential integrands as Meijer's G- functions [21, eqs. 11] and using [22, eq. 07.34.21.0012.01], integral in (1.28) is solved in closed-form:

$$\frac{\langle C \rangle_{ora}}{B} = \frac{\beta}{2\Gamma(L)(\Xi\bar{\gamma})^{L\beta/2} \ln 2} H_{2,3}^{3,1} \left((\Xi\bar{\gamma})^{-\beta/2} \left| \begin{matrix} (-L\beta/2, \beta/2), (1-L\beta/2, \beta/2) \\ (0,1), (-L\beta/2, \beta/2), (-L\beta/2, \beta/2) \end{matrix} \right. \right) \tag{1.29}$$

with:

$$H_{p,q}^{m,n} \left(x \left| \begin{matrix} (a_1, \alpha_1) \dots (a_p, \alpha_p) \\ (b_1, \beta_1) \dots (b_q, \beta_q) \end{matrix} \right. \right) \tag{1.30}$$

denoting the Fox H function [23].

4. Channel inversion with fixed rate

Channel inversion with fixed rate policy (CIFR protocol) is quite different than the first two protocols as it maintains constant rate and adapts its power to the inverse of the channels fading. CIFR protocol achieves what is known as the outage capacity of the system; that is the maximum constant data rate that can be supported for all channel conditions with some probability of outage. However, the capacity of channel inversion is always less than the capacity of the previous two protocols as the transmission rate is fixed. On the other hand, constant rate transmission is required in some applications and is worth the loss in achievable capacity. CIFR is adaptation technique based on inverting the channel fading. It is the least complex technique to implement assuming that the transmitter on this way adapts its power to maintain a constant SNR at the receiver. Since a large amount of the transmitted power is required to compensate for the deep channel fades, channel inversion with fixed rate suffers a certain capacity penalty compared to the other techniques.

The channel capacity with this technique is derived from the capacity of an AWGN channel and is given in [8]:

$$\langle C \rangle_{cifr} = B \log_2 \left(1 + 1 / \int_0^{\infty} (p_{\gamma}(\gamma) / \gamma) d\gamma \right). \quad (1.31)$$

4.1 κ - μ fading channels

After substituting (1.1) into (1.31), and by using [16, eq. 6.643.2]:

$$\int_0^{\infty} x^{\mu - \frac{1}{2}} e^{-\alpha x} I_{2\nu}(2\beta\sqrt{x}) dx = \frac{\Gamma\left(\mu + \nu + \frac{1}{2}\right)}{\Gamma(2\nu + 1)} \beta^{-1} e^{-\frac{\beta^2}{2\alpha}} \alpha^{-\nu} M_{-\mu, \nu}\left(\frac{\beta^2}{\alpha}\right) \quad (1.32)$$

where $M_{k,m}(z)$ is the Whittaker's function, we can obtain expression for the CIFR channel capacity in the form of:

$$\langle C \rangle_{cifr} = B \log_2 \left(1 + \frac{(\mu - 1)}{e^{-\frac{\mu k}{2}} \left(\frac{1+k}{k\bar{\gamma}}\right)^{\frac{\mu}{2}} M_{1-\frac{\mu}{2}, \frac{\mu-1}{2}}(\mu k)} \right). \quad (1.33)$$

Case when MRC diversity is applied can be modelled by:

$$\langle C \rangle_{cifr}^{MRC} = B \log_2 \left(1 + \frac{(L\mu - 1)}{e^{-\frac{\mu k L}{2}} \left(\frac{1+k}{kL\bar{\gamma}}\right)^{\frac{L\mu}{2}} M_{1-\frac{L\mu}{2}, \frac{L\mu-1}{2}}(\mu k L)} \right). \quad (1.34)$$

Similarly, after substituting (1.5) into (1.31), with respect to [16, eqs. 8.531, 7.552.5, 9.14]:

$$\Lambda(a, x) = \frac{x^a}{a} e^{-x} {}_1F_1(1; 1 + a; x) \tag{1.35}$$

$$\int_0^\infty e^{-x} x^{s-1} {}_pF_q(a_1, \dots, a_p; b_1, \dots, b_q; \alpha x) dx = \Gamma(s) {}_{p+1}F_q(s, a_1, \dots, a_p; b_1, \dots, b_q; \alpha x) \tag{1.36}$$

$${}_1F_1(a; b; x) = \sum_{k=0}^\infty \frac{(a)_k x^k}{(b)_k k!} \tag{1.37}$$

expressions for the CIFR channel capacity over κ - μ fading with SC diversity applied for dual and triple branch combining at the receiver can be obtained in the form of:

$$\begin{aligned} \langle C \rangle_{cifr}^{SC-2} &= B \log_2 \left(1 + 1 / \sum_{p=0}^\infty \sum_{q=0}^\infty f_1 \right) \\ f_1 &= \frac{\mu^{p+q+1} \kappa^{p+q} (1 + \kappa) \Gamma(p + q + 2\mu - 1)}{2^{p+q+2\mu-2} e^{2\mu\kappa} \bar{\gamma} \Gamma(p + \mu) p! \Gamma(q + \mu) q! (q + \mu)} \\ &\quad {}_2F_1 \left(p + q + 2\mu - 1, 1; 1 + q + \mu; \frac{1}{2} \right) \end{aligned} \tag{1.38}$$

$$\begin{aligned} \langle C \rangle_{cifr}^{SC-3} &= B \log_2 \left(1 + 1 / \sum_{p=0}^\infty \sum_{q=0}^\infty \sum_{r=0}^\infty \sum_{s=0}^\infty f_2 \right) \\ f_2 &= \frac{\mu^{p+q+r+1} \kappa^{p+q+r} (1 + \kappa) \Gamma(p + q + r + s + 3\mu - 1)}{3^{p+q+r+s+3\mu-3} e^{3\mu\kappa} \bar{\gamma} \Gamma(p + \mu) p! \Gamma(q + \mu) q! (q + \mu) \Gamma(r + \mu) r! (r + \mu) (1 + r + \mu)_s} \\ &\quad \times {}_2F_1 \left(p + q + r + s + 3\mu - 1, 1; 1 + q + \mu; \frac{1}{3} \right) \end{aligned} \tag{1.39}$$

Number of terms that need to be summed in (1.38) and (1.39) to achieve accuracy at 5th significant digit for some values of system parameters is presented in Table 1 in the section Numerical results.

4.2 Weibull fading channels

After substituting (1.6) into (1.31) we can obtain expression for the CIFR channel capacity when MRC diversity is applied in the form of:

$$\frac{\langle C \rangle_{cifr}^{MRC}}{B} = \log_2 \left(1 + \frac{(\Xi \bar{\gamma}) \Gamma(L)}{\Gamma(L - 2 / \beta)} \right) \tag{1.40}$$

5. Truncated channel inversion with fixed rate

The channel inversion and truncated inversion policies use codes designed for AWGN channels, and are therefore the least complex to implement, but in severe fading conditions they exhibit large capacity losses relative to the other techniques.

The truncated channel inversion policy inverts the channel fading only above a fixed cutoff fade depth γ_0 . The capacity with this truncated channel inversion and fixed rate policy $\langle C \rangle_{tifr}/B$ is derived in [8]:

$$\langle C \rangle_{tifr} = B \log_2 \left(1 + 1 / \int_{\gamma_0}^{\infty} (p_{\gamma}(\gamma) / \gamma) d\gamma \right) (1 - P_{out}). \quad (1.41)$$

5.1 κ - μ fading channels

After substituting (1.2) into (1.40) we can obtain expression for the CIFR channel capacity over κ - μ fading channel in the following form:

$$\langle C \rangle_{tifr} = B \log_2 \left(1 + 1 / \sum_{p=0}^{\infty} f_3 \right) \left(1 - \sum_{i=0}^{\infty} \frac{(k\mu)^i}{e^{\mu k} \Gamma(i + \mu) i!} \Lambda \left(\mu + i, \frac{\mu(1+k)\gamma_0}{\gamma} \right) \right) \quad (1.42)$$

$$f_3 = \sum_{p=0}^{\infty} \frac{\mu^{p+1} \kappa^p (1 + \kappa) \Lambda \left(p + \mu - 1, \frac{\mu(1+k)\gamma_0}{\gamma} \right)}{e^{\mu k} \gamma \Gamma(p + \mu) p!}$$

Case when MRC diversity is applied can be modelled by:

$$\langle C \rangle_{tifr}^{MRC} = B \log_2 \left(1 + 1 / \sum_{p=0}^{\infty} f_4 \right) \left(1 - \sum_{i=0}^{\infty} \frac{(kL\mu)^i}{e^{L\mu k} \Gamma(i + L\mu) i!} \Lambda \left(\mu L + i, \frac{\mu L(1+k)\gamma_0}{\gamma} \right) \right) \quad (1.43)$$

$$f_4 = \sum_{p=0}^{\infty} \frac{\mu^{p+1} \kappa^p (1 + \kappa) L^p \Lambda \left(p + \mu L - 1, \frac{\mu L(1+k)\gamma_0}{\gamma} \right)}{e^{\mu k L} \gamma \Gamma(p + \mu L) p!}$$

Convergence of infinite series expressions in (1.42) and (1.43) is rapid, since we need about 10-15 terms to be summed in order to achieve accuracy at the 5th significant digit.

5.2 Weibull fading channels

After substituting (1.6) into (1.41) we can obtain expression for the CIFR channel capacity over Weibull fading channels when MRC diversity is applied in the form of:

$$\frac{\langle C \rangle_{ufr}^{MRC}}{B} = \log_2 \left(1 + \frac{\Xi \bar{\gamma} \Gamma(L)}{\Gamma(L - 2 / \beta, (\gamma_0 / \Xi \bar{\gamma})^{\beta/2})} \right) \frac{\Gamma(L, (\gamma_0 / \Xi \bar{\gamma})^{\beta/2})}{\Gamma(L)}. \quad (1.44)$$

6. Numerical results

In order to discuss usage of diversity techniques and adaptation policies and to show the effects of various system parameters on obtained channel capacity, numerically obtained results are graphically presented.

In Figs. 1.1 and 1.8 channel capacity without diversity, $\langle C \rangle_{ora}$ given by (1.22), for the cases when κ - μ and Weibull fading are affecting channels, for various system parameters are plotted against γ . These figures also display the capacity per unit bandwidth of an AWGN channel, C_{AWGN} given by:

$$C_{AWGN} = B \log_2(1 + \gamma). \quad (1.45)$$

Considering obtained results, with respect that $C_{AWGN} = 3.46$ dB for average received SNR of 10dB we find that depending of fading parameters of κ - μ and Weibull distribution, channel capacity could be reduced up to 30 %. From Fig. 1.1 we can see that channel capacity is less reduced for the cases when fading severity parameter μ , and dominant/scattered components power ratio κ , have higher values, since for smaller κ and μ values the dynamics in the channel is larger. Also from Fig. 1.8 we can observe that channel capacity is less reduced in the areas where Weibull fading parameter β has higher values.

Figures 1.2-1.4,1.6 show the channel capacity per unit bandwidth as a function of $\bar{\gamma}$ for the different adaptation policies with MRC diversity over κ - μ fading channels. It can be seen that as the number of combining branches increases the fading influence is progressively reduced, so the channel capacity improves remarkably. However, as L increases, all capacities of the various policies converge to the capacity of an array of L independent AWGN channels, given by:

$$C_{AWGN}^{MRC} = B \log_2(1 + L\gamma) \quad (1.46)$$

Thus, in practice it is not possible to entirely eliminate the effects of fading through space diversity since the number of diversity branches is limited. Also considering downlink (base station to mobile) implementation, we found that mobile receivers are generally constrained in size and power.

In Fig. 1.5 comparison of the channel capacity per unit bandwidth with CIFR adaptation policy, when SC and MRC diversity techniques are applied at the reception is shown. As expected, better performances are obtained when MRC reception over κ - μ fading channels is applied.

Figure 1.7 shows the calculated channel capacity per unit bandwidth as a function of $\bar{\gamma}$ for different adaptation policies. From this figure we can see that the OPRA protocol yields a small increase in capacity over constant transmit power adaptation and this small increase

in capacity diminishes as $\bar{\gamma}$ increases. However, greater improvement is obtained in going from complete to truncated channel inversion policy. Truncated channel inversion policy provides better diversity gain compared to complete channel inversion varying any of parameters.

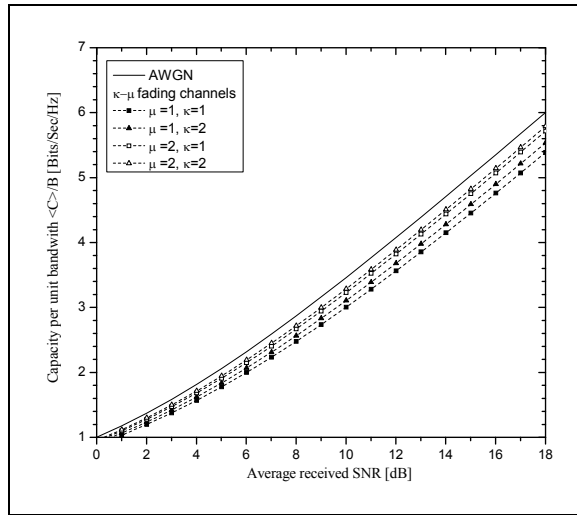


Fig. 1.1 Average channel capacity per unit bandwidth for a κ - μ fading and an AWGN channel versus average received SNR.

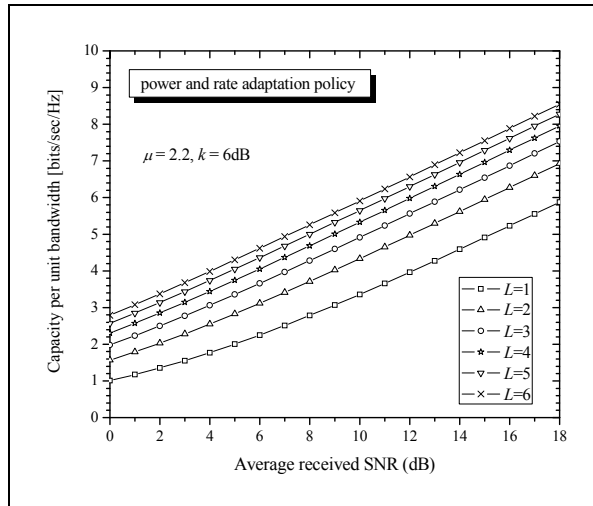


Fig. 1.2 Power and rate adaptation policy capacity per unit bandwidth over κ - μ fading channels, for various values of diversity order.

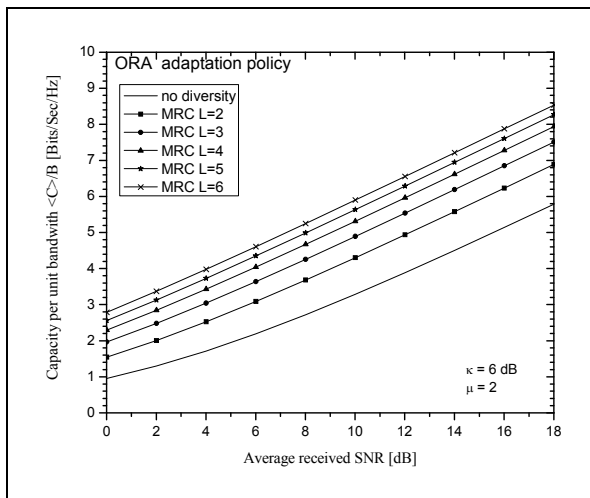


Fig. 1.3 ORA policy capacity per unit bandwidth over κ - μ fading channels, for various values of MRC diversity order.

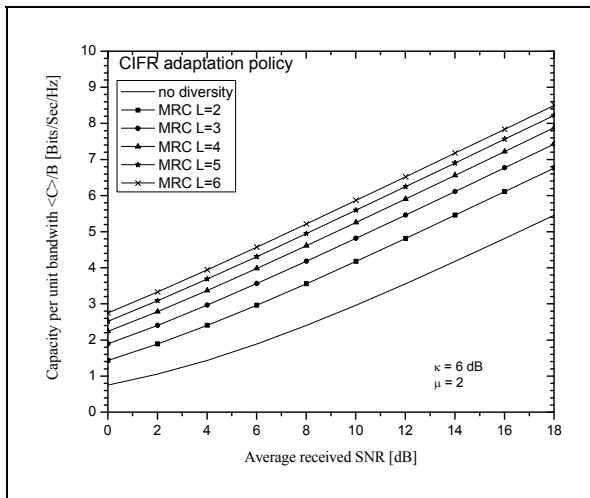


Fig. 1.4 CIFR policy capacity per unit bandwidth over κ - μ fading channels, for various values of MRC diversity order.

Similar results are presented considering channels affected by Weibull fading. Figures 1.9-1.12 show the channel capacity per unit bandwidth as a function of $\bar{\gamma}$ for the different adaptation policies with L -branch MRC diversity applied. Comparison of adaptation policies is presented at Fig. 1.13.

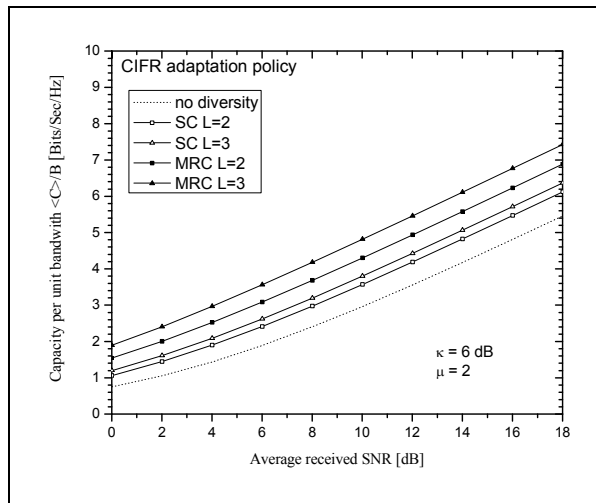


Fig. 1.5 CIFR policy capacity per unit bandwidth over κ - μ fading channels, for MRC and SC diversity techniques various orders.

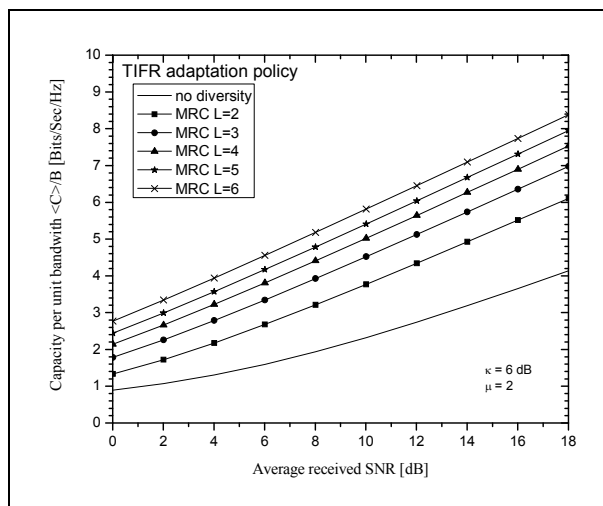


Fig. 1.6 TIFR policy capacity per unit bandwidth over κ - μ fading channels, for various values of MRC diversity order.

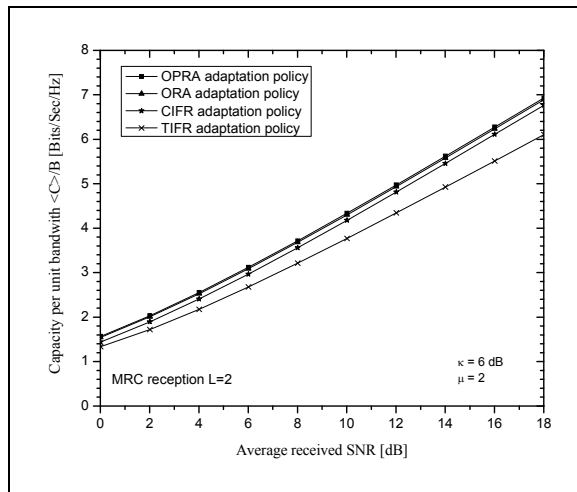


Fig. 1.7 Comparison of adaptation policies over MRC diversity reception in the presence of κ - μ fading.

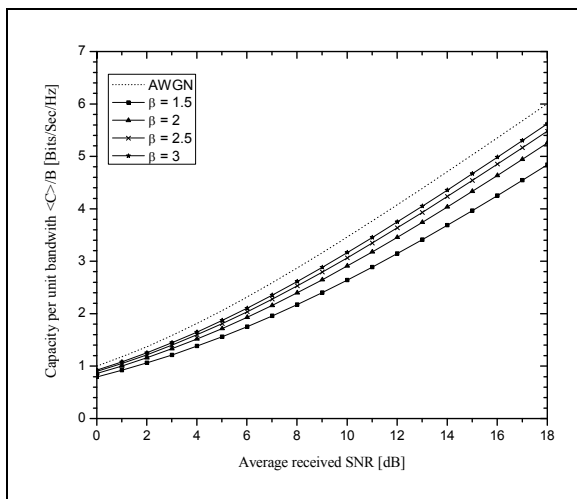


Fig. 1.8 Average channel capacity per unit bandwidth for a Weibull fading for various values of system parameters and an AWGN channel versus average received SNR [dB].

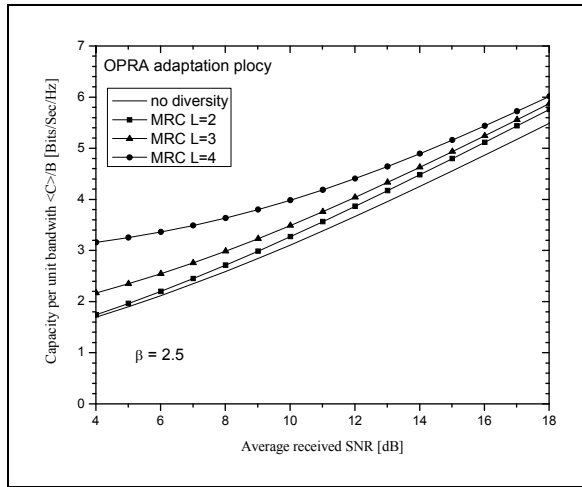


Fig. 1.9 ORPA policy capacity per unit bandwidth over Weibull fading channels, for various values of MRC diversity order.

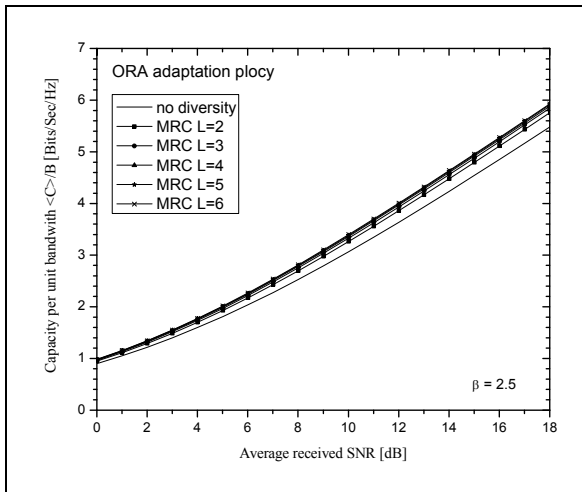


Fig. 1.10 ORA policy capacity per unit bandwidth over Weibull fading channels, for various values of MRC diversity order.

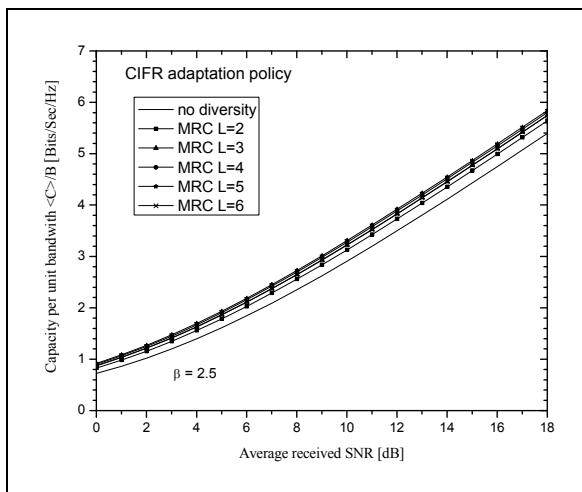


Fig. 1.11 CIFR policy capacity per unit bandwidth over Weibull fading channels, for various values of MRC diversity order.

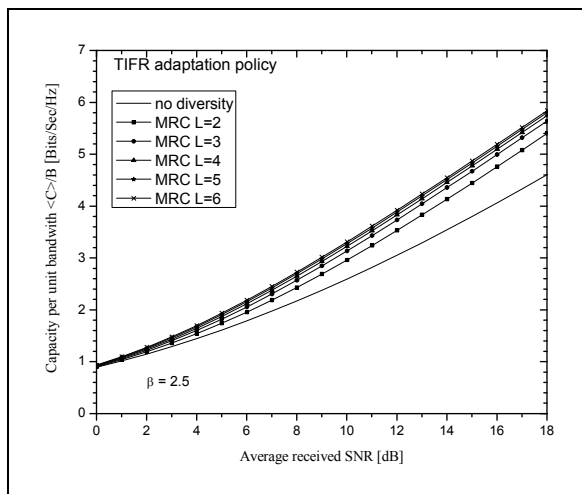


Fig. 1.12 TIFR policy capacity per unit bandwidth over Weibull fading channels, for various values of MRC diversity order.

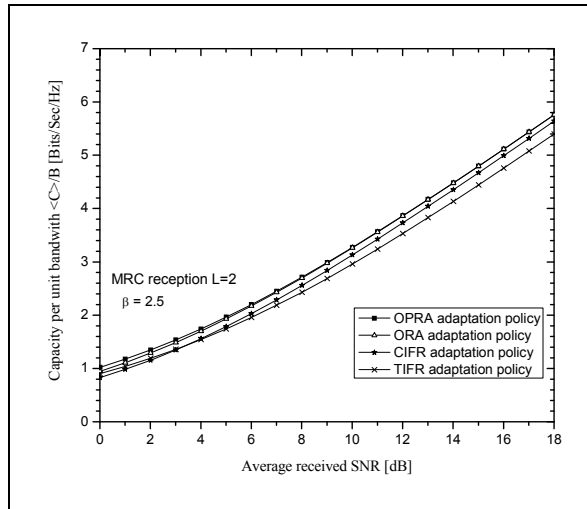


Fig. 1.13 Comparison of adaptation policies over MRC diversity reception in the presence of Weibull fading.

The nested infinite sums in (1.38) and (1.39), as can be seen from Table 1, for dual and triple branch diversity case, converge for any value of the parameters κ , μ and $\bar{\gamma}$. As it is shown in this Table 1, the number of the terms need to be summed to achieve a desired accuracy, depends strongly on these parameters and it increases as these parameter values increase.

Expression (1.38) 6 th significant digit		$\bar{\gamma} = 5$ dB	$\bar{\gamma} = 10$ dB	$\bar{\gamma} = 15$ dB
$\kappa = 1$	$\mu = 1$	8	9	10
$\kappa = 2$	$\mu = 2$	15	15	16
Expression (1.39) 6 th significant digit		$\bar{\gamma} = 5$ dB	$\bar{\gamma} = 10$ dB	$\bar{\gamma} = 15$ dB
$\kappa = 1$	$\mu = 1$	19	21	24
$\kappa = 2$	$\mu = 2$	23	26	28

Table 1. Number of terms that need to be summed in (1.38) and (1.39) to achieve accuracy at the specified significant digit for some values of system parameters.

7. Conclusion

Cases when wireless channels are affected by general and nonlinear fading distributions are discussed in this chapter. The analytical study of the κ - μ fading channel capacity, e.g., under the OPRA, ORA, CIFR and TIFR adaptation policies and MRC and SC diversity techniques is performed. The main contribution are closed-form expressions derived for the proposed adaptation policies and diversity techniques. Based on them, numerically obtained results are graphically presented in order to show the effects of various system parameters. Since κ - μ model as general physical fading model includes Rayleigh, Rician, and Nakagami- m fading models, as special cases, the generality and applicability of this analysis are more than obvious. Nonlinear fading scenario is discussed in the similar manner, as an analytical

study of the Weibull fading channel capacity, under the OPRA, ORA, CIFR and TIFR adaptation policies and MRC diversity technique.

8. Acknowledgment

This paper was supported by the Serbian Ministry of Education and Science (projects: III44006 and TR32023).

9. References

- [1] Goldsmith, A. & Varaiya, P. (1997). Capacity of fading channels with channel side information. *IEEE Transactions on Information Theory*, vol. 43, no. 6, (November 1997), pp. 1896–1992.
- [2] Freeman, L. R. (2005). *Fundamentals of telecommunications*, John Wiley & sons, Hoboken, New Jersey, 2005.
- [3] Sampei, S. ; Morinaga, N. & Kamio, Y. (1995). Adaptive modulation/TDMA with a BDDFE for 2 Mbit/s multimedia wireless communication systems, *Proceedings of the IEEE VTC'95*, pp. 311-315, 1995.
- [4] Lee, W.C.Y. (2001). *Mobile communications engineering*, Mc-Graw-Hill, New York 2001.
- [5] Ibnkahla, M. (2000). *Signal processing for mobile communications*, CRC Press LLC, Boca Raton, Florida, 2000.
- [6] Brennan, D. (1959). Linear diversity combining techniques, *Proceedings of IRE*, vol.47, (June 1959), pp. 1075-1102.
- [7] Alouini, M. S. & Goldsmith, A. (1999). Capacity of Rayleigh fading channels under different adaptive transmission and diversity-combining techniques, *IEEE Transactions on Vehicular Technology*, vol. 48, no. 4, (July 1999), pp. 1165-1181.
- [8] Simon, M. & Alouini, M. S. (2000). *Digital communications over fading channels*, John Wiley & sons, New York, 2000.
- [9] Shao, J.; Alouini, M. & Goldsmith A. (1999). Impact of fading correlation and unequal branch gains on the capacity of diversity systems. In: *Proceedings of the IEEE vehicular technology conference (VTC-Spring'99)*, Houston, TX, May 1999. pp. 2159–2163.
- [10] Bhaskar, V. (2008). Capacity evaluation for equal gain diversity schemes over Rayleigh fading channels, *AEÜ - International Journal of Electronics and Communications*, vol. 63, no. 9, pp. 235-240.
- [11] Anastasov, J., Panic, S., Stefanovic M. & Milenkovic V. Capacity of correlative Nakagami-m fading channels under adaptive transmission and maximal-ratio combining diversity technique, accepted for publication in *Journal of Communications Technology and Electronics*
- [12] Subadar, R. & Sahu, P. (2011). Channel capacity of adaptive transmission schemes using equal gain combining receiver over Hoyt fading channels, *Communications (NCC), 2011 National Conference on*, 28-30 Jan. 2011, Bangalore, pp 1-5
- [13] Sagias, N. (2006) Capacity of dual-branch selection diversity receivers in correlative Weibull fading," *European Transactions on Telecommunications*, vol. 16, no. 1, (February 2006), pp. 37-43.
- [14] Yacoub, M. (2007). The κ - μ distribution and the η - μ distribution, *IEEE Antennas and Propagation Magazine*, vol. 49 no. 1, January 2007, pp. 68-81.

- [15] Milisic, M.; Hamza, M. & Hadzialic M. (2009). BEP/SEP and Outage Performance Analysis of L -Branch Maximal-Ratio Combiner for κ - μ Fading, *International Journal of Digital Multimedia Broadcasting*, vol. 2009, Article ID 573404, pp 1-8.
- [16] Gradshteyn, I. & Ryzhik, I. (1994). *Table of Integrals, Series, and Products*, Academic Press, 5th ed., Orlando, 1994.
- [17] Sagias, N.; Zogas, D. ; Karagiannidis, G. & Tombras, G. (2004). Channel Capacity and Second Order Statistics in Weibull Fading, *IEEE Communications Letters*, vol. 8, no. 6, (June 2004) pp. 377-379.
- [18] Filho, J. & Yacoub, M. (2006). Simple Precise Approximations to Weibull Sums. *IEEE Communication Letters*, vol. 10, no. 8, (August 2006), pp. 614-616.
- [19] Sagias, N. & Mathiopoulos, P. (2005). Switched diversity receivers over generalized Gamma fading channels, *IEEE Communications Letters*, vol. 9, no. 10, (October 2005), pp. 871-873.
- [20] Abramowitz, M. & Stegun, I. (1970). *Handbook of Mathematical Functions*, Dover Publications, Inc., New York, 1970.
- [21] Akademik V. & Marichev, O. (1990). The algorithm for calculating integrals of hypergeometric type functions and its realization in REDUCE system in *Proc. Int. Conf. on Symbolic and Algebraic Computation*, Tokyo, Japan, 1990., pp. 212-224.
- [22] The Wolfram Functions Site, 2008. [Online] Available: <http://functions.wolfram.com>
- [23] Prudnikov, A.; Brychkov, Y. & Marichev, O. (1990). *Integral and Series: Volume 3, More Special Functions*. CRC Press Inc., 1990.

Part 4

Wireless Communication Performance Analysis Tools and Methods

Generalized Approach to Signal Processing in Wireless Communications: The Main Aspects and some Examples

Vyacheslav Tuzlukov
Kyungpook National University
South Korea

1. Introduction

The additive and multiplicative noise exists forever in any wireless communication system. Quality and integrity of any wireless communication systems are defined and limited by statistical characteristics of the noise and interference, which are caused by an electromagnetic field of the environment. The main characteristics of any wireless communication system are deteriorated as a result of the effect of the additive and multiplicative noise. The effect of addition of noise and interference to the signal generates an appearance of false information in the case of the additive noise. For this reason, the parameters of the received signal, which is an additive mixture of the signal, noise, and interference, differ from the parameters of the transmitted signal. Stochastic distortions of parameters in the transmitted signal, attributable to unforeseen changes in instantaneous values of the signal phase and amplitude as a function of time, can be considered as multiplicative noise. Under stimulus of the multiplicative noise, false information is a consequence of changed parameters of transmitted signals, for example, the parameters of transmitted signals are corrupted by the noise and interference. Thus, the impact of the additive noise and interference may be lowered by an increase in the signal-to-noise ratio (SNR). However, in the case of the multiplicative noise and interference, an increase in SNR does not produce any positive effects.

The main functional characteristics of any wireless communication systems are defined by an application area and are often specific for distinctive types of these systems. In the majority of cases, the main performance of any wireless communication systems are defined by some initial characteristics describing a quality of signal processing in the presence of noise: the precision of signal parameter measurement, the definition of resolution intervals of the signal parameters, and the probability of error.

The main idea is to use the generalized approach to signal processing (GASP) in noise in wireless communication systems (Tuzlukov, 1998; Tuzlukov, 1998; Tuzlukov, 2001; Tuzlukov, 2002; Tuzlukov, 2005; Tuzlukov, 2012). The GASP is based on a seemingly abstract idea: the introduction of an additional noise source that does not carry any information about the signal and signal parameters in order to improve the qualitative performance of wireless communication systems. In other words, we compare statistical data defining the statistical parameters of the probability distribution densities (pdfs) of the observed input stochastic samp-

les from two independent frequency-time regions – a "yes" signal is possible in the first region and it is known a priori that a "no" signal is obtained in the second region. The proposed GASP allows us to formulate a decision-making rule based on the determination of *the jointly sufficient statistics of the mean and variance* of the likelihood function (or functional). Classical and modern signal processing theories allow us to define *only the mean* of the likelihood function (or functional). Additional information about the statistical characteristics of the likelihood function (or functional) leads us to better quality signal detection and definition of signal parameters in compared with the optimal signal processing algorithms of classical or modern theories.

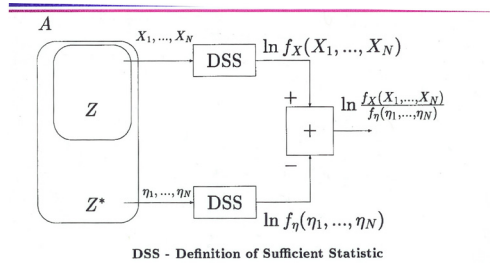
Thus, for any wireless communication systems, we have to consider two problems – analysis and synthesis. The first problem (analysis) – the problem to study a stimulus of the additive and multiplicative noise on the main principles and performance under the use of GASP – is an analysis of impact of the additive and multiplicative noise on the main characteristics of wireless communication systems, the receivers in which are constructed on the basis of GASP. This problem is very important in practice. This analysis allows us to define limitations on the use of wireless communication systems and to quantify the additive and multiplicative noise impact relative to other sources of interference present in these systems. If we are able to conclude that the presence of the additive and multiplicative noise is the main factor or one of the main factors limiting the performance of any wireless communication systems, then the second problem – the definition of structure and main parameters and characteristics of the generalized detector or receiver (GD or GR) under a dual stimulus of the additive and multiplicative noise – the problem of synthesis – arises.

GASP allows us to extend the well-known boundaries of the potential noise immunity set by classical and modern signal processing theories. Employment of wireless communication systems, the receivers of which are constructed on the basis of GASP, allows us to obtain high detection of signals and high accuracy of signal parameter definition with noise components present compared with that systems, the receivers of which are constructed on the basis of classical and modern signal processing theories. The optimal and asymptotic optimal signal processing algorithms of classical and modern theories, for signals with amplitude-frequency-phase structure characteristics that can be known and unknown a priori, are constituents of the signal processing algorithms that are designed on the basis of GASP.

2. GASP: Brief description

GASP is based on the assumption that the frequency-time region Z of the noise exists where a signal may be present; for example, there is an observed stochastic sample from this region, relative to which it is necessary to make the decision a "yes" signal (the hypothesis H_1) or a "no" signal (the hypothesis H_0). We now proceed to modify the initial premises of the classical and modern signal processing theories. Let us suppose there are two independent frequency-time regions Z and Z^* belonging to the space A . Noise from these regions obeys the same pdf with the same statistical parameters (for simplicity of considerations). Generally, these parameters are differed. A "yes" signal is possible in the noise region Z as before. *It is known a priori that a "no" signal is obtained in the noise region Z^* .* It is necessary to make the decision a "yes" signal (the hypothesis H_1) or a "no" signal (the hypothesis H_0) in the observed stochastic sample from the region Z , by comparing statistical parameters of pdf of this

sample with those of the sample from the reference region Z^* . Thus, there is a need to accumulate and compare statistical data defining the statistical parameters of pdf of the observed input stochastic samples from two independent frequency-time regions Z and Z^* . If statistical parameters for two samples are equal or agree with each other within the limits of a given before accuracy, then the decision of a "no" signal in the observed input stochastic process X_1, \dots, X_N is made - the hypothesis H_0 . If the statistical parameters of pdf of the observed input stochastic sample from the region Z differ from those of the reference sample from the region Z^* by a value that exceeds the prescribed error limit, then the decision of a "yes" signal in the region Z is made - the hypothesis H_1 .



It is known *a priori* that a "no" signal obtains in the noise region Z^* .

Fig. 1. Definition of sufficient statistics under GASP.

The simple model of GD in form of block diagram is represented in Fig.2. In this model, we use the following notations: MSG is the model signal generator (the local oscillator), the AF is the additional filter (the linear system) and the PF is the preliminary filter (the linear system) A detailed discussion of the AF and PF can be found in (Tuzlukov, 2001 and Tuzlukov, 2002).

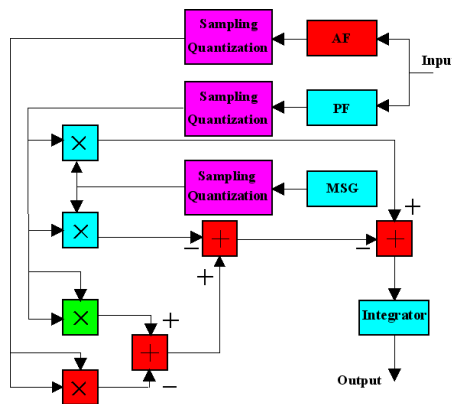


Fig. 2. Principal flowchart of GD.

Consider briefly the main statements regarding the AF and PF. There are two linear systems at the GD front end that can be presented, for example, as bandpass filters, namely, the PF with the impulse response $h_{PF}(\tau)$ and the AF with the impulse response $h_{AF}(\tau)$. For simplicity of analysis, we think that these filters have the same amplitude-frequency responses and bandwidths. Moreover, a resonant frequency of the AF is detuned relative to a resonant frequency of PF on such a value that signal cannot pass through the AF (on a value that is higher the signal bandwidth). Thus, the signal and noise can be appeared at the PF output and *the only noise* is appeared at the AF output. It is well known, if a value of detuning between the AF and PF resonant frequencies is more than $4 \div 5\Delta f_a$, where Δf_a is the signal bandwidth, the processes forming at the AF and PF outputs can be considered as independent and uncorrelated processes (in practice, the coefficient of correlation is not more than 0.05). In the case of signal absence in the input process, the statistical parameters at the AF and PF outputs will be the same, because the same noise is coming in at the AF and PF inputs, and we may think that the AF and PF do not change the statistical parameters of input process, since they are the linear GD front end systems.

By this reason, the AF can be considered as a generator of reference sample with *a priori* information *a "no" signal is obtained in the additional reference noise* forming at the AF output. There is a need to make some comments regarding the noise forming at the PF and AF outputs. If the Gaussian noise $n(t)$ comes in at the AF and PF inputs (the GD linear system front end), the noise forming at the AF and PF outputs is Gaussian, too, because the AF and PF are the linear systems and, in a general case, take the following form:

$$n_{PF}(t) = \int_{-\infty}^{\infty} h_{PF}(\tau)n(t-\tau)d\tau \quad \text{and} \quad n_{AF}(t) = \int_{-\infty}^{\infty} h_{AF}(\tau)n(t-\tau)d\tau. \quad (1)$$

If, for sake of simplicity, the additive white Gaussian noise (AWGN) with zero mean and two-sided power spectral density $2N_0$ is coming in at the AF and PF inputs (the GD linear system front end), then the noise forming at the AF and PF outputs is Gaussian with zero mean and variance given by $\sigma_n^2 = \frac{2N_0\omega_0^2}{8\Delta_F}$ (Tuzlukov, 2002) where in the case if AF (or PF) is the RLC oscillatory circuit, the AF (or PF) bandwidth Δ_F and resonance frequency ω_0 are defined in the following manner $\Delta_F = \pi\beta$, $\omega_0 = \frac{1}{\sqrt{LC}}$, $\beta = \frac{R}{2L}$. The main functioning condition of GD is an equality over the whole range of parameters between the model signal $u^*(t)$ at the GD MSG output and the transmitted signal $u(t)$ forming at the GD input liner system (the PF) output, i.e. $u(t) = u^*(t)$. How we can satisfy this condition in practice is discussed in detail in (Tuzlukov, 2002; Tuzlukov, 2012). More detailed discussion about a choice of PF and AF and their impulse responses is given in (Tuzlukov, 1998).

3. Diversity problems in wireless communication systems with fading

In the design of wireless communication systems, two main disturbance factors are to be properly accounted for, i.e. fading and additive noise. As to the former, it is usually taken into account by modeling the propagation channel as a linear-time-varying filter with random impulse response (Bello, 1963 & Proakis, 2007). Indeed, such a model is general enough to encompass the most relevant instances of fading usually encountered in practice, i.e.

frequency- and/or time-selective fading, and flat-flat fading. As to the additive noise, such a disturbance has been classically modeled as a possibly correlated Gaussian random process.

However, the number of studies in the past few decades has shown, through both theoretical considerations and experimental results, that Gaussian random processes, even though they represent a faithful model for the thermal noise, are largely inadequate to model the effect of real-life noise processes, such as atmospheric and man-made noise (Kassam, 1988 & Webster, 1993) arising, for example, in outdoor mobile communication systems. It has also been shown that non-Gaussian disturbances are commonly encountered in indoor environments, for example, offices, hospitals, and factories (Blankenship & Rappaport, 1993), as well as in underwater communications applications (Middleton, 1999). These disturbances have an impulsive nature, i.e. they are characterized by a significant probability of observing large interference levels.

Since conventional receivers exhibit dramatic performance degradations in the presence of non-Gaussian impulsive noise, a great attention has been directed toward the development of non-Gaussian noise models and the design of optimized detection structures that are able to operate in such hostile environments. Among the most popular non-Gaussian noise models considered thus far, we cite the alpha-stable model (Tsihrintzis & Nikias, 1995), the Middleton Class-A and Class-B noise (Middleton, 1999), the Gaussian-mixture model (Garth & Poor, 1992) which, in turn, is a truncated version, at the first order, of the Middleton Class-A noise, and the compound Gaussian model (Conte et al., 1995). In particular, in the recent past, the latter model, subsuming, as special cases, many marginal probability density functions (pdfs) that have been found appropriate for modeling the impulsive noise, like, for instance, the Middleton Class-A noise, the Gaussian-mixture noise (Conte, 1995), and the symmetric alpha-stable noise (Kuruoglu, E. et al., 1998). They can be deemed as the product of a Gaussian, possibly complex random process times a real non-negative one.

Physically, the former component, which is usually referred to as speckle, accounts for the conditional validity of the central limit theorem, whereas the latter, the so-called texture process, rules the gross characteristics of the noise source. A very interesting property of compound-Gaussian processes is that, when observed on time intervals whose duration is significantly shorter than the average decorrelation time of the texture component, they reduce to spherically invariant random processes (SIRPs) (Yao, 1973), which have been widely adopted to model the impulsive noise in wireless communications (Gini, F et al., 1998), multiple access interference in direct-sequence spread spectrum cellular networks (Sousa, 1990), and clutter echoes in radar applications (Sangston & Gerlach, 1994).

We consider the problem of detecting one of M signals transmitted upon a zero-mean fading dispersive channel and embedded in SIRP noise by GD based on the GASP in noise. The similar problem has been previously addressed. In (Conte, 1995), the optimum receiver for flat-flat Rayleigh fading channels has been derived, whereas in (Buzzi et al., 1999), the case of Rayleigh-distributed, dispersive fading has been considered. It has been shown therein that the receiver structure consists of an estimator of the short-term conditional, i.e. given the texture component, noise power and of a bank of M estimators-correlators keyed to the estimated value of the noise power. Since such a structure is not realizable, a suboptimum detection structure has been introduced and analyzed in (Buzzi et al., 1997).

We design the GD extending conditions of (Buzzi et al., 1997) and (Buzzi et al., 1999) to the case that a diversity technique is employed. It is well known that the adoption of diversity techniques is effective in mitigating the negative effects of the fading, and since conventional diversity techniques can incur heavy performance loss in the presence of impulsive disturbance (Kassam & Poor, 1985), it is of interest to envisage the GD for optimized diversity reception in non-Gaussian noise. We show that the optimum GD is independent of the joint pdf of the texture components on each diversity branch. We also derive a suboptimum GD, which is amenable to a practice. We focus on the relevant case of binary frequency-shift-keying (BFSK) signaling and provide the error probability of both the optimum GD and the suboptimum GD. We assess the channel diversity order impact and noise spikiness on the performance.

3.1 Problem statement

The problem is to derive the GD aimed at detecting one out of M signals propagating through single-input multiple-output channel affected by dispersive fading and introducing the additive non-Gaussian noise. In other words, we have to deal with the following M -ary hypothesis test:

$$H_i \Rightarrow \begin{cases} x_1(t) = s_{1,i}(t) + n_1(t) \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ x_p(t) = s_{p,i}(t) + n_p(t) \end{cases} \quad i = 1, \dots, M \quad t \in [0, T], \quad (2)$$

where P is the channel diversity order and $[0, T]$ is the observation interval; the waveforms $\{x_p(t)\}_{p=1}^P$ are the complex envelopes of the P distinct channel outputs; $\{s_{p,i}(t)\}_{p=1}^P, i = 1, \dots, M$ represent the baseband equivalents of the useful signal received on the P diversity branches under the i th hypothesis. Since the channel is affected by dispersive fading, we may assume (Proakis, 2007) that these waveforms are related to the corresponding transmitted signals $u_i(t)$

$$s_{p,i}(t) = \int_{-\infty}^{\infty} h_p(t, \tau) u_i(t - \tau) d\tau, \quad t \in [0, T] \quad (3)$$

where $h_p(t, \tau), p = 1, \dots, P$ is the random impulse response of the channel p th diversity branch and is modeled as a Gaussian random process with respect to the variable t . In keeping with the uncorrelated-scattering model, we assume that the random processes $h_p(t, \tau), p = 1, \dots, P$ are all statistically independent; as a consequence, the waveforms $\{s_{p,i}(t)\}_{p=1}^P$ are themselves independent complex Gaussian random processes that we assume to be zero-mean and with the covariance function

$$Cov(t, \tau) = E[s_{p,i}(t) s_{p,i}^*(\tau)], \quad i = 1, \dots, M \quad t, \tau \in [0, T] \quad (4)$$

independent of p (the channel correlation properties are identical of each branch) and upper bounded by a finite positive constant. This last assumption poses constraint on the average receive energy in the i -th hypothesis $E_i = \int_0^T Cov_i(t, t) dt < \infty$. We also assume in keeping with

the model (Van Trees, 2001) that $E[s_{p,i}(t)s_{p,i}(\tau)] = 0$. This is not a true limitation in most practical instances, and it is necessarily satisfied if the channel is wide sense stationary. Finally, as to the additive non-Gaussian disturbances $\{n_p(t)\}_{p=1}^P$, we resort to the widely adopted compound model, i.e. we deem the waveform $n_p(t)$ as the product of two independent processes:

$$n_p(t) = v_p(t)g_p(t) \quad , \quad p = 1, \dots, P \quad (5)$$

where $v_p(t)$ is a real non-negative random process with marginal pdf $f_{v_p}(\cdot)$ and $g_p(t)$ is a zero-mean complex Gaussian process. If the average decorrelation time of $v_p(t)$ is much larger than the observation interval $[0, T]$, then the disturbance process degenerates into SIPR (Yao, 1973)

$$n_p(t) = v_p g_p(t) \quad , \quad p = 1, \dots, P \quad (6)$$

From now on, we assume that such a condition is fulfilled, and we refer to (Conte, 1995) for further details on the noise model, as well as for a list of all of the marginal pdfs that are compatible with (5). Additionally, we assume $E[v_p^2] = 1$ and that the correlation function of the random process $g_p(t)$ is either known or has been perfectly estimated based on (5). While previous papers had assumed that the noise realization $n_1(t), \dots, n_p(t)$ were statistically independent, in this paper, this hypothesis is relaxed. To be more definite, we assume that the Gaussian components $g_1(t), \dots, g_p(t)$ are uncorrelated (independent), whereas the random variables v_1, \dots, v_p are arbitrary correlated. We thus denote by $f_{v_1, \dots, v_p}(v_1, \dots, v_p)$ their joint pdf. It is worth pointing out that the above model subsumes the special case that the random variables v_1, \dots, v_p are either statistically independent or fully correlated, i.e. $v_1 = \dots = v_p$. Additionally, it permits modeling a much wider class of situations that may occur in practice. For instance, if one assumes that the P diversity observations are due to a temporal diversity, it is apparent that if the temporal distance between consecutive observations is comparable with the average decorrelation time of the process $v(t)$, then the random variables v_1, \dots, v_p can be assumed to be neither independent nor fully correlated. Such a model also turns out to be useful in clutter modeling in that if the diversity observations are due to the returns from neighboring cells, the corresponding texture components may be correlated (Barnard & Weiner, 1996). For sake of simplicity, consider the white noise case, i.e. $n_p(t)$ possesses an impulsive covariance $\forall p$

$$Cov_n(t, \tau) = 2N_0 E[v_p^2] \delta(t - \tau) = 2N_0 \delta(t - \tau) \quad , \quad (7)$$

where $2N_0$ is the power spectral density (PSD) of the Gaussian component of the noise processes $g_1(t), \dots, g_p(t)$. Notice that this last assumption does not imply any loss of generality should the noise possess a non-impulsive correlation. Then, due to the closure of SIRP with respect to linear transformations, the classification problem could be reduced to the above form by simply preprocessing the observables through a linear whitening filter. In such a situation, the $s_{p,i}(t)$ represent the useful signals at the output of the cascade of the channel and of the whitening filter. Due to the linearity of such systems, they are still Gaussian processes with known covariance functions. Finally, we highlight here that the assumption that

the useful signals and noise covariance functions (3) and (6) are independent of the index p has been made to simplify notation.

3.2 Synthesis and design

3.2.1 Optimum GD structure design

Given the M -ary hypothesis test (1), the synthesis of the optimum GD structure in the sense of attaining the minimum probability of error P_E requires evaluating the likelihood functionals under any hypothesis and adopting a maximum likelihood decision-making rule. Formally, we have

$$\hat{H} = H_i \Rightarrow \Lambda[\mathbf{x}(t); H_i] > \max_{k \neq i} \Lambda[\mathbf{x}(t); H_k] \quad (8)$$

with $\mathbf{x}(t) = [x_1(t), \dots, x_p(t)]^T$. The above functionals are usually evaluated through a limiting procedure. We evaluate the likelihood $f_{\mathbf{x}_Q|H_i}(\mathbf{x}_Q)$ of the Q -dimensional random vector $\mathbf{x} = [x_1, \dots, x_Q]^T$ whose entries are the projections of the received signal along the first Q elements of suitable basis \mathcal{B}_i . Therefore, the likelihood functional corresponding to H_i is

$$\Lambda[\mathbf{x}(t); H_i] = \lim_{Q \rightarrow \infty} \frac{f_{\mathbf{x}_Q|H_i}(\mathbf{x}_Q)}{f_{\mathbf{n}_{AFQ}}(\mathbf{n}_{AFQ})}, \quad (9)$$

where $f_{\mathbf{n}_{AFQ}}(\mathbf{n}_{AFQ})$ is the likelihood corresponding to the reference sample with *a priori* information a “no” signal is obtained in the additional reference noise forming at the AF output, i.e. no useful signal is observed at the P channel outputs. In order to evaluate the limit (9), we resort to a different basis for each hypothesis. We choose for the i -th hypothesis the Karhunen-Loeve basis \mathcal{B}_i determined by the covariance function of the useful received signal under the hypothesis H_i . Projecting the waveform received on the p -th diversity branch along the first N axes of the i -th basis yields the following N -dimensional vector:

$$\mathbf{x}_{N,p}^i = \mathbf{s}_{N,p}^i + v_p \mathbf{g}_{N,p}^i, \quad p = 1, \dots, P \quad (10)$$

where $\mathbf{s}_{N,p}^i$ and $\mathbf{g}_{N,p}^i$ are the corresponding projections of the waveforms $s_{p,i}(t)$ and $g_p(t)$. Since \mathcal{B}_i is the Karhunen-Loeve basis for the random processes $s_{1,i}(t), \dots, s_{p,i}(t)$, the entries of $\mathbf{s}_{N,p}^i$ are a sequence of uncorrelated complex Gaussian random variables with the variances $(\sigma_{s_{1,i}}^2, \dots, \sigma_{s_{N,i}}^2)$ which are the first N eigenvalues of the covariance function $Cov_i(t, u)$, whereas the entries of $\mathbf{g}_{N,p}^i$ are a sequence of uncorrelated Gaussian variables with variance $2N_0$. Here we adopt the common approach of assuming that any complete orthonormal system is an orthonormal basis for white processes (Conte, 1995 and Poor, 1988). Upon defining the following NP -dimensional vector

$$\mathbf{x}_N^i = [\mathbf{x}_{N,1}^{iT}, \mathbf{x}_{N,2}^{iT}, \dots, \mathbf{x}_{N,P}^{iT}]^T \quad (11)$$

the likelihood functional taking into consideration subsection 3.1 and (Tuzlukov, 2001) can be written in the following form

$$\Lambda[\mathbf{x}_N^i; H_i] = \frac{f_{\mathbf{x}_N^i | H_i}(\mathbf{x}_N^i)}{f_{\mathbf{n}_{AFN}^i | H_0}(\mathbf{n}_{AFN}^i)} = \frac{\int \prod_{p=1}^P \prod_{j=1}^N \frac{1}{\sigma_{s_{j,i}}^2 + 4\sigma_n^4 y_p^2} \exp\left[-\frac{|x_{j,p}^i|^2}{\sigma_{s_{j,i}}^2 + 4\sigma_n^4 y_p^2}\right] f_v(\mathbf{y}) d\mathbf{y}}{\int \prod_{p=1}^P \frac{1}{(4\sigma_n^4 y_p^2)^N} \exp\left[-\frac{|\mathbf{n}_{AF,p}^i|^2}{4\sigma_n^4 y_p^2}\right] f_v(\mathbf{y}) d\mathbf{y}}, \quad (12)$$

where $x_{j,p}^i$ is the j -th entry of the vector $\mathbf{x}_{N,p}^i$, the integrals in (12) are over the set $[0, \infty)^P$, $\mathbf{v} = [v_1, \dots, v_p]$, $\mathbf{y} = [y_1, \dots, y_p]$, and $d\mathbf{y} = \prod_{i=1}^p dy_i$. The convergence in measure of (12) for increasing N to the likelihood functional $\Lambda[\mathbf{x}(t); H_i]$ is ensured by the Grenander theorem (Poor, 1988). In order to evaluate the above functional, we introduce the substitution

$$y_p = \frac{\|\mathbf{x}_{N,p}^i\|}{\sqrt{4\sigma_n^4 z_p}}, \quad p = 1, 2, \dots, P \quad (13)$$

where $\|\cdot\|$ denotes the Euclidean norm. Applying the same limiting procedure as in (Buzzi, 1999), we come up with the following asymptotical expression:

$$\Lambda[\mathbf{x}(t); H_i] = \lim_{N \rightarrow \infty} \prod_{p=1}^P \Lambda_{gN}^p \left[\mathbf{x}_{N,p}^i, \frac{\|\mathbf{x}_{N,p}^i\|^2}{4\sigma_n^4 N}; H_i \right], \quad (14)$$

where

$$\Lambda_{gN}^p(\mathbf{x}_{N,p}^i, y_p^2; H_i) = \exp \left\{ \sum_{j=1}^N \left[\frac{\sigma_{s_{j,i}}^2 |x_{j,p}^i|^2}{4\sigma_n^4 y_p^2 (\sigma_{s_{j,i}}^2 + 4\sigma_n^4 y_p^2)} - \ln \left(1 + \frac{\sigma_{s_{j,i}}^2}{4\sigma_n^4 y_p^2} \right) \right] \right\} \quad (15)$$

represents the ratio between the conditional likelihoods for H_i and H_0 based on the observation of the signal received on the p -th channel output only. Equation (14) also requires evaluating

$$Z_p = \lim_{N \rightarrow \infty} \frac{\|\mathbf{x}_{N,p}^i\|^2}{N}, \quad (16)$$

that, following in (Buzzi, 1999), can be shown to converge in the mean square sense to the random variable $4\sigma_n^4 v_p^2$ for any of the Karhunen-Loeve basis $B_i, i = 1, \dots, M$. Due to the fact that the considered noise is white, this result also holds for the large signal-to-noise ratios even though, in this case, a large number of summands is to be considered in order to achieve a given target estimation accuracy. Notice also that $4\sigma_n^4 v_p^2$ can be interpreted as a short-term noise power spectral density (PSD), namely, the PSD that would be measured on sufficiently short time intervals on the p -th channel output. Thus, the classification problem under study admits the sufficient statistics

$$\ln \Lambda[\mathbf{x}(t); H_i] = \sum_{p=1}^P \sum_{j=1}^{\infty} \left[\frac{\sigma_{s_{j,i}}^2 |x_{j,p}^i|^2}{Z_p(\sigma_{s_{j,i}}^2 + Z_p)} - \ln \left(1 + \frac{\sigma_{s_{j,i}}^2}{Z_p} \right) \right]. \quad (17)$$

The above equations demonstrate that the optimum GD structure for the problem given in (1) is completely canonical in that for any $f_{\nu_1, \dots, \nu_p}(\nu_1, \dots, \nu_p)$ and, for any noise model in the class of compound-Gaussian processes and for any correlation of the random variables ν_1, \dots, ν_p , the likelihood functional is one and the same. Equation (17) can be interpreted as a bank of P estimator-GDs (Van Trees, 2003) plus a bias term depending on the eigenvalues of the signal correlation under the hypothesis H_i . The optimum test based on GASP can be written in the following form:

$$\begin{aligned} \hat{H} = H_i &\Rightarrow \sum_{p=1}^P \frac{1}{Z_p} \left\{ \int_0^T [2x_p(t)\hat{s}_{p,i}^*(t) - x_p(t)x_p(t-\tau)]dt + \int_0^T n_{AF_p}^2(t)dt \right\} - b_{p,i} \\ &> \sum_{p=1}^P \frac{1}{Z_p} \left\{ \int_0^T [2x_p(t)\hat{s}_{p,k}^*(t) - x_p(t)x_p(t-\tau)]dt + \int_0^T n_{AF_p}^2(t)dt \right\} - b_{p,k}, \quad \forall k \neq i \end{aligned} \quad (18)$$

where $\hat{s}_{p,i}(t)$ is the linear minimum mean square estimator of $s_{p,i}(t)$ embedded in white noise with PSD Z_p , namely,

$$\hat{s}_{p,i}(t) = \int_0^T h_{p,i}(t, u)x_p(u)du, \quad (19)$$

where $h_{p,i}(t, u)$ is the solution to the Wiener-Hopf equation

$$\int_0^T \text{Cov}_i(t, z)h_{p,i}(z, \tau)dz + Z_p h_{p,i}(t, \tau) = \text{Cov}_i(t, \tau). \quad (20)$$

As to the bias terms $b_{p,i}$, they are given by

$$b_{p,i} = \sum_{j=1}^{\infty} \ln \left[1 + \frac{\sigma_{s_{j,i}}^2}{Z_p} \right] \quad i = 1, \dots, M \quad \text{and} \quad p = 1, \dots, P. \quad (21)$$

The block diagram of the corresponding GD is outlined in Fig.3. The received signals $x_1(t), \dots, x_p(t)$ are fed to P estimators of the noise short-term PSD, which are subsequently used for synthesizing the bank of MP minimum mean square error filters $h_{p,i}(\cdot, \cdot), \forall p = 1, \dots, P, \forall i = 1, \dots, M$ to implement the test (18). The newly proposed GD structure is a generalization, to the case of multiple observations, of that proposed in (Buzzi, 1999), to which it reduces to $P = 1$.

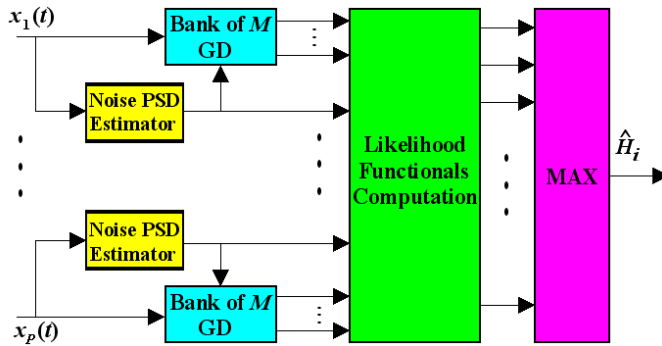


Fig. 3. Flowchart of optimum GD in compound Gaussian noise.

3.2.2 Suboptimal GD: Low energy coherence approach

Practical implementation of the decision rule (18) requires an estimation of the short-term noise PSDs on each diversity branch and evaluation of the test statistic. This problem requires a real-time design of MP estimator-GDs that are keyed to the estimated values of the short-term PSDs. This would require a formidable computational effort, which seems to prevent any practical implementation of the new receiving structure. Accordingly, we develop an alternative suboptimal GD structure with lower complexity. Assume that the signals $\{s_{p,i}(t) : \forall p = 1, \dots, P, \forall i = 1, \dots, M\}$ possess a low degree of coherence, namely, that their energy content is spread over a large number of orthogonal directions. Since

$$\bar{E}_i = \sum_{j=1}^{+\infty} \sigma_{s_{j,i}}^2, \quad (22)$$

The low degree of coherence assumption implies that the covariance functions $Cov_i(t, \tau)$ have a large number of nonzero eigenvalues and do not have any dominant eigenvalue. Under these circumstances, it is plausible to assume that the following low energy coherence condition is met:

$$\sigma_{s_{j,i}}^2 \ll 2N_0 \quad i = 1, \dots, M \quad j = 1, 2, \dots \quad (23)$$

If this is the case, we can approximate the log-likelihood functional (17) with its first-order McLaurin series expansion with starting point $\sigma_{s_{j,i}}^2 / Z_p = 0$. Following the same steps as in (Buzzi, 1997), we obtain the following suboptimal within the limits GASP decision-making rule:

$$\begin{aligned} \hat{H} = H_i &\Rightarrow \sum_{p=1}^P \frac{1}{Z_p^2} \left\{ \int_0^T \int_0^T x_p(t) x_p^*(\tau) Cov_i(t, \tau) dt d\tau + \int_0^T n_{AF_p}^2(t) dt \right\} - \frac{\bar{E}_i}{Z_p} \\ &> \sum_{p=1}^P \frac{1}{Z_p^2} \left\{ \int_0^T \int_0^T x_p(t) x_p^*(\tau) Cov_k(t, \tau) dt d\tau + \int_0^T n_{AF_p}^2(t) dt \right\} - \frac{\bar{E}_k}{Z_p} \quad \forall k \neq i. \end{aligned} \quad (24)$$

The new GD again requires estimating the short-term noise PSDs Z_1, \dots, Z_p . Unlike the optimum GD (18), in the suboptimum GD (24), the MP minimum mean square error filters $h_{p,i}(\cdot, \cdot), \forall p = 1, \dots, P, \forall i = 1, \dots, M$ whose impulse responses depend on Z_1, \dots, Z_p through (20) are now replaced with M filters whose impulse response $Cov_i(t, \tau)$ is independent of the short-term noise PSDs realizations, which now affect the decision-making rule as mere proportionality factors. The only difficulty for practical implementation of such a GD scheme is the short-term noise PSD estimation through (16). However, as already mentioned, such a drawback can be easily circumvented by retaining only a limited number of summands.

3.3 Special cases

3.3.1 Channels with flat-flat Rayleigh fading

Let us consider the situation where the fading is slow and non-selective so that the signal observed on the p -th channel output under the hypothesis H_i takes the form

$$s_{p,i}(t) = A_p \exp\{j\theta_p\} u_i(t), \quad (25)$$

where $A_p \exp\{j\theta_p\}$ is a complex zero-mean Gaussian random variable. The signal covariance function takes a form:

$$Cov_i(t, \tau) = \bar{E}_i u_i(t) u_i^*(\tau), \quad (26)$$

where the assumption has been made that $u_i(t)$ possesses unity norm. Notice that this equation represents the Mercer expansion of the covariance in a basis whose first unit vector is parallel to $u_i(t)$. It should be noted that since the Mercer expansion of the useful signal covariance functions contains just one term, the low energy coherence condition is, in this case, equivalent to a low SNR condition. It thus follows that the low energy coherence GD can be now interpreted as a locally optimum GD, thus implying that for large $SNRs$, its performance is expectedly much poorer than that of the optimum GD. The corresponding eigenvalues are

$$\sigma_{s_{1,i}}^2 = \bar{E}_i, \quad \sigma_{s_{k,i}}^2 = 0, \quad \forall k \neq 1. \quad (27)$$

Accordingly, the minimum mean square error filters to be substituted in (18) have the following impulse responses:

$$h_{p,i}(t, \tau) = \frac{\bar{E}_i}{\bar{E}_i + Z_p} u_i(t) u_i^*(\tau), \quad (28)$$

where the bias term is simply $b_{p,i} = \ln\left\{1 + \frac{\bar{E}_i}{Z_p}\right\}$. We explicitly notice here that such a bias term turns out to depend on the estimated PSD Z_p . Substituting into (18), we find the optimum test

$$\hat{H} = H_i \Rightarrow \sum_{p=1}^P \frac{\bar{E}_i}{Z_p(\bar{E}_i + Z_p)} \left| \int_0^T [2x_p(t) u_i^*(t) - x_p(t) x_p(t - \tau)] dt + \int_0^T n_{AF_p}^2(t) dt \right|^2 - b_{p,i}$$

$$> \max_{k \neq i} \sum_{p=1}^P \frac{\bar{E}_k}{Z_p(\bar{E}_k + Z_p)} \left| \int_0^T [2x_p(t)u_k^*(t) - x_p(t)x_p(t-\tau)]dt + \int_0^T n_{AF_p}^2(t)dt \right|^2 - b_{p,k}, \quad (29)$$

whereas its low energy coherence suboptimal approximation can be written in the following form:

$$\begin{aligned} \hat{H} = H_i &\Rightarrow \sum_{p=1}^P \frac{1}{Z_p^2} \left| \int_0^T [2x_p(t)u_i^*(t) - x_p(t)x_p(t-\tau)]dt + \int_0^T n_{AF_p}^2(t)dt \right|^2 - \frac{\bar{E}_i}{Z_p} \\ &> \max_{k \neq i} \sum_{p=1}^P \frac{1}{Z_p^2} \left| \int_0^T [2x_p(t)u_i^*(t) - x_p(t)x_p(t-\tau)]dt + \int_0^T n_{AF_p}^2(t)dt \right|^2 - \frac{\bar{E}_i}{Z_p}. \end{aligned} \quad (30)$$

It is worth pointing out that both GDs are akin to the “square-law combiner” (Tuzlukov, 2001) GD that is well known to be the optimum GD in GASP (Tuzlukov, 2005 and Tuzlukov, 2012) viewpoint for array signal detection in Rayleigh flat-flat fading channels and Gaussian noise. The relevant difference is due to the presence of short-term noise PSDs Z_1, \dots, Z_p which weigh the contribution from each diversity branch. In the special case of equienergy signals, the bias terms in the above decision-making rules end up irrelevant, and the optimum GD test (29) reduces to a generalization of the usual incoherent GD, with the exception that the decision statistic depends on the short-term noise PSD realizations.

3.3.2 Channels with slow frequency-selective Rayleigh fading

Now, assume that the channel random impulse response can be written in the following form:

$$\chi_p(t, \tau) = \chi_p(\tau) = \sum_{k=0}^{L-1} A_{p,k} \exp\{j\theta_{p,k}\} \delta(\tau - kW^{-1}), \quad (31)$$

where $A_{p,k} \exp\{j\theta_{p,k}\}$ is a set of zero-mean, independent complex Gaussian random variables, and L is the number of paths. Equation (31) represents the well known tapped delay line channel model, which is widely encountered in wireless mobile communications. It is readily shown that in such a case, the received useful signal, upon transmission of $u_i(t)$, has the following covariance function:

$$\text{Cov}_i(t, \tau) = \sum_{k=0}^{L-1} \overline{A_k^2} u_i(t - kW^{-1}) u_i^*(\tau - kW^{-1}), \quad i = 1, \dots, M \quad (32)$$

where $\overline{A_k^2}$ is the statistical expectation (assumed independent of p) of the random variables $A_{p,k}^2$. These correlations admit L nonzero eigenvalues, and a procedure for evaluating their eigenvalues and eigenfunctions can be found in (Matthews, 1992). In the special case that the L paths are resolvable, i.e. $T \leq W^{-1}$, the optimum GD (18) assumes the following simplified form:

$$\begin{aligned}
\hat{H} = H_i &\Rightarrow \sum_{p=1}^P \sum_{j=0}^{L-1} \left\{ \frac{\left| \overline{A_j^2} \int_0^T x_p(t) u_i^*(t - jW^{-1}) dt + \int_0^T n_{A_{F_p}}^2(t) dt \right|^2}{Z_p (Z_p + \overline{E_i A_j^2})} - \ln \left\{ 1 + \frac{\overline{E_i A_j^2}}{Z_p} \right\} \right\} \\
&> \max_{k \neq i} \sum_{p=1}^P \sum_{j=0}^{L-1} \left\{ \frac{\left| \overline{A_j^2} \int_0^T x_p(t) u_k^*(t - jW^{-1}) dt + \int_0^T n_{A_{F_p}}^2(t) dt \right|^2}{Z_p (Z_p + \overline{E_k A_j^2})} - \ln \left\{ 1 + \frac{\overline{E_k A_j^2}}{Z_p} \right\} \right\}, \tag{33}
\end{aligned}$$

where $\overline{E_i}$ is the energy of the signal $u_i(t)$. The low energy coherence suboptimal GD (24) is instead written as

$$\begin{aligned}
\hat{H} = H_i &\Rightarrow \sum_{p=1}^P \sum_{j=0}^{L-1} \left\{ \frac{\left| \frac{\overline{A_j^2}}{Z_p^2} \int_0^T x_p(t) u_i^*(t - jW^{-1}) dt + \int_0^T n_{A_{F_p}}^2(t) dt \right|^2 - \frac{\overline{E_i A_j^2}}{Z_p}}{Z_p} \right\} \\
&> \max_{k \neq i} \sum_{p=1}^P \sum_{j=0}^{L-1} \left\{ \frac{\left| \frac{\overline{A_j^2}}{Z_p^2} \int_0^T x_p(t) u_k^*(t - jW^{-1}) dt + \int_0^T n_{A_{F_p}}^2(t) dt \right|^2 - \frac{\overline{E_k A_j^2}}{Z_p}}{Z_p} \right\}. \tag{34}
\end{aligned}$$

Optimality of (33) obviously holds for one-shot detection, namely, neglecting the intersymbol interference induced by the channel band limitedness.

3.4 Performance assessment

In this section, we focus on the performance of the proposed GD structures. A general formula to evaluate the probability of error P_E of any receiver in the presence of spherically invariant disturbance takes the following form:

$$P_E = \int P_E(e | \mathbf{v}) f_{\mathbf{v}}(\mathbf{v}) d\mathbf{v}, \tag{35}$$

where $P_E(e | \mathbf{v})$ is the receiver probability of error in the presence of Gaussian noise with PSD on the p -th diversity branch $2N_0 \sigma_p^2$. The problem to evaluate P_E reduces to that of first analyzing the Gaussian case and then carrying out the integration (35). In order to give an insight into the GD performance, we consider a BFSK signaling scheme, i.e. the baseband equivalents of the two transmitted waveforms are related as

$$u_2(t) = u_1(t) \exp\{j2\pi\Delta f t\}, \tag{36}$$

where $\Delta f = T^{-1}$ denotes the frequency shift. Even for this simple case study, working out an analytical expression for the probability of error of both the optimum GD and of its low energy coherence approximation is usually unwieldy even for the case of Gaussian noise. With

regard to the optimum GD structure, upper and lower bounds for the performance may be established via Chernoff-bounding techniques. Generalizing to the case of multiple observations, the procedure in (Van Trees, 2003), the conditional probability of error given ν_1, \dots, ν_p can be bounded as

$$\frac{\exp\{2\mu(0.5|\mathbf{v})\}}{2\left(1+\sqrt{0.25\pi\dot{\mu}(0.5|\mathbf{v})}\right)} \leq P_E(e|\mathbf{v}) \leq \frac{\exp\{2\mu(0.5|\mathbf{v})\}}{2\left(1+\sqrt{0.25\pi\dot{\mu}(0.5|\mathbf{v})}\right)}, \quad (37)$$

where $\mu(\cdot|\mathbf{v})$ is the following conditional semi-invariant moment generating the function

$$\begin{aligned} \mu(x|\mathbf{v}) &= \lim_{N \rightarrow \infty} \ln E \left\{ \exp \left[x \sum_{p=1}^P \ln \Lambda_{gN}^p(\mathbf{x}_{N,p}; \nu_p^2 : H_1) \right] \middle| H_0, \mathbf{v} \right\} \\ &= \sum_{j=1}^{\infty} \sum_{p=1}^P \left\{ (1-x) \ln \left[1 + \frac{\sigma_{s_j}^2}{4\sigma_n^4 \nu_p^2} \right] - \ln \left[1 + \frac{\sigma_{s_j}^2(1-x)}{4\sigma_n^4 \nu_p^2} \right] \right\} \end{aligned} \quad (38)$$

with $\{\sigma_{s_j}^2\}_{j=1}^{\infty}$ being the set of common eigenvalues. Substituting this relationship into (37) and averaging with respect to ν_1, \dots, ν_p yields the unconditional bounds on the probability of error for the optimum GD (18).

3.5 Simulation results

To proceed further in the GD performance there is a need to assign both the marginal pdf, as well as the channel spectral characteristics. We assume hereafter the generalized Laplace noise, i.e. the marginal pdf of the p -th noise texture component takes the following form:

$$f_{\nu_p}(x) = \frac{2\nu^\nu}{\Gamma(\nu)} x^{2\nu-1} \exp\{-\nu x^2\}, \quad x > 0 \quad (39)$$

where ν is a shape parameter, ruling the distribution behavior. In particular, the limiting case $\nu \rightarrow \infty$ implies $f_{\nu_p}(x) = \delta(x-1)$ and, eventually, Gaussian noise, where increasingly lower values of ν account for increasingly spikier noise distribution. Regarding the channel, we consider the case of the frequency-selective, slowly fading channel, i.e. the channel random impulse response is expressed by (31), implying that the useful signal correlation is that given in (32). For simplicity, we also assume that the paths are resolvable. In the following plots the P_E is evaluated a) through a semianalytic procedure, i.e. by numerically averaging the Chernoff bound (37) with respect to the realizations of the ν_1, \dots, ν_p , and b) by resorting to a Monte Carlo counting procedure. In this later case, the noise samples have been generated by multiplying standard, i.e. with zero-mean and unit-variance, complex Gaussian random variates times the random realizations of ν_1, \dots, ν_p .

The Chernoff bound for the optimum GD versus the averaged received radio-frequency energy contrast that is defined as $\gamma_0 = P \sum_{j=1}^L \sigma_{s_j}^2 (4\sigma_n^4)^{-0.5}$ at $P = 2$ and for two values of the noise shape parameter ν is shown in Fig.4. The noise texture components have been assumed to be independent. Inspecting the curves, we see that the Chernoff bound provides a very reliable

estimate of the actual P_E , as the upper and lower bound very tightly follow each other. As expected, the results demonstrate that in the low P_E region, the spikier the noise, i.e. the lower ν , the worse the GD performance. Conversely, the opposite behavior is observed for small values of γ_0 . This fact might appear, at a first look, surprising. It may be analytically justified in light of the local validity of Jensen's inequality (Van Trees, 2003) and is basically the same phenomenon that makes digital modulation schemes operating in Gaussian noise to achieve, for low values of γ_0 , superior performance in Rayleigh flat-flat fading channels than in no-fading channels. Notice, this phenomenon is in accordance with that observed in (Conte 1995). In order to validate the Chernoff bound, we also show, on the same plots, some points obtained by Monte Carlo simulations. These points obviously lie between the corresponding upper and lower probability of error bounds. Additionally, we compare the GD Chernoff bound with that for the conventional optimum receiver (Buzzi et al., 2001). A superiority of GD structure is evident.

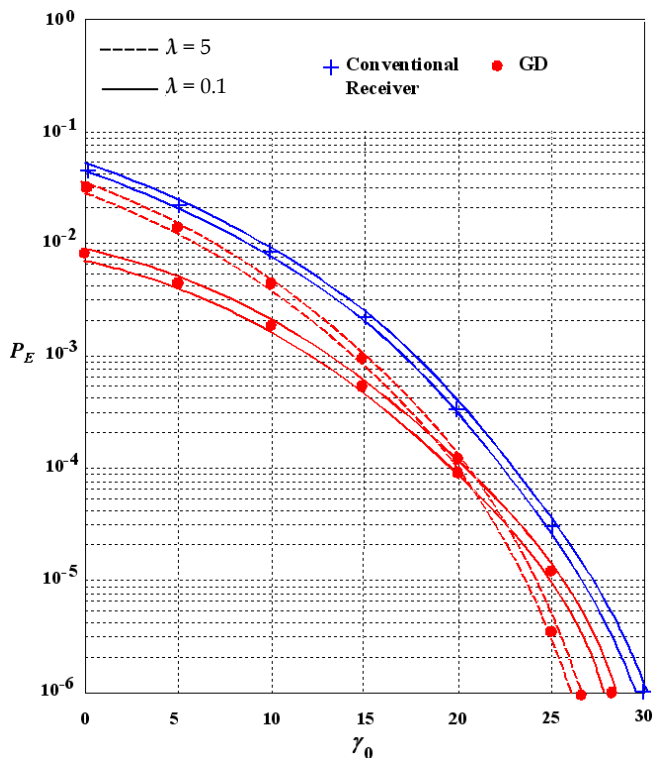


Fig. 4. Chernoff bounds for P_E of the optimum GD.

In Fig. 5, the effect of the channel diversity order is investigated. Indeed, the optimum GD performance versus γ_0 is represented for several values of P and with $\nu = 1$. The Z_1, \dots, Z_P have been assumed exponentially correlated with correlation coefficient $\rho = 0.2$. A procedure for generating these exponentially correlated random variables for integer and semi-integer values of ν is reported in (Lombardo et al., 1999). As expected, as P increases, the GD performance ameliorates, thus confirming that diversity represents a suitable means to restore performance in severely hostile scenarios. Also, we compare the GD performance with that for the conventional optimum receiver (Buzzi et al., 2001) and we see that the GD keeps superiority in this case, too.

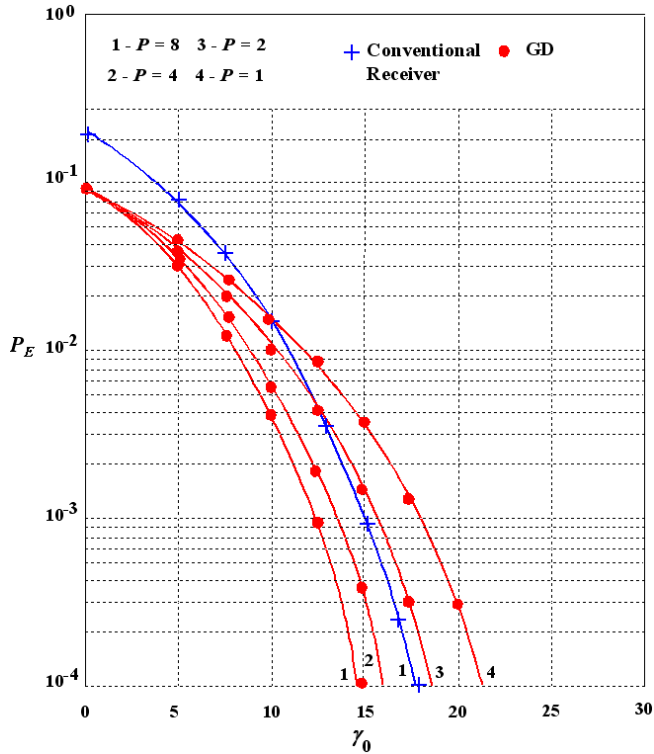


Fig. 5. P_E at several values of P .

The optimum GD performance versus γ_0 for the generalized Laplace noise at $\nu = 1, P = 4$ and for several values of the correlation coefficient ρ is demonstrated in Fig. 6. It is seen that the probability of error improves for vanishingly small ρ . For small ρ , the GD takes much advantage of the diversity observations. For high values of ρ , the realizations Z_1, \dots, Z_p are very similar and much less advantage can be gained through the adoption of a diversity strategy. Such GD performance improvement is akin to that observed in signal diversity detection in the presence of flat-flat fading and Gaussian noise. We see that the GD outperforms the conventional optimum receiver (Buzzi et al., 2001) by the probability of error.

In Fig. 7, we compare the optimum GD performance versus that of the low energy coherence GD. We assumed $\rho = 0.2$ and $P = 4$. It is seen that the performance loss incurred by the low energy coherence GD with respect to the optimum GD is kept within a fraction of 1 dB at $P_E = 10^{-4}$. Simulation results that are not presented in the paper show that the crucial factor ruling the GD performance is the noise shape parameter, whereas the particular noise distribution has a rather limited effect on the probability of error.

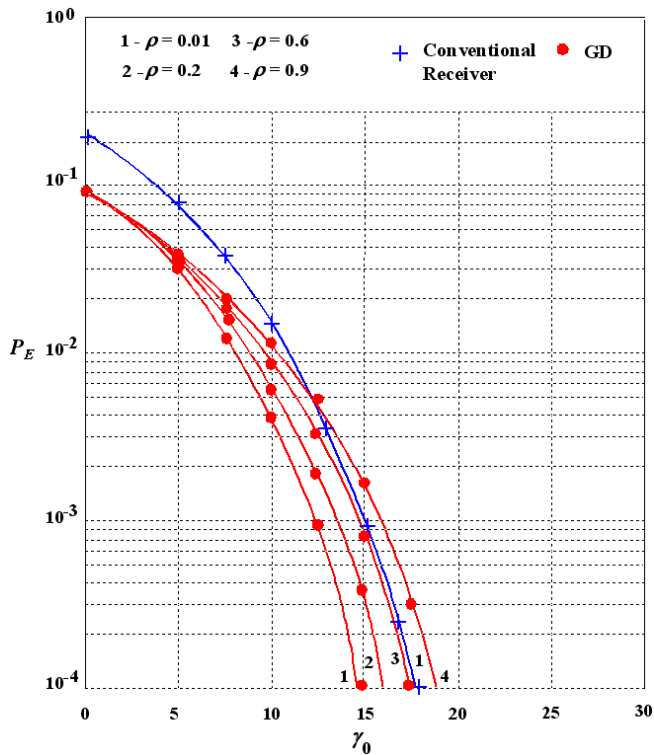


Fig. 6. P_E at several values of the correlation coefficient.

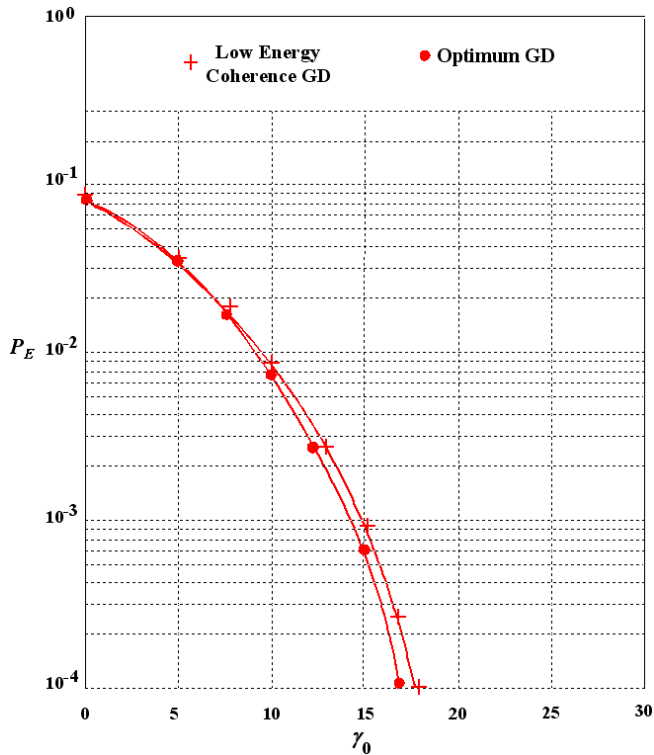


Fig. 7. P_E for the optimum and low energy coherence GDs.

3.6 Discussion

We have considered the problem of diversity detection of one out of M signals transmitted over a fading dispersive channel in the presence of non-Gaussian noise. We have modeled the additive noise on each channel diversity branch through a spherically invariant random process, and the optimum GD has been shown to be independent of the actual joint pdf of the noise texture components present on the channel diversity outputs. The optimum GD is similar to the optimum GD for Gaussian noise, where the only difference is that the noise PSD $2N_0$ is substituted with a perfect estimate of the short-term PSD realizations of the impulsive additive noise. We also derived a suboptimum GD matched with GASP based on the low energy coherence hypothesis. At the performance analysis stage, we focused on frequency-selective slowly fading channels and on a BFSK signaling scheme and evaluated the GD performance through both a semianalytic bounding technique and computer simulations. Numerical results have shown that the GD performance is affected by the average received energy contrast, by the channel diversity order, and by the noise shape parameter, whereas it is only marginally affected by the actual noise distribution. Additionally, it is seen that in impulsive environments, diversity represents a suitable strategy to improve GD performance.

4. MIMO radar systems applied to wireless communications based on GASP

Multiple-input multiple-output (MIMO) wireless communication systems have received a great attention owing to the following viewpoints: a) MIMO wireless communication systems have been deemed as efficient spatial multiplexers and b) MIMO wireless communication systems have been deemed as a suitable strategy to ensure high-rate communications on wireless channels (Foschini, 1996). Space-time coding has been largely investigated as a viable means to achieve spatial diversity, and thus to contrast the effect of fading (Tarokh, et al., 1998 and Hochwald, et al., 2000). We apply GASP to the design and implementation of MIMO wireless communication systems used space-time coding technique. Theoretical principles of MIMO wireless communication systems were discussed and the potential advantages of MIMO wireless communication systems are thoroughly considered in (Fishler, et al., 2006).

MIMO architecture is able to provide independent diversity paths, thus yielding remarkable performance improvements over conventional wireless communication systems in the medium-high range of detection probability. As was shown in (Fishler, et al., 2006), the MIMO mode can be conceived as a means of bootstrapping to obtain greater coherent gain. Some practical issues concerning implementation (equipment specifications, dynamic range, phase noise, system stability, isolation and spurs) of MIMO wireless communication systems are discussed in (Skolnik, 2008).

MIMO wireless communication systems can be represented by m transmit antennas, spaced several wavelengths apart, and n receive antennas, not necessarily collocated, and possibly forwarding, through a wired link, the received echoes to a fusion center, whose task is to make the final decision about the signal in the input waveform. If the spacing between the transmit antennas is large enough and so is the spacing between the receive antennas, a rich scattering environment is generated, and each receive antenna processes l statistically independent copies of incoming signal. The concept of rich scattering environment is borrowed from communication theory, and models a situation where the MIMO architecture yields interchannel interference, eventually resulting into a number of independent random channels. Unlike a conventional wireless communication array system, which attempts to maximize the coherent processing, MIMO wireless communication system resorts to the fading diversity in order to improve the detection performance. Indeed, it is well known that, in conventional wireless communication array system, multiple access interference (MAI) of the order of 10 dB may arise. This effect leads to severe degradations of the detection performance, due to the high signal correlation at the array elements. This drawback might be partially circumvented under the use of MIMO wireless communication system, which exploits the channel diversity and fading. Otherwise, uncorrelated signals at the array elements are available. Based on mentioned above statements, it was shown in (Fishler, et al., 2006) that in the case of additive white Gaussian noise (AWGN), transmitting orthogonal waveforms result into increasingly constrained fluctuation of interference.

Our approach is based on implementation of GASP and employment of some key results from communication theory, and in particular, the well-known concept that, upon suitably space-time encoding the transmitted waveforms, a maximum diversity order given by $m \times n$ can be achieved. Importing these results in a wireless communication system scenario poses

a number of problems, which forms the object of the present study, and in particular: a) the issue of waveform design, which exploits the available knowledge as to space-time codes; b) the issue of designing a suitable detection structure based on GASP, also in the light of the fact that the disturbance can no longer be considered as AWGN, due to the presence of interferences; and c) at the performance assessment level, the issue of evaluating the maximum diversity order that can be achieved and the space-time coding ensuring it under different types of interferences. The first and third tasks are merged in the unified problem of determining the space-time coding achieving maximum diversity order in signal detection, for constrained BER, and for given interference covariance. As to the second task, the decision-making criterion exploiting by GASP is employed.

Unlike (Fishler, et al., 2006), no assumption is made on either the signal model or the disturbance covariance. Thus, a family of detection structures is derived, depending upon the number of transmitting and receiving antennas and the disturbance covariance. A side result, which paves the way to further investigations on the feasibility of fully adaptive MIMO wireless communication systems is that the decision statistic, under the null hypothesis of no signal, is an ancillary statistic, in the sense that it depends on the actual interference covariance matrix, but its probability density function (pdf) is functionally independent of such a matrix. Therefore, threshold setting is feasible with no prior knowledge as to the interference power spectrum. As to the detection performance, a general integral form of the probability of detection P_D is provided, holding independent of the signal fluctuation model. The formula is not analytically manageable, nor does it appear to admit general approximate expressions, that allow us to give an insightful look in the wireless communication system behavior. We thus restrict our attention to the case of Rayleigh-distributed attenuation, and use discussed in (De Maio & Lops, 2007) an information-theoretic approach to code construction, which, surprisingly enough, leads to the same solution found through the optimization of the Chernoff bound.

4.1 System model

We consider MIMO radar system composed of m fixed transmitters and n fixed receivers and assume that the antennas as the two ends of the wireless communication system are sufficiently spaced such that a possible incoming message and/or interference provides uncorrelated reflection coefficients between each transmit/receive pair of sensors. Denote by $s_i(t)$ the baseband equivalent of the coherent pulse train transmitted by the i -th antenna, for example,

$$s_i(t) = \sum_{j=1}^N a_{i,j} p[t - (j-1)T_p], \quad i = 1, \dots, m \quad (40)$$

where $p(t)$ is the signature of each transmitted pulse, which we assume, without loss of generality, with unit energy and duration τ_p ; T_p is the pulse repetition time;

$$\mathbf{a}_i = [a_{i,1}, \dots, a_{i,N}]^T \quad (41)$$

is an N -dimensional column vector whose entries are complex numbers which modulate both in amplitude and in phase the N pulses of the train, where $(\cdot)^T$ denotes transpose. In the sequel, we refer to \mathbf{a}_i as the code word of the i -th antenna. The baseband equivalent of the signal received by the i -th sensor, from a target with two-way time delay τ , can be presented in the following form

$$x_i(t) = \sum_{l=1}^m a_{i,l} \sum_{j=1}^N a_{i,j} p[t - \tau - (j-1)T_p] + n_i(t), \quad i = 1, \dots, n, \quad (42)$$

where $a_{i,l}$, $i = 1, \dots, n$ and $l = 1, \dots, m$, are complex numbers accounting for both the target backscattering and the channel propagation effects between the l -th transmitter and the i -th receiver; $n_i(t)$, $i = 1, \dots, n$, are zero-mean, spatially uncorrelated, complex Gaussian random processes accounting for both the external and the internal disturbance. For simplicity, we assume a zero-Doppler target, but all the derivations can be easily extended to account for a possible known Doppler shift. We explicitly point out that the validity of the above model requires the narrowband assumption

$$\frac{d_{\max}^m + d_{\max}^n}{c} \ll \frac{1}{B} \quad (43)$$

where B is the bandwidth of the transmitted pulse, d_{\max}^m and d_{\max}^n denote the maximum spacing between two sensors at the transmitter and the receiver end, respectively. The signal $x_i(t)$, at each of the receive elements, is matched filtered to the pulse $p(t)$ by preliminary filter of the GD and the filter output is sampled at the time instants $\tau + (k-1)T_p$, $k = 1, \dots, N$. Thus, denote by $x_i(k)$ the k -th sample, i.e.,

$$x_i(k) = \sum_{l=1}^m a_{i,l} a_{l,k} + n_i(k), \quad (44)$$

where $n_i(k)$ is the filtered noise sample. Define the N -dimensional column vectors

$$\mathbf{x}_i = [x_i(1), \dots, x_i(N)]^T \quad (45)$$

and rewrite them as

$$\mathbf{x}_i = \mathbf{A} \boldsymbol{\alpha}_i + \boldsymbol{\xi}_{PF_i}, \quad i = 1, \dots, n \quad (46)$$

where

$$\boldsymbol{\xi}_{PF_i} = [\xi_{PF_i}(1), \dots, \xi_{PF_i}(N)]^T, \quad (47)$$

$$\boldsymbol{\alpha}_i = [a_{i,1}, \dots, a_{i,m}]^T, \quad (48)$$

and the $(N \times m)$ -dimensional matrix \mathbf{A} , defined in the following form

$$\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_m] \quad (49)$$

has the code words as columns. This last matrix is referred to as the code matrix. We assume that \mathbf{A} is full rank matrix. It is worth underlining that the model given by (46) applies also to the case that space-time coding is performed according to (De Maio & Lops, 2007), namely, by dividing a single pulse in N sub-pulses. The code matrix \mathbf{A} thus defines m different code words of length N , which can be received by a single receive antenna, thus defining the multiple-input single-output (MISO) structure, as well as by a set of n receive antennas, as in the present study.

4.2 GD design for MIMO radar systems applied to wireless communications

The problem of detecting a target return signal with a MIMO radar system can be formulated in terms of the following binary hypothesis test

$$\begin{cases} H_0 \Rightarrow \mathbf{x}_i = \xi_{PF_i}, & i = 1, \dots, n \\ H_1 \Rightarrow \mathbf{x}_i = \mathbf{A}\alpha_i + \xi_{PF_i}, & i = 1, \dots, n \end{cases} \quad (50)$$

where ξ_{PF_i} , $i = 1, \dots, n$, are statistically independent and identically distributed (i.i.d.) zero-mean complex Gaussian vectors with covariance matrix

$$E[\xi_{PF_i} \xi_{PF_i}^*] = E[\xi_{AF_i} \xi_{AF_i}^*] = \mathbf{M}. \quad (51)$$

Here $E[\cdot]$ denotes the statistical expectation and $(*)$ denotes conjugate transpose. The covariance matrix (51) is assumed positive definite and known. According to the Neyman-Pearson criterion, the optimum solution to the hypotheses testing problem (50) must be the likelihood ratio test. However, for the case at hand, it cannot be implemented since total ignorance of the parameters α_i is assumed. One possible way to circumvent this drawback is to resort to the generalized likelihood ratio test (GLRT) (Van Trees, 2003), which is tantamount to replacing the unknown parameters with their maximum likelihood (ML) estimates under each hypothesis. Applying GASP to the GLRT, we obtain the following decision rule

$$\frac{\max_{\alpha_1, \dots, \alpha_n} f(\mathbf{x}_1, \dots, \mathbf{x}_n | H_1, \mathbf{M}, \alpha_1, \dots, \alpha_n)}{f(\xi_{AF_1}, \dots, \xi_{AF_n} | H_0, \mathbf{M})} \underset{H_0}{\overset{H_1}{>}} K_g, \quad (52)$$

where $f(\mathbf{x}_1, \dots, \mathbf{x}_n | H_1, \mathbf{M}, \alpha_1, \dots, \alpha_n)$ is the probability density function (pdf) of the data under the hypothesis H_1 and $f(\xi_{AF_1}, \dots, \xi_{AF_n} | H_0, \mathbf{M})$ is pdf of the data under the hypothesis H_0 , respectively, K_g is a suitable modification of the original threshold. Previous assumptions imply that the aforementioned pdfs can be written in the following form:

$$f(\xi_{AF_1}, \dots, \xi_{AF_n} | H_0, \mathbf{M}) = \frac{1}{\pi^{Nn} \det^n(\mathbf{M})} \exp \left[-\sum_{i=1}^n \xi_{AF_i}^* \mathbf{M}^{-1} \xi_{AF_i} \right] \quad (53)$$

at the hypothesis H_0 and

$$f(\mathbf{x}_1, \dots, \mathbf{x}_n | H_1, \mathbf{M}, \boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n) = \frac{1}{\pi^{Nn} \det^n(\mathbf{M})} \exp \left[-\sum_{i=1}^n (\mathbf{x}_i - \mathbf{A}\boldsymbol{\alpha}_i)^* \mathbf{M}^{-1} (\mathbf{x}_i - \mathbf{A}\boldsymbol{\alpha}_i) \right] \quad (54)$$

under the hypothesis H_1 , where $\det(\cdot)$ denotes the determinant of a square matrix. Substituting (16) and (17) in (15), we can recast the GLRT based on GASP, after some mathematical transformations, in the following form

$$\sum_{i=1}^n \boldsymbol{\xi}_{AF_i}^* \mathbf{M}^{-1} \boldsymbol{\xi}_{AF_i} - \sum_{i=1}^n \min_{\boldsymbol{\alpha}_i} (\mathbf{x}_i - \mathbf{A}\boldsymbol{\alpha}_i)^* \mathbf{M}^{-1} (\mathbf{x}_i - \mathbf{A}\boldsymbol{\alpha}_i) \underset{H_0}{\overset{H_1}{>}} K_g. \quad (55)$$

In order to solve the n minimization problems in (55) we have to distinguish between two different cases.

Case 1: $N > m$. In this case, the quadratic forms in (55) achieve the minimum at

$$\hat{\boldsymbol{\alpha}}_i = (\mathbf{A}^* \mathbf{M}^{-1} \mathbf{A})^{-1} \mathbf{A}^* \mathbf{M}^{-1} \mathbf{x}_i, \quad i = 1, \dots, n \quad (56)$$

and, as a consequence, the GLRT based on GASP at the main condition of GD functioning, i.e., equality in whole range of parameters between the transmitted information signal and reference signal (signal model) in the receiver part, becomes

$$2 \sum_{i=1}^n \mathbf{x}_i^* \mathbf{M}^{-1} \mathbf{A} (\mathbf{A}^* \mathbf{M}^{-1} \mathbf{A})^{-1} \mathbf{A}^* \mathbf{M}^{-1} \mathbf{x}_i - \sum_{i=1}^n \mathbf{x}_i^* \mathbf{M}^{-1} \mathbf{A} \mathbf{A}^* \mathbf{M}^{-1} \mathbf{x}_i + \sum_{i=1}^n \boldsymbol{\xi}_{AF_i}^* \mathbf{M}^{-1} \mathbf{M}^{-1} \boldsymbol{\xi}_{AF_i} \underset{H_0}{\overset{H_1}{>}} K_g. \quad (57)$$

Case 2: $N \leq m$. In this case, the minimum of the quadratic forms in (55) is zero, since each linear system

$$\mathbf{A} \hat{\boldsymbol{\alpha}}_i = \mathbf{x}_i, \quad i = 1, \dots, n \quad (58)$$

is determined. As a consequence the GLRT based on GASP at the main condition of GD functioning, i.e., equality in whole range of parameters between the transmitted information signal and reference signal (signal model) in the receiver part, becomes

$$\sum_{i=1}^n \boldsymbol{\xi}_{AF_i}^* \mathbf{M}^{-1} \mathbf{M}^{-1} \boldsymbol{\xi}_{AF_i} - \sum_{i=1}^n \mathbf{x}_i^* \mathbf{M}^{-1} \mathbf{A} \mathbf{A}^* \mathbf{M}^{-1} \mathbf{x}_i \underset{H_0}{\overset{H_1}{>}} K_g. \quad (59)$$

4.3 Performance analysis

In order to define possible design criteria for the space-time coding, it is useful to establish a direct relationship between the probability of detection P_D and the transmitted waveform, which is thus the main goal of the present section. Under the hypothesis H_0 , the left hand side of the GLRT based on GASP can be written in the following form

$$\sum_{i=1}^n \boldsymbol{\xi}_{AF_i}^* \mathbf{M}^{-1} \boldsymbol{\xi}_{AF_i} - \sum_{i=1}^n \boldsymbol{\xi}_{PF_i}^* \mathbf{M}^{-1} \boldsymbol{\xi}_{PF_i} \quad (60)$$

and, represents the GD background noise. It follows from (Tuzlukov 2005) that the decision statistic is defined by the modified second-order Bessel function of an imaginary argument or, as it is also called, McDonald's function with $m \times n$ degrees of freedom. Thus, the decision statistic is independent of dimensionality N of the column vector given by (41) whose entries are complex numbers, which modulate both in amplitude and in phase the N pulses of the train. Consequently, the probability of false alarm P_{FA} can be evaluated in the following form

$$P_{FA} = \exp(-K_g) \sum_{k=0}^n \frac{(K_g)^k}{k!}. \tag{61}$$

This last expression allows us to note the following observations: a) the decision statistic is ancillary, in the sense that it depends on the actual interference covariance matrix, but its pdf is functionally independent of such a matrix; and b) the threshold setting is feasible with no prior knowledge as to the interference power spectrum, namely, the GLRT based on GASP ensures the constant false alarm (CFAR) property.

Under the hypothesis H_1 , given α_i , the vectors $\mathbf{x}_i, i = 1, \dots, n$, are statistically independent complex Gaussian vectors with the mean value $\mathbf{M}^{-1} \mathbf{A} \alpha_i$ and identity covariance matrix. It follows that, given α_i , the GLRT based on GASP is no the central distributed modified second-order Bessel function of an imaginary argument, with the no centrality parameter $\sum_{i=1}^n \alpha_i^* \mathbf{A}^* \mathbf{M}^{-1} \mathbf{A} \alpha_i$ and degrees of freedom $m \times n$. Consequently, the conditional probability of detection P_D based on statements in (Van Trees, 2003) and discussion in (Tuzlukov, 2005) can be represented in the following form

$$P_D = Q_{m \times n}(\sqrt{2q}, \sqrt{2K_g}), \tag{62}$$

where

$$q = \sum_{i=1}^n \alpha_i^* \mathbf{A}^* \mathbf{M}^{-1} \mathbf{A} \alpha_i \tag{63}$$

and $Q_k(\cdot, \cdot)$ denotes the generalized Marcum Q function of order k . An alternative expression for the conditional probability of detection P_D , in terms of an infinite series, can be also written in the following form:

$$P_D = \sum_{k=0}^{\infty} \frac{\exp(-q) q^k}{k!} [1 - \Gamma_{inc}(K_g, k + m \times n)], \tag{64}$$

where

$$\Gamma_{inc}(p, r) = \frac{1}{\Gamma(r)} \int_0^w \exp(-z) z^{r-1} dz \tag{65}$$

is the incomplete Gamma function. Finally, the unconditional probability of detection P_D can be obtained averaging the last expression over the pdf of $\alpha_i, i = 1, \dots, n$.

4.4 Code design by information-theoretic approach

In principle, the basic criterion for code design should be the maximization of the probability of detection P_D given by (62) over the set of admissible code matrices, i.e.,

$$\arg \max_{\mathbf{A}} E \left[Q_{m \times n} \left(\sqrt{2q}, \sqrt{2K_g} \right) \right] = \arg \max_{\mathbf{A}} E \left[Q_{m \times n} \left(\sqrt{2 \sum_{i=1}^n \alpha_i^* \mathbf{A}^* \mathbf{M}^{-1} \mathbf{A} \alpha_i}, \sqrt{2K_g} \right) \right], \quad (66)$$

where $\arg \max_{\mathbf{A}}(\cdot)$ denotes the value of \mathbf{A} , which maximizes the argument and the statistical average is over $\alpha_i, i = 1, \dots, n$. Unfortunately, the above maximization problem does not appear to admit a closed-form solution, valid independent of the fading law, whereby we prefer here to resort to the information-theoretic criterion supposed in (De Maio & Lops, 2007). Another way is based on the optimization of the Chernoff bound over the code matrix \mathbf{A} . As was shown in (De Maio & Lops, 2007), these ways lead to the same solution, which subsumes some well-known space-time coding, such as Alamouti code and, more generally, the class of space-time coding from orthogonal design (Alamouti, 1998) and (Tarokh et al., 1999), which have been shown to be optimum in the framework of communication theory. In subsequent derivations, we assume that $\alpha_i, i = 1, \dots, n$, are independent and identically distributed (i.i.d.) zero-mean complex Gaussian vectors with scalar covariance matrix, i.e.,

$$E[\alpha_i \alpha_i^*] = \sigma_a^2 \mathbf{I}, \quad (67)$$

where σ_a^2 is a real factor accounting for the backscattered useful power, and \mathbf{I} denotes the identity matrix.

Roughly speaking, the GLRT strategy overcomes the prior uncertainty as to the target fluctuations by ML estimating the complex target amplitude, and plugging the estimated value into the conditional likelihood in place of the true value. Also, it is well known that, under general consistency conditions, the GLRT converges towards the said conditional likelihood, thus achieving a performance closer and closer to the perfect measurement bound, i.e., the performance of an optimum test operating in the presence of known target parameters. Diversity, on the other hand, can be interpreted as a means to transform an amplitude fluctuation in an increasingly constrained one. It is well known, for example that, upon suitable receiver design, exponentially distributed square target amplitude may be transformed into a central chi-square fluctuation with d degrees of freedom through a diversity of order d in any domain. More generally, a central chi-square random variable with $2m$ degrees of freedom may be transformed into a central chi-square with $2m \times d$ degrees of freedom. In this framework, a reasonable design criterion for the space-time coding is the maximization of the mutual information between the signals received from the various diversity branches and the fading amplitudes experienced thereupon. Thus, denoting by $I(\alpha, \mathbf{X})$ the mutual information (Cover & Thomas, 1991) between the random matrices

$$\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n] \quad (68)$$

and

$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_k] = \mathbf{A}\boldsymbol{\alpha} + \boldsymbol{\Xi} \quad (69)$$

the quantity to be maximized is

$$I(\boldsymbol{\alpha}, \mathbf{X}) = H(\mathbf{X}) - H(\mathbf{X} | \boldsymbol{\alpha}), \quad (70)$$

where

$$\boldsymbol{\Xi} = [\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n], \quad (71)$$

$H(\mathbf{X})$ denotes the entropy of the random matrix $\boldsymbol{\Xi}$, and $H(\mathbf{X} | \boldsymbol{\alpha})$ is the conditional entropy of \mathbf{X} given $\boldsymbol{\alpha}$ (Cover & Thomas, 1991). Exploiting the statistical independence between $\boldsymbol{\alpha}$ and \mathbf{X} , we can write (70) in the following form

$$I(\boldsymbol{\alpha}, \mathbf{X}) = H(\mathbf{X}) - H(\mathbf{X} | \boldsymbol{\alpha}) = H(\mathbf{X}) - H(\boldsymbol{\Xi}), \quad (72)$$

where $H(\boldsymbol{\Xi})$ is the entropy of the random matrix $\boldsymbol{\Xi}$. Assuming that the columns of $\boldsymbol{\alpha}$ are i.i.d. zero-mean complex Gaussian vectors with covariance matrix $\sigma_a^2 \mathbf{I}$, we can write $H(\mathbf{X})$ and $H(\boldsymbol{\Xi})$, respectively, in the following form:

$$H(\mathbf{X}) = x \lg[(\pi e)^N \det(\mathbf{M} + \sigma_a^2 \mathbf{A} \mathbf{A}^*)] \quad (73)$$

and

$$H(\boldsymbol{\Xi}) = x \lg[(\pi e)^N \det(\mathbf{M})]. \quad (74)$$

As design criterion we adopt the maximization of the minimum probability of detection P_D , which can be determined as the lower Chernoff bound, under an equality constraint for the average signal-to-clutter power ratio (SCR) given by

$$SCR = \frac{1}{Nmn} E \left[\sum_{i=1}^n \boldsymbol{\alpha}_i^* \mathbf{A}^* \mathbf{M}^{-1} \mathbf{A} \boldsymbol{\alpha}_i \right] = \frac{\sigma_a^2}{Nm} \text{tr}(\mathbf{A}^* \mathbf{M}^{-1} \mathbf{A}) = \frac{\sigma_a^2}{Nm} \sum_{j=1}^m \lambda_j, \quad (75)$$

where $\text{tr}(\cdot)$ denotes the trace of a square matrix and λ_j are the elements or corresponding ordered (in decreasing order) eigenvalues of the diagonal matrix $\boldsymbol{\Lambda}$ defined by the eigenvalue decomposition $\mathbf{V}^* \mathbf{A} \mathbf{V}$ of the matrix $\mathbf{M}^{-1} \mathbf{A} \mathbf{A}^* \mathbf{M}^{-1}$, where \mathbf{V} is an $N \times N$ unitary matrix. The considered design criterion relies on the maximization of the mutual information (70) under equality constraint (75) for SCR. This is tantamount to solving the following constrained minimization problem since $H(\boldsymbol{\Xi})$ does not exhibit any functional dependence on \mathbf{A} .

$$\min_{\lambda_1, \dots, \lambda_m} \prod_{j=1}^m \left[\frac{1}{1 + \gamma(\lambda_j \sigma_a^2 + 1)} \right]^n \quad \text{and} \quad \frac{\sigma_a^2}{Nm} \sum_{j=1}^m \lambda_j = \mu \quad (76)$$

which, taking the logarithm, is equivalent

$$\max_{\lambda_1, \dots, \lambda_m} \sum_{j=1}^m \lg[1 + \gamma(\sigma_a^2 \lambda_j + 1)] \quad \text{and} \quad \sum_{j=1}^m \lambda_j = \frac{\mu m N}{\sigma_a^2}, \quad (77)$$

where γ is the variable defining the upper Chernoff bound (Benedetto & Biglieri, 1999).

Since $\lg[1 + \gamma(\sigma_a^2 y + 1)]$ is a concave function of y , we can apply Jensen's inequality (Cover & Thomas, 1991) to obtain

$$\sum_{j=1}^m \lg[1 + \gamma(\sigma_a^2 \lambda_j + 1)] \leq m \lg \left[1 + \gamma \left(\frac{1}{m} \sum_{j=1}^m \lambda_j \sigma_a^2 + 1 \right) \right]. \quad (78)$$

Moreover, forcing in the right hand side of (78), the constraint of (77), we obtain

$$\sum_{j=1}^m \lg[1 + \gamma(\sigma_a^2 \lambda_j + 1)] \leq m \lg[1 + \gamma(\mu N + 1)]. \quad (79)$$

The equality in (79) is achieved if

$$\lambda_k = \frac{\mu N}{\sigma_a^2}, \quad k = 1, \dots, m \quad (80)$$

implying that an optimum code must comply with the condition

$$\mathbf{M}^{-1} \mathbf{A} \mathbf{A}^* \mathbf{M}^{-1} = \begin{cases} \frac{\mu N}{\sigma_a^2} [2\mathbf{A}(\mathbf{A}^* \mathbf{M}^{-1} \mathbf{A})^{-1} \mathbf{A}^* - \mathbf{A} \mathbf{A}^*] & \text{Case 1} \\ \frac{\mu N}{\sigma_a^2} \mathbf{I} & \text{Case 2.} \end{cases} \quad (81)$$

In particular, if the additive disturbance is white, i.e., $\mathbf{M} = \sigma_n^2 \mathbf{I}$, the above equation reduces to

$$\mathbf{A} \mathbf{A}^* = \begin{cases} \frac{4\sigma_n^4 \mu N}{\sigma_a^2} (\mathbf{A}^* \mathbf{M}^{-1} \mathbf{A})^{-1} & \text{Case 1} \\ \frac{4\sigma_n^4 \mu N}{\sigma_a^2} \mathbf{I} & \text{Case 2.} \end{cases} \quad (82)$$

The last equation subsumes, as a relevant case, the set of orthogonal space-time codes. Indeed, assuming $N = n = m$, the condition (82) yields, for the optimum code matrix,

$$\mathbf{A} \mathbf{A}^* = \frac{4\sigma_n^4 \mu N}{\sigma_a^2} \mathbf{I}, \quad (83)$$

i.e., the code matrix \mathbf{A} should be proportional to any unitary $N \times N$ matrix. Thus, any orthonormal basis of F^N can be exploited to construct an optimum code under the Case 2 and

white Gaussian noise. If, instead, we restrict our attention to code matrices built upon Galois Fields (GF), there might be limitations to the existing number of optimal codes. Deffering to (Tarokh, 1999) and to the Urwitz-Radon condition exploited therein, we just remind here that, under the constraint of binary codes, unitary matrices exist only for limited values of N : for 2×2 coding, we find the normalized Alamouti code (Alamouti, 1998), which is an orthonormal basis, with elements in $\text{GF}(2)$, for F^2 .

Make some comments. First notice, that under the white Gaussian noise, both performance measures considered above are invariant under unitary transformations of the code matrix, while at the correlated clutter they are invariant with respect to right multiplication of \mathbf{A} by a unitary matrix. Probably, these degrees of freedom might be exploited for further optimization in different radar functions. Moreover, (70) represents the optimum solution for the case that no constraint is forced upon the code alphabet; indeed, the code matrices turn out in general to be built upon the completely complex field. If, instead, the code alphabet is constrained to be finite, then the optimum solution (70) may be no longer achievable for arbitrary clutter covariance. In fact, while for the special case of white clutter and binary alphabet the results of (Tarokh, 1999) may be directly applied for given values of m and n , for arbitrary clutter covariance and (or) transmit/receive antennas number, a code matrix constructed on $\text{GF}(q)$ and fulfilling the conditions (70) is no longer ensured to exist. In these situations, which however form the object of current investigations, a brute-force approach could consist of selecting the optimum code through an exhaustive search aimed at solving (66), which would obviously entail a computational burden $O(q^{mN})$ floating point operations. Herein we use the usual Landau notation. $O(n)$; hence, an algorithm is $O(n)$ if its implementation requires a number of floating point operations proportional to n (Golub & Van Loan, 1996). Fortunately, the exhaustive search has to be performed off line. The drawback is that the code matrix would inevitably depend on the target fluctuation law; moreover, if one would account for possible nonstationarities of the received clutter, a computationally acceptable code updating procedure should be envisaged so, as to optimally track the channel and clutter variations.

4.5 Simulation

The present section is aimed at illustrating the validity of the proposed encoding and detection schemes under diverse scenarios. In particular, we first assume uncorrelated disturbance, whereby orthogonal space-time codes are optimal. In this scenario, simulations have been run, and the results have been compared to the Chernoff bounds of the conventional GLRT receiver discussed in (De Maio & Lops, 2007) and to the GD performance achievable through a single-input single-output (SISO) radar system. Next, the effect of the disturbance correlation is considered, and the impact of an optimal code choice is studied under different values of transmit/receiver antenna numbers. In all cases, the behavior of the mutual information between the observations and the target replicas can be also represented, showing that such a measure is itself a useful tool for system design and assessment, but this analysis is outside of a scope of the present chapter.

Figure 8 represents the white Gaussian disturbance and assesses the performance of the GLRT GD. To elicit the advantage of waveform optimization, we consider both the optimum

coded wireless communication system and the uncoded one, corresponding to pulses with equal amplitudes and phases. The probability of detection P_D is plotted versus SCR assuming $P_{FA} = 10^{-4}$ and $N = m = n = 2$. This simulation setup implies that the Alamouti code is optimum in the sense specified by (82). For comparison purposes, we also plot the performance of the uncoded SISO GD. We presented the performance of the conventional GLRT to underline a superiority of GD employment.

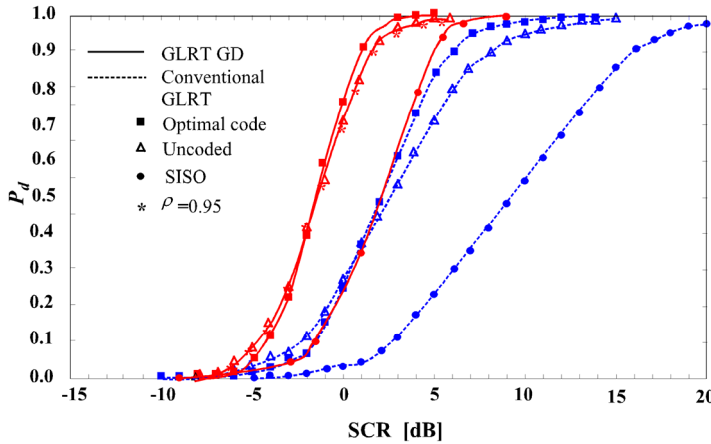


Fig. 8. P_D versus SCR; white Gaussian disturbance and disturbance with exponentially shaped covariance matrix ($\rho = 0.95$); $P_{FA} = 10^{-4}$; $N = m = n = 2$.

The curves highlight that the optimum coded wireless communication system employing the GD and exploiting the Alamouti code, achieves a significant performance gain with respect to both the uncoded and the SISO radar systems. Precisely, for $P_D = 0.9$, the performance gain that can be read as the horizontal displacement of the curves corresponding to the analyzed wireless communication systems, is about 1 dB with reference to the uncoded GLRT GD wireless communication system and 5 dB with respect to the SISO GD. Superiority of employment GD with respect to the conventional GLRT wireless communication systems achieves 6 dB for the optimum coded wireless communication system, 8 dB for the uncoded wireless communication systems, and 12 dB for SISO wireless communication systems. It is worth pointing out that the uncoded wireless communication system performs slightly better the coded one for low detection probabilities. This is a general trend in detection theory, which predicts that less and less constrained fluctuations are detrimental in the high SCR region, while being beneficial in the low SCR region. On the other hand, the code optimization results in a more constrained fluctuation, which, for low SCRs, leads to slight performance degradation as compared with uncoded systems. The effect of disturbance correlation is elicited in Fig.8 too, where the analysis is produced assuming an overall disturbance with exponentially shaped covariance matrix, whose one-lag correlation coefficient ρ is set to 0.95. In this case, the Alamouti code is no longer optimum. The plots show that the performance gain of the optimum coded GLRT GD wireless communication system over both the uncoded and the SISO GD detector is almost equal to that resulting when the disturbance is

white. On the other hand, setting $N = m = n = 2$ in (81), shows that, under correlated disturbance, the optimum code matrix is proportional to \mathbf{M} : namely, an optimal code tends to restore the “white disturbance condition.” This also explains why the conventional Alamouti code follows rather closely the performance of the uncoded GLRT GD wireless communication system.

The effect of number n of receive antennas on the performance is analyzed in Fig.9, where P_D is plotted versus SCR for $N = m = 8$, exponentially shaped clutter covariance matrix with $\rho = 0.95$, and several values of n . The curves highlight that the higher n , namely the higher the diversity order, the better the performance. Specifically, the performance gap between the case $n = 8$ and the case of a MISO GLRT GD radar system (i.e., $n = 1$) is about 2.5 dB, while, in the case of the conventional GLRT radar systems, is about 7 dB for $P_D = 0.9$. A great superiority between the radar systems employing GLRT GD and conventional GLRT is evident and estimated at the level of 6 dB at $n = 8$ and 10 dB in the case of a MISO (i.e., $n = 1$) for $P_D = 0.9$. Notice that this performance trend is also in accordance with the expression of the mutual information that exhibits a linear, monotonically increasing, dependence on n . The same qualitative, but not quantitative, performance can be presented under study of the number m of available transmit antennas on the GLRT GD wireless communication system performance.

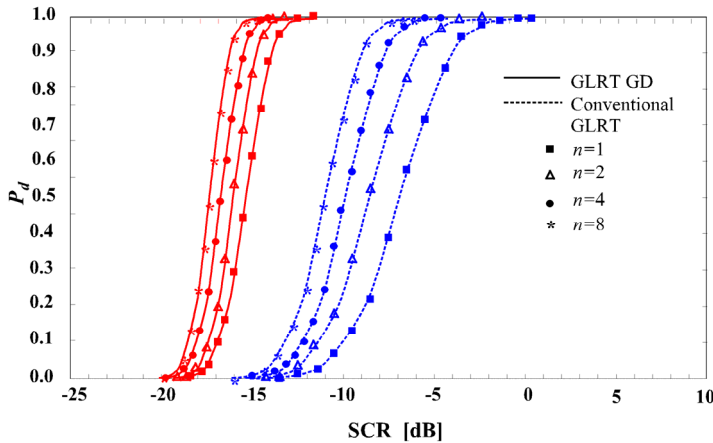


Fig. 9. P_D of optimum coded system versus SCR; disturbance with exponentially shaped covariance matrix ($\rho = 0.95$); and several values of m ; $P_{FA} = 10^{-4}$; $N = m = 8$.

4.6 Discussion

We have addressed the synthesis and the analysis of MIMO radar systems employing the GD and exploiting space-time coding. To this end, after a short description of the MIMO radar signal model applied to wireless communications, we have devised the GLRT GD under the assumption of the additive white Gaussian disturbance. Remarkably, the decision statistic is ancillary and, consequently, CFAR property is ensured, namely, the detection thresh-

old can be set independent of the disturbance spectral properties. We have also assessed the performance of the GLRT GD providing closed-form expressions for both P_D and P_{FA} . Lacking a manageable expression for P_D under arbitrary target fluctuation models, we restricted our attention to the case of Rayleigh distributed amplitude fluctuation. The performance assessment that has been undertaken under several instances of number of receive and transmit antennas, and of clutter covariance, has confirmed that MIMO GD radar systems with a suitable space-time coding achieve significant performance gains over SIMO, MISO, SISO, or conventional SISO radar systems employing the conventional GLRT detector. Also, these MIMO GD radar systems outperform the listed above systems employing the conventional GD. Future research might concern the extension of the proposed framework to the case of an unknown clutter covariance matrix, in order to come up with a fully adaptive detection system. Moreover, another degree of freedom, represented by the shapes of the transmitted pulses could be exploited to further optimize the performance. More generally, the impact of space-time coding in MIMO CD radar systems to estimate the target parameters is undoubtedly a topic of primary concern. Finally, the design of GD and space-time coding strategies might be of interest under the very common situation of non-Gaussian radar clutter.

5. Acknowledgment

This research was supported by Kyungpook National University research Grant, 2011.

6. References

- Barnard, T. & Weiner, D. (1996). Non-Gaussian Clutter Modelling with Generalized Spherically Invariant Random Vectors. *IEEE Transactions on Signal Processing*, Vol. SP-44, No. 10, pp. 2384-2390, ISSN 1053-587X.
- Bello, P. (1963), Characterization of Randomly Time-Invariant Linear Channels. *IEEE Transactions on Communications Systems*, Vol. CS-11, No. 12, pp. 360-393, ISSN 0096-1965.
- Benedetto, S. & Biglieri, E. (1999). *Principle of Digital Transmission with Wireless Applications*, Plenum Press, ISBN 0-3064-5753-9, New York, USA.
- Blankenship, T. & Rappaport, T. (1998). Characteristics of Impulsive Noise in the 450-MHz band in hospitals and clinics. *IEEE Transactions on Antennas and Propagation*, Vol. 46, No. 2, pp.194-203, ISSN 0018-926X.
- Buzzi, S. et al. (1999). Optimum Detection over Rayleigh-Fading, Dispersive Channels with Non-Gaussian Noise. *IEEE Transactions on Communications*, Vol. COM-35, No. 7, pp. 926-934, ISSN 0096-1965.
- Buzzi, S. et al. (1997). Signal Detection over Rayleigh-Fading Channels with Non-Gaussian Noise. In *Proc. Inst. Elect. Eng., Commun.*, Vol. 144, No. 6, pp. 381-386.
- Buzzi, S. et al. (2001). Optimum Diversity Detection over Fading Disperive Channels with Non-Gaussian Noise. *IEEE Transactions on Communications*, Vol. COM-49, No. 4, pp. 767-775, ISSN 0096-1965.
- Gini, F. et al. (1998). The Modified Cramer-Rao Bound in Vector Parameter Estimation. *IEEE Transactions on Communications*, Vol. COM-46, No. 1, pp. 52-60, ISSN 0096-1965.
- Conte, E. et al. (1995). Canonical Detection in Spherically Invariant Noise. *IEEE Transactions on Communications*, Vol. COM-43, No. 2-4, pp. 347-353, ISSN 0096-1965.

- Conte, E. et al. (1995). Optimum Detection of Fading Signals in Impulsive Noise. *IEEE Transactions on Communications*, Vol. COM-43, No. 2-4, pp. 869-876, ISSN 0096-1965.
- Cover, T. & Thomas, J. (1991). *Elements of Information Theory*, Wiley & Sons, Inc., ISBN 0-4710-6259-6, New York, USA.
- De Maio, A. & Lops, M. (2007). Design Principles of MIMO Radar Detectors. *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-43, No. 3, pp. 886-898, ISSN 0018-9251.
- Fishler, E. et al. (2006). Spatial Diversity in Radars – Models and Detection Performance. *IEEE Transactions on Signal Processing*, Vol. SP-54, No. 3, pp. 823-838, ISSN 1053-587X.
- Foschini, G. (1996). Layered Space-Time Architecture for Wireless Communication in a Fading Environment Using Multi-Element Antennas. *BLTJ*, Vol. 1, No. 2, pp. 41-59, ISSN 1538-7305.
- Golub, G. & Van Loan, C. (1996). *Matrix Computations*, 3rd Ed. The John Hopkins Press, ISBN 0-8018-5414-8, Baltimore, MD, USA.
- Kassam, S. (1988), *Signal Detection in Non-Gaussian Noise*, Springer-Verlag, ISBN 0-3879-6680-3, New York, USA.
- Kassam, S. & Poor, H. (1985). Robust Technique for Signal Processing: A Survey. *Proceedings IEEE*, Vol. 73, No. pp. 433-481, ISSN 0018-9219.
- Kuruoglu, E. et al. (1998). Near Optimal Detection of Signals in Impulsive Noise modelled with a Symmetric α -Stable Distribution. *IEEE Communications Letters*, Vol. 2, No. 10, pp. 282-284.
- Lombardo, P et al. (1999). MRC Performance for Binary Signals in Nakagami Fading with General Branch Correlation. *IEEE Transactions on Communications*, Vol. COM-47, No. 1, pp. 44-52, ISSN 0096-1965.
- Matthews, J. (1992). Eigenvalues and Troposcatter Multipath Analysis, *IEEE Journal of Selected Areas in Communications*, Vol. 10, No. 4, pp. 497-505, ISSN 0733-8716.
- Middleton, D. (1999). New Physical-Statistical Methods and Models for Clutter and Reverberation: The KA-Distribution and Related Probability Structures. *IEEE Journal of Oceanic Engineering*, Vol. 24, No. 7, pp. 261-284, ISSN 0364-9059.
- Poor, H. (1988). *An Introduction to Signal Detection and Estimation*. Springer-Verlag, ISBN 0-3879-4173-8, New York, USA.
- Proakis, J. (2007), *Digital Communications*, 5th Ed. McGraw-Hill, ISBN 0-0729-5716-6, New York, USA.
- Sangston, K. & Gerlach, K. (1994). Coherent Detection of Radar Targets in a Non-Gaussian Background. *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-30, No. 4, pp. 330-334, ISSN 0018-9251.
- Skolnik, M. (2008). *Radar Handbook*, 3rd Ed. McGraw-Hill, ISBN 978-0-07-148547-0, New York, USA
- Sousa, E. (1990). Interference Modelling in a Direct-Sequence Spread-Spectrum Packet Radio Network. *IEEE Transactions on Communications*, Vol. COM-38, No. 9, pp. 1475-1482, ISSN 0096-1965.
- Tuzlukov, V. (1998). A New Approach to Signal Detection Theory. *Digital Signal Processing*, Vol. 8, No. 3, pp. 166-184, ISSN 1051-2204
- Tuzlukov, V. (1998), *Signal Processing in Noise: A New Methodology*, IEC, ISBN 985-6453-16-X, Minsk, Belarus

- Tuzlukov, V. (2001), *Signal Detection Theory*, Springer-Verlag, ISBN 0-8176-4152-1, New York, USA
- Tuzlukov, V. (2002), *Signal Processing Noise*, CRC Press, Taylor & Francis Group, ISBN 0-8493-1025-3, Boca Raton, USA
- Tuzlukov, V. (2005), *Signal and Image Processing in Navigational Systems*, CRC Press, Taylor & Francis Group, ISBN 0-8493-1598-0, Boca Raton, USA
- Tuzlukov, V. (2012), *Signal Processing in Radar Systems*, CRC Press, Taylor & Francis Group, ISBN 0-8493-.....-, Boca Raton, USA (in press).
- Van Trees, H. (2003), *Detection, Estimation, and Modulation Theory*. 2nd Ed. Wiley & Sons, Inc., ISBN 978-0-471-44967-6, New York, USA.
- Webster, R. (1993), Ambient Noise Statistics. *IEEE Transactions on Signal Processing*, Vol. SP-41, No. 6, pp. 2249-2253, ISNN 1053-587X.
- Yao, K. (1973). A Representation Theorem and Its Applications to Spherically Invariant Random Processes. *IEEE Transactions on Information Theory*, Vol. IT-19, No. 7-8, pp. 151-155, ISNN 0018-9448.

Engineering of Communication Systems and Protocols

Pero Latkoski and Borislav Popovski
*Faculty of Electrical Engineering and
Information Technologies / Ss Cyril and Methodius University – Skopje
Macedonia*

1. Introduction

The complexity of the communication systems and protocols is increasing constantly, while the communication products' time-to-market is becoming shorter. Afterthoughts communication system redesign due to lack of performance is financially and time expensive, and it is unacceptable. This book chapter proposes a method for improving the telecommunication systems, by means of enhancement the performance of the protocols they rely on. The proposed engineering of communication systems is based on a formal method and it provides an early-phase performance evaluation of the underlying communication protocols. The methodology is illustrated through a hands-on case study conducted on an existing wireless communication system.

The development and standardization of new telecommunication and information technologies is a rather complex process which requires a comprehensive framework (Sherif, 2001). The result of such a process is a new agreement that should satisfy all of the involved parties, such as: vendors, providers, and most importantly service users. To create a comprehensive communication standard and consequently a reliable communication system, many strategic and tactical issues need to be considered. The missing question is how to produce a standard that specifies a protocol or a system with high performance. The lack of performance issue might be a major cause for pitfall of entire communication systems. Most of the problems result from poor protocol specifications or from its enormous complexity. Furthermore, the design errors caused by the short and intensive creation period usually remain hidden until the testing and implementation phases of the communication product development. Fixing the problems after product's delivery for communication software and hardware increases the cost of the product by factor of 100 to 1000 compared to the fixing of the problem in the analysis phase.

High performance communication protocols which are untainted of functional errors are crucial in the telecommunications sector where product expectation cycle is denominated in decades instead of years. In order to develop such a protocol, two aspects should be fulfilled: introduction of formal methods during the specification process and integration of the performance-related activities in the early phases of the communication system specification and development. The former one is already taking place as a result of the need

for clarity and accuracy in the telecommunication standards, but the last aspect is commonly avoided or even neglected.

The formal methods are always advised for the development process when early functional error detection is needed. Formal Descriptive Techniques (FDTs) provide corrective actions in the more abstract phases by introducing formal syntax and what is more important, precise semantics. In combination with the computer-aided software engineering, FDTs offer a delivery of better communication protocols and systems, sooner. The introduction of the FDTs has brought correctness and reliability into the protocol development, which has been recognized long time ago (Wing, 1990), (Hall, 1990). Today there are many formal languages and tools used in the protocol development process: Specification and Description Language-SDL (SDL, 2011), Simple ProMeLa Interpreter (Spin), Estelle (Estelle, 1989), Language of Temporal Ordering Specifications (LOTOS, 2000), Petri Nets (Petri, 1996), Uppaal (Larsen, 1997), Message Sequence Chart (MCS, 2001) and Unified Modelling Language (Booch, 2000). Among them, SDL has achieved widespread success because of its friendly graphical notation, its standardization by the International Telecommunication Union (ITU-T) as the major specification tool for standards and protocols, and because of its support for other popular notations such as ASN.1 (ASN.1, 1993), MSC and TTCN (TTCN, 2006). The effectiveness of SDL and its ability to develop unambiguous protocols have won it a widespread popularity and have led the standardization institutes, such as ETSI (European Telecommunications Standards Institute) (ETSI), 3GPP (Third Generation Partnership Project) (3GPP) and IEEE (The Institute of Electrical and Electronics Engineers) (IEEE) to include SDL diagrams in their protocols specification. SDL also provides powerful analysis of communication protocols, along with design, comprehensive modelling, protocol prototyping, exhaustive validation and verification, and all that by a user-friendly graphical notation. Along with Message Sequence Chart (MSC) description language, SDL is the most widely used FDT not only in the communication protocol specification area, but also in the industry systems engineering domain. Because of the previously stated advantages, SDL was selected as a protocol description method for the purpose of this chapter's analysis.

The aim of this chapter is to emphasize the importance of conducting an early performance evaluation of the communication protocols and systems, and to suggest an appropriate solution for carrying out such an activity. Performance evaluation activity denotes the actions to evaluate the protocol under development regarding its performance. This process can take place in different phases of the development, and can be based on modelling or measurements. If the designer can control the performance of the product, rather than just manage its functionality, the result will be a much superior creation. This problem is treated in this chapter through a tangible wireless communication protocol example.

The chapter is organized as follows. Section 2 presents the most relevant and most recent work which relates to the target topic of the chapter. In Section 3, the proposed and used methodology is elaborated in details. This methodology is demonstrated in Section 4, where a real engineering problem is provided, involving an IEEE 802.16 wireless communication protocol. Section 5 contains the conclusions of the chapter.

2. Related work

The following section provides an overview of what has been done by other researchers, related to the chapter's topic. Only a part of the most relevant and most recent work has

been selected, which is needed for proper introduction of the proposed methodology in Section 3 and for presentation of the example in Section 4.

Engineering of a communication system means to describe, to analyze and to optimize the dynamic, time dependent behavior of the system and its inherent communication protocols. However, as it says in (Mitschele-Thiel, 2001) it is common for a system to be fully designed and functionally tested before an attempt is made to determine its performance characteristics. But it is a necessity to integrate the performance engineering into the design process from the very beginning. In (Mitschele-Thiel, 2001) the author addresses an improvement of the run-time properties by taking into account the characteristics of the applications (communication protocols) and different process scheduling and management strategies. The author concentrates on efficient implementation of behavioural concepts. For the treatment of issues arising from object-oriented concepts the author applies the traditional flattening approach of the language standard. Finally, it is obvious that the book lacks of actual communication system engineering examples, through which the engineering process would have been successfully explained.

The usage of FDT for protocol development has also arisen as a promising way of dealing with the increasing complexity of next generation mobile protocols. In (Showk, 2009) a rudimentary version of the Long Term Evolution (LTE) protocol for the access stratum user plane is modelled using SDL. The LTE radio communication is the upgrade of the current 3G mobile technology with a more complex protocol in order to enable very high data rates. This related work presents a tool which shows easy understanding of the model as well as easy testing of its functionality by simulation in cooperation with Message Sequence Chart. The simulation result presented in (Showk, 2009) shows that the implemented SDL guarantees a good consistency with the target scenarios. The system implementation is mapped to multiple threads and integrated with the operating system to enable execution in multi core hardware platforms. The only obvious drawback of the work is the usability of the created model, as it is only used for functional validation and not for performance evaluation of the analyzed communication protocol.

When developing modern communication systems, the energy consumption is a major concern, especially in the case of wireless networks consisting of battery-powered nodes. In (Gotzhein, 2009) the authors study possibilities of specifying energy aspects in the system a designing phase, with SDL as design language. In particular, they strive for suitable abstractions, by establishing a design view that is largely platform-independent. This objective is achieved by identifying and realizing energy mode signalling and energy scheduling as two complementary approaches to incorporate energy aspects into SDL. A case study illustrates the use of both approaches in a wireless networked control system. These approaches are applied and tested on a hardware platform, but again, the paper does not provide in a sufficient manner any performance metrics of the implemented wireless network.

The security of communication systems is another important aspect which must be considered in the protocol development. In order to study this aspect, (Lopez 2005) have developed a methodology for application of the formal analysis techniques, commonly used in communication protocols, to the analysis of cryptographic protocols. In particular, (Lopez, 2005) have extended the design and analysis phases with security properties. This

related work uses a specification notation based on one of the most commonly used standard requirement languages HMSC/MS, which can be automatically translated into a generic SDL specification. The obtained SDL system can then be used for the analysis of the addressed security properties, by using an observer process scheme. Besides the main goal to provide a notation for describing the formal specification of security systems, (Lopez, 2005) studies the possible attacks to the system, and the possibility of re-using the produced specifications to describe and analyse more complex systems.

The related work (Chen Hui, 2010) analyzes a Networked Control System (NCS), which governs the communication activities and directly affects the communication Quality of Service (QoS). Full or partial reconfiguration of protocol stack offers both optimized communication service and system performance. (Chen Hui, 2010) proposes a formal approach for the design and implementation of reconfiguration protocol stack based on Specification and Description Language for NCS. In Telelogic TAU environment, detail SDL models to support communication and reconfiguration functions of communication link layer, network transmission layer and application layer are discussed respectively. Similarly to the most of the presented related papers, only MS verification results validate the effectiveness of the reconfiguration concepts of the protocol implementation for NCS.

The methodology which is presented in the following section differs from all previously presented work, as it extends the performance evaluating aspect of the communication systems engineering process. The methodology tries to maintain the functional correctness efforts regarding developing communication entity (similarly to the most of the related work), but at the same time provides the developers with a realistic insight of its performance capabilities.

3. Methodology overview

In order to obtain a performance evaluation-based analysis of telecommunications protocols and systems, the proposed methodology extracts all the necessary information from the available form of the analyzed standard. Taking into consideration that we are talking about an early stage of communication system development, the standard for such particular system is usually available as a draft version which combines the work provided by different working groups. The aim is to build an appropriate model from which the performance of the communication system will be assessed. This kind of model is commonly referred to as performance evaluating model. The communication standard under evaluation generally contains three basic parts: textual, SDL-represented and MS-represented. Depending on the standard and standardization body, the proportion of these parts can vary. Mainly, the textual part dominates as “in the standardization process, words are still the final product” (Sherif, 1992).

But there is a major setback of the textual part of the standards caused by its inherent ambiguousness which is more deeply related to the natural languages’ doublethink. This is the reason why the textual part of standards lacks of scientific foundation and is commonly the reason for miss-communication between standard developers, regarding the communication system requirements and expectations.

On the other hand, the SDL and MSC parts of the standard introduce more rigorous protocol or system specification, brought to a mathematical precision, and these formal parts of the standard are the guarantee for a correct system requirements presentation, as well as for an unambiguous definition of the system behavior. As it was previously said, different standards contain variable amount of formal representation, e.g. IEEE 802.16 (WiMAX, 2010) contains only sequences of formal protocol behavior, IEEE 802.11 (WiFi, 2007) encloses entire Medium Access Control (MAC) Layer presented in SDL, and IEEE 802.15 (Bluetooth, 2005) is completely offered in formal representation.

Using the three basic types of standard representations, the proposed methodology creates a so called behavior model of the analyzed standard, as it is presented in Fig. 1. The behavior model describes the protocol behavior and its abstract data structure by using SDL. In particular, the behavior model evaluates the relationship of single stimuli - response pair applied to the analysed protocol stack. This model is ideal for testing of protocol entity's functional correctness.

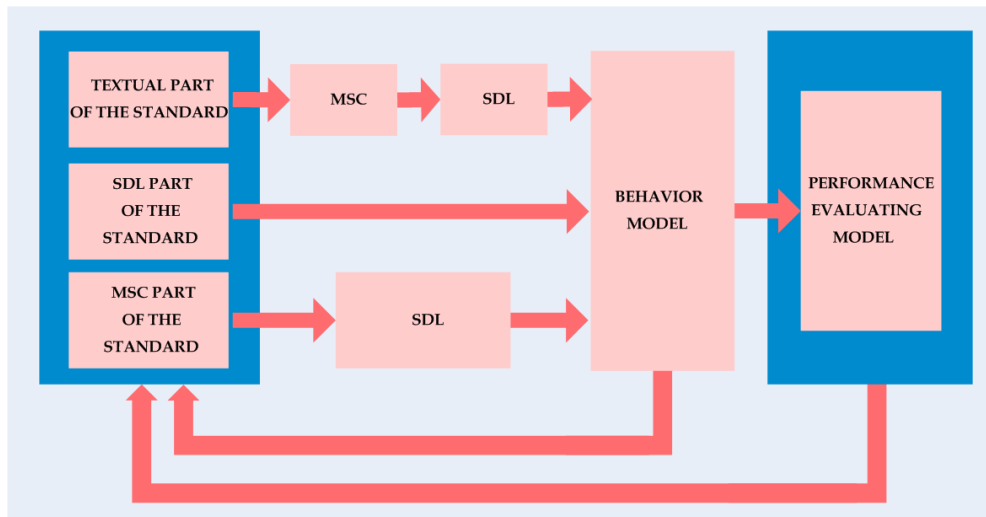


Fig. 1. Transformation of a communication standard into a performance model.

As it can be seen in Fig. 1, the last step of the methodology is the conversion of the behavior model into a so called performance evaluating model. The performance model can emulate a real scenario of communicating devices implementing the described protocol. It is built into a standalone executable that embeds the created behavior model and channelizes its preciseness into an accurate event driven type of simulator. This type of simulator is used for performance evaluation of the communication system or protocol, where every change and tuning of the specification can be easily evaluated. For instance one can measure the achieved data throughput or delay when a group of protocol based devices are communicating between each other.

Both, behavior and performance models provide valuable information regarding the functional and performance issues of the developing communication system, which is then looped-back to the specification process for further improvement of the system.

3.1 Detailed steps in the methodology

The behavior model can be defined as a SDL representation of the requirements, behavior and capabilities of the specified system or protocol. As it was explained earlier, it is created using all three representation parts of standards (text, SDL, and MSC). The SDL part of a standard is easily incorporated in the behavior model. This is the reason why the SDL part of the standard formulates the backbone of this model. Obviously it is the most mathematically rigorous component of the behaviour model. The MSC part of the standard includes all signal exchange sequences occurring among the protocol entities defined by the standard. To incorporate this part into the behavior model, it is necessary first to convert the MSCs into SDL code sequences. Although automated tools for such a conversion exist, the manual step-by-step translation is preferred, as the SDL code produced by the automated tools is not optimized, and also for providing nomenclature and style consistency of the SDL code. The trickiest part of the behavior model development is to convert all the informal textual representation of the standard into SDL code. This is an unavoidable step, as long as all the missing parts of a complete functional system description are given in a textual form. The SDL code sequences produced from the textual part of the standard act as a glue that connects all the previously created parts of the behavior model. The need to convert text into SDL is also unavoidable because of the fact that for most of the communication standards the SDL and MSC parts are supplementary, while the textual part is mandatory. These are the reason why this step should be taken as the one with greatest importance. Additionally, the communicating signals which are exchanged among entities, along with the signal parameters, are most of the times constructed according to the text of the standard. The practice have shown that it is much easier if the textual represented requirements of the standard are firstly translated into MSC sequences, and after that converted into SDL code. This principle proves to be especially suitable for capturing the complex communication system or protocol behavior.

For the completeness of the previous explanation, this section will present an example of a generic behavior model of an abstract communication standard (Fig. 2). As it can be seen in Figure 2, SDL copes with the protocol complexity by using a hierarchical decomposition and by implying several levels of abstraction. The highest level of abstraction is called the system level. The system level is composed of multiple SDL blocks connected through unidirectional or bidirectional SDL channels. SDL channels are transferring SDL signals, which can carry additional signal parameters. Inside the SDL blocks lays the second level of abstraction represented by groups of processes located in the blocks. The processes use signal routes for transferring the signals among them or to the higher-level channels. Inside the SDL processes, Extended Communicating Finite State Machines (ECFSMs) are used for description of each protocol entity behavior. This is the lowest and the most detailed level of the behavior model. The functionalities of the protocol entity are presented unambiguously by using SDL discrete states, triggers, transitions, tasks, procedures, decisions, manipulation of variables, management of

signals, etc. In the same manner, all protocol primitives are described, along with the signal's parameters exchanged among the protocol entities.

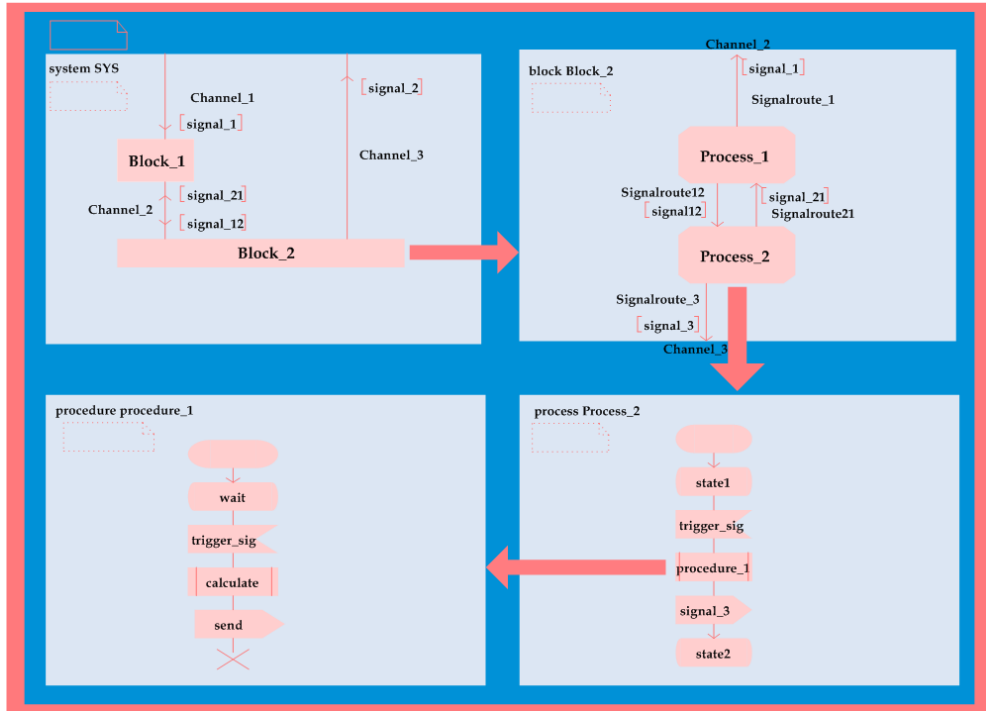


Fig. 2. Insight of a generic behavior model.

In order to assess the real performance of the analyzed system or protocol, it is necessary to build a performance evaluating model. The SDL performance model can emulate real working scenarios of communicating devices. It is a standalone model that embeds the behavior model and translates its preciseness into an accurate event driven type of simulator. The results obtained by the performance model are reliable indicator of the expected performance of the future communication device, which will be built according to the analysed standard.

The process of 'assimilation' and upgrade of a generic behavior model into a generic performance model is depicted in Fig. 3. The presented generic behavior model contains several protocol layers, represented by SDL processes. The following layers are included: Convergence layer, MAC layer, PHY layer, and a vertical Management layer. The behavior model describes all possible interactions between these entities in a single stimuli-response manner.

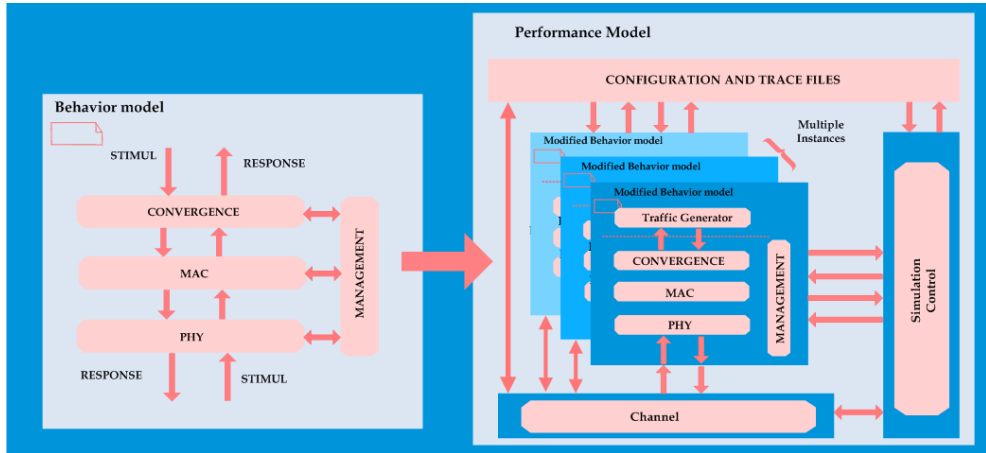


Fig. 3. Building a generic performance model.

Unlike the behavior model where only single stimuli - response pair of a protocol entity is evaluated, the performance model introduces new entities which are needed for completing the communication scenario emulation, both on system and block level. On the system level besides multiple instances of the modified behavior model, it is crucial to introduce a simulation control block and a channel block. The control block manages the generation of block-instances and controls the simulation time, while the channel block ensures a proper emulation of exchanging packets (e.g. radio frequency packets) among the communicating entities. Each block instance is characterized by a unique process identifier (PID), which enables differentiation and addressing of the identical entities.

As it was previously stated, in the foundation of the performance model lays the behavior model. It is necessarily modified (i.e. upgraded), in order to conduct its expected role of real communicating device emulation. Many new processes with an appropriate purpose should be introduced in this upgraded version of the behaviour model: controller of the primitive exchange process, procedures for queuing and prioritization of the signals, processes for segmentation and reassemble of the messages, manipulators of simulation time (timers and clocks), etc. All these processes are introduced according to the textual part of the protocol specification.

After building the performance model using the SDL Graphical Representation (SDL-GR), abstract Data Types (ADTs) are added in order to introduce important functionalities (e.g. reading and writing to file, different kinds of random number generators, etc.).

The analyzer then runs the performance model for detecting all the ambiguities. Next step is the conversion of the built model into a so called Phrasal Representation of SDL (SDL-PR). Using SDL-PR, the code generator produces C++ source code, compiles it and links it with appropriate libraries. The result of these steps is a standalone simulator executable which requires as an input only the configuration files, needed for the description of the desired network scenario.

4. Case study

This section contains an implementation case study of the previously proposed methodology. In particular, the case study extends the findings proposed by (Latkoski, 2010), and provides the needed validation of the analytical and numerical analysis contained in (Latkoski, 2010).

The targeted communication system is based on WiMAX technology, standardized by IEEE 802.16. It belongs to the group of wireless metropolitan area networks (WMANs), which are on the steady track of widespread deployment in many urban environments. This worldwide trend is facilitated by the ever growing demand for “last-kilometre” network connectivity in every part of those urban environments with guaranteed service quality. A significant portion of the WMAN installations are based on the IEEE 802.16 technology, which is mature enough for seamless and low-cost deployment.

The focus of the analysis provided here, is the protocol which is responsible for bandwidth allocation among the WiMAX users. The WiMAX channel access is controlled by one of the several available Medium Access Control (MAC) procedures. (Latkoski, 2010) studies the contention-based bandwidth request procedure based on original analytical model, facilitated by numerical analysis. In (Latkoski, 2010) the key parameters of the contention procedure are optimized in order to minimize the average bandwidth access delay, thus ensuring the highest possible quality of service (QoS) to the WiMAX users.

WiMAX supports several QoS classes: UGS - Unsolicited Grant Service (E1/T1), real-time Polling Service - rtPS (MPEG), non-real-time Polling Service - nrtPS (FTP) and Best Effort - BE (HTTP). Except for the UGS that uses dedicated uplink transmission slots, the remaining service classes use the bandwidth request procedures over the uplink to the base station (BS). Depending on the service class, the access scheme can be either contention-based or based on unicast polling. The vendor-specific implementation can offer two optional non-mandatory procedures: piggybacking and bandwidth stealing procedures. Here, we focus on the IEEE 802.16 contention-based bandwidth request access scheme, which supports the BE class of traffic, generated by most Internet applications (web surfing, FTP, etc.). Additionally we will compare this scheme with the round-robin polling scheme, as well as with the multicast-groping-based principal of bandwidth management. All three access schemes are briefly explained in the following subsections.

4.1 Contention based bandwidth access

The IEEE 802.16 standard supports a mandatory Point-to-Multipoint (PMP) architecture operating in Time Division Duplex (TDD) mode. In such network conditions, the frames are divided to downlink (DL) and uplink (UL) subframes. The BS transmits uplink map (UL-

MAP) messages at the beginning of the DL subframe, in order to schedule the uplink traffic from the subscriber stations (SSs) to the BS. The beginning of UL subframe contains Information Elements (IEs) dedicated for initial ranging and bandwidth request procedures, followed by slots for the actual data transmission. The MAC layer of IEEE 802.16 specifies the rules for the contention-mode bandwidth request procedure. A contention period, as mentioned, is allocated at the beginning of the uplink subframe. It is divided into an integer number of transmission slots and is called an information element. Each transmission slot can be used for a transmission of only one bandwidth request. The SSs use the contention slots to send bandwidth request messages. If a SS's request message transmission is successful, the BS grants contention-free data transmission slot for that particular SS in one of the following frames by placing the SS's Connection ID (CID) in the UL-MAP message.

If more than one SS tries to transmit its request in the same transmission slot, a collision happens. Since it is not practically possible for SSs to sense the UL channel and to detect a collision, the SSs can only know of the success of their bandwidth request transmission if they receive a response in the form of a bandwidth grant from the BS in the subsequent frames. A subscriber station that does not receive a response to its bandwidth request by a certain deadline assumes that either a collision happened or resources are not available at the BS. In either case, since the SS can not determine the cause, it assumes that a collision happened and uses an exponential binary back-off procedure to resolve the collision. In particular, the SS starts a contention-based procedure by setting a so called initial backoff window which is an integer number. Next step is selection of a random number within the window, which determines the number of contention slot for which the SS will defer its next request message transmission. Only the slots for which the SS is eligible to send are counted. When the SS's counter reaches zero, the SS sends its request message. The SS considers the contention transmission as lost if no data grant has been given within the period of time defined by a timer. Then the SS enters in the next stage of the backoff algorithm by doubling the size of the backoff window and selecting another random number. This repeats with each loss of the request message, until the backoff window size reaches its maximum size.

4.2 Multicast-grouping based bandwidth access

The BS controls the access rights of the SSs for each contention slot. If the BS declares one contention slot as a *broadcast* type of slot, then all SSs have the right to transmit their bandwidth request messages in that particular slot. Contrary to this principle, the BS can mark certain slot as a *multicast* type of slot. In this case only the members of the specified multicast group can access the slot. The BS controls the membership of each SS into multicast groups. For this purpose, it uses a special MAC message called MCA-REQ (Multicast Polling Assignment Request). Each MCA-REQ message contains three basic parameters: the basic CID of the SS, the index of the multicast group, and one of the two possible commands, *join* or *leave*. One SS can belong to several multicast groups.

In order to evaluate the influence of multicast-group implementation over the system's performance, we will calculate the ratio of the number of successful bandwidth request transmissions per frame and the number of active SSs ($maxN_{suc}/n$). The value of this ratio $maxN_{suc}/n = 1$ means that all active SSs are served in one TDD frame. The following figure (Fig. 4) presents the results regarding $maxN_{suc}/n$ for different number of active SSs (n) and

different number of transmission slots per frame (N_r). The results are obtained by using the analytical equations provided by (Latkoski, 2010).

Furthermore, Fig. 4 provides additional insight regarding the maximization of the bandwidth procedure success rate. It is obvious that instead of using all N_r slots for the contention of all n subscriber stations, it is better to split the SSs into M groups, and to give each group a portion of N_r/M slots for contention. The colored lines in Fig. 4, give the possibilities for implementation of this idea. For example, instead of using $N_r = 16$ slots for $n = 16$ users, it is more efficient to use $M = 8$ multicast groups, as the N_{suc}/n is higher for $(n, N_r) = (2, 2)$ compared to the case where $(n, N_r) = (16, 16)$.

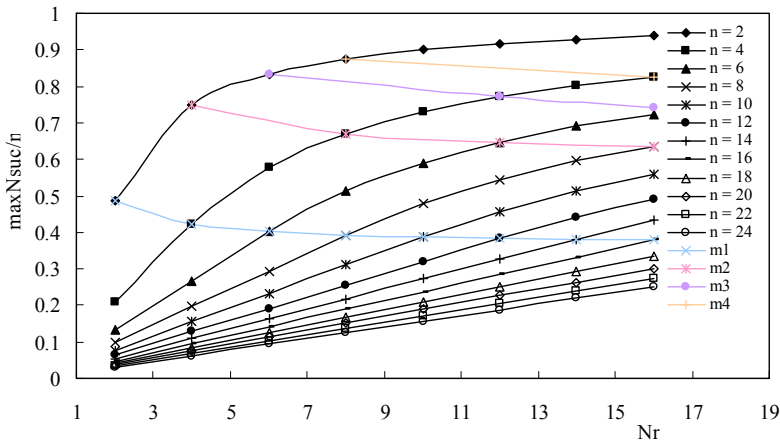


Fig. 4. Normalized success rate for conation-based scheme.

The obvious challenge here is to obtain a precise estimation of the number of active SSs (n) by the serving BS, which controls the contention and multicasting.

4.3 Round robin polling bandwidth access

Instead of using contention based bandwidth distribution among the SSs, the BS has an option to use the round-robin polling based procedure for bandwidth access. In this case, the BS asks each of the registered SSs whether they need bandwidth, starting with the first SS and ending with the last N_{all} SS. Then the circle of polling repeats again from the first SS. Considering that not all N_{all} SSs need bandwidth at a time, but only n of them have such need, we can calculate the efficiency of this method through the performance parameter defined as utilization of the slots. We can compute the utilization of the transmission slots in the case of round-robin polling (RR_{util}) as:

$$RR_{util} = \frac{n}{N_{all}}, \tag{1}$$

while the average bandwidth access delay seen by the SSs (RR_{Tid}) is:

$$RR_{Td} = \frac{N_{all}}{N_r} t_{frm}, \tag{2}$$

where the N_r represents the total number of transmitting slots, and t_{frm} is the TDD frame duration. These simple equations reveal that the utilization in the case of round-robin polling scheme does not depend on N_r , while the delay does not depend on n . Consequently, this bandwidth allocation scheme is expected to have higher performance in scenarios where the number of active SSs (SSs which need bandwidth) is close to the number of registered SSs.

The previous conclusion can be proven by making a comparison of the transmission slot utilization in the case of round-robin polling and contention based schemes for different numbers of active users. For this purpose we have used the equations provided by (Latkoski, 2010) for the maximal utilization of the transmitting slots provided by contention scheme. The results presented in the following figure are obtained for different values of N_r and N_{all} .

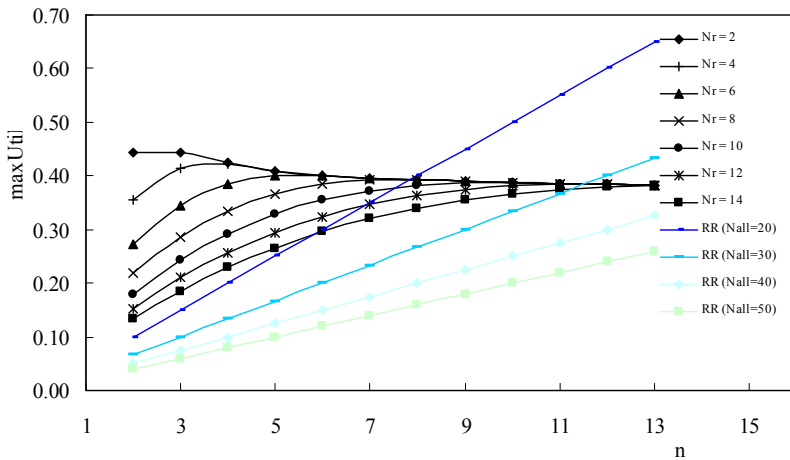


Fig. 5. Comparison of the schemes performance.

The range of values where $n \approx N_{all}$ (right part of the figure) is more appropriate for round-robin polling scheme utilization, compared to the contention-based scheme.

4.4 SDL models

For the purpose of analytical results validation, as well as for testing and improvement of the communication protocol for bandwidth allocation, we have created according to the methodology presented in the previous section, both behavior and performance evaluating models. Actually, we have built several behavior models for different MAC-layer processes involved by the communication protocol, located in both base station and subscriber station. After the functional testing of each protocol entity, the behavior models are implemented into fully operational performance model. The highest level of this model is presented in the

following Fig. 6. It contains a behavior model of the BS which contains several processes: Optimizer, Estimator and MsgCreator. The Optimizer determines the optimal values of the following contention parameters: the initial contention window, the number of allowed consequent unsuccessful bandwidth request transmissions, the number of multicast groups, and the number of contention slots per frame. These parameters are sent to the process which constructs the MAC management messages (MsgCreator), as well as to the process Estimator. The purpose of the Estimator is to estimate several network condition related parameters, such as: the number of active users, the probability of collision, the probability of transmission, etc. These parameters are needed for an accurate operation of the Optimizer.

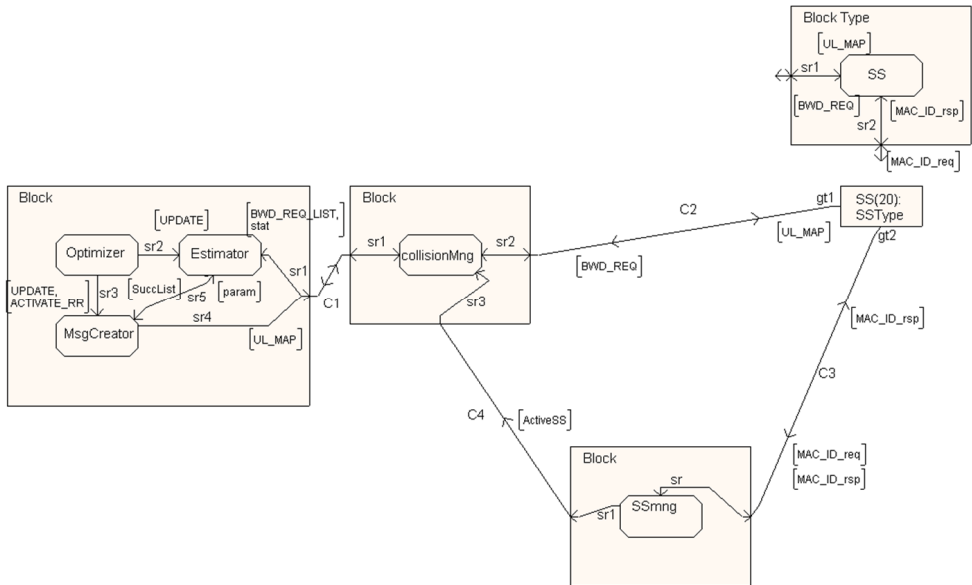


Fig. 6. Case study Performance model.

The performance model also contains blocks for channel emulation (collisionMng) and simulation control block (SSrng) which provides SSs PID management.

Besides the BS, the performance model contains multiple instances of the subscriber station block. All instances of the SS block operate as independent user equipment stations. Through the SSrng block we are able to define and control the number of active stations in the simulation scenario. The BS through the MsgCreator block controls the mode of operation (contention or round-robin, along with the number of multicast groups), according to the Optimizer commands. The Estimator operates dynamically, and feeds the Optimizer with the necessary information regarding the network conditions. The active stations are sending bandwidth request messages and then register the outcome of every attempt (success or failure).

In this communication protocol engineering case study, the Optimizer is the newly proposed entity which operation will be described in details. Actually, the Optimizer performs several steps presented formally in Fig. 7. These steps are:

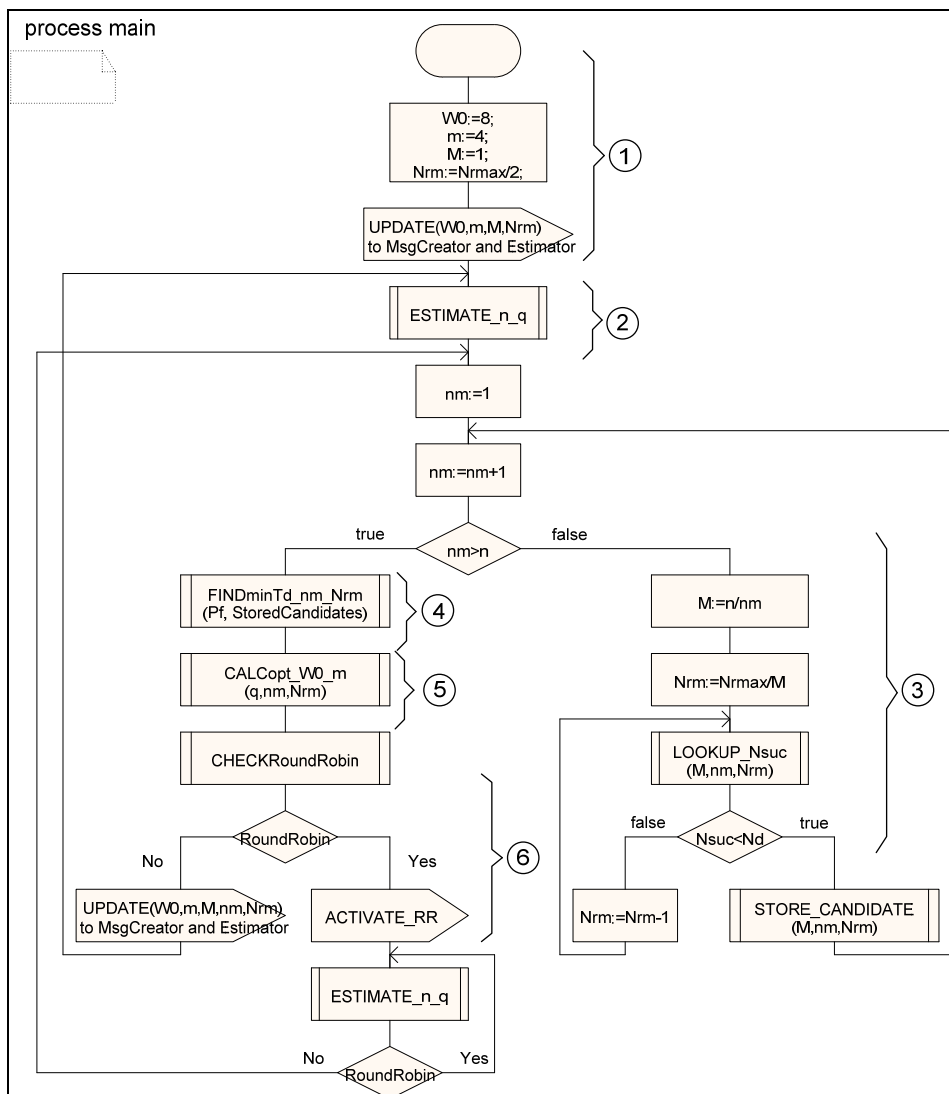


Fig. 7. Functional steps of the Optimizer.

1. The Optimizer initiates the bandwidth request procedure using predefined initial values for the contention parameters, without using multicast ($M = 1$).
2. After a training period of 5 seconds, the Estimator sends to the Optimizer estimated values for the number of active SSs (n) and information regarding their activity dynamics (g), please refer to (Latkoski, 2010).
3. The Optimizer finds the most suitable values for the number of multicast groups and number of SSs per group. For this purpose the Optimizer searches through a LOOKUP table for the possible candidate values of the contention parameters which can provide number of successful bandwidth requests per frame (N_{suc}) such as $N_{suc} < N_d$, where N_d is the number of uplink data slots per frame.
4. From all candidate parameter values, the Optimizer selects those which will provide the lowest value for the average bandwidth access delay.
5. Then the Optimizer calculates the optimal values for the contention window and the number of consecutive unsuccessful attempts.
6. Finally, the Optimizer checks whether the round-robin polling method could provide better performance. After this, it sends its final decision by a command to the MsgCreator.

4.5 Results

The performance model was tested in simulation scenario where the number of active stations (n) is changed with the time, as presented in Fig. 8. The scenario simulates 24 hour network operation. The next two figures (Fig. 9 and Fig. 10) provide the measured performance of the bandwidth request procedure for three different modes of operation: round-robin polling, contention without multicast grouping, and contention with multicast grouping. Fig. 9 presents the transmission slots utilization, while Fig. 10 presents the average bandwidth access delay.

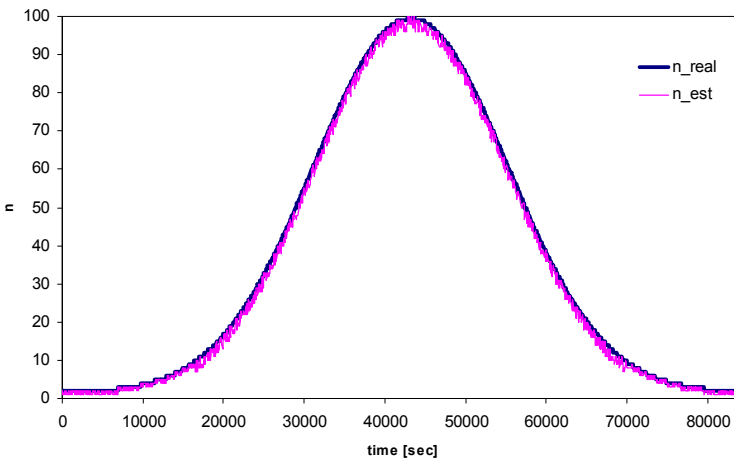


Fig. 8. Number of active SSs during the simulation (actual and estimated).

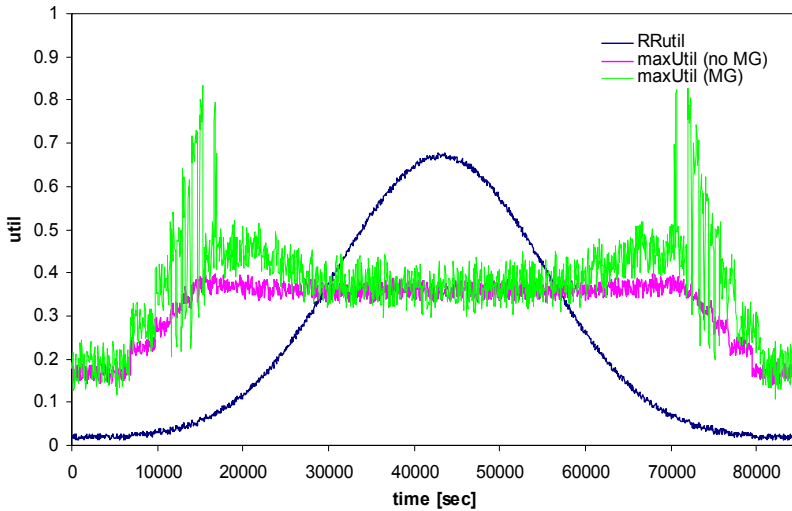


Fig. 9. Measured utilization of the transmission slots.

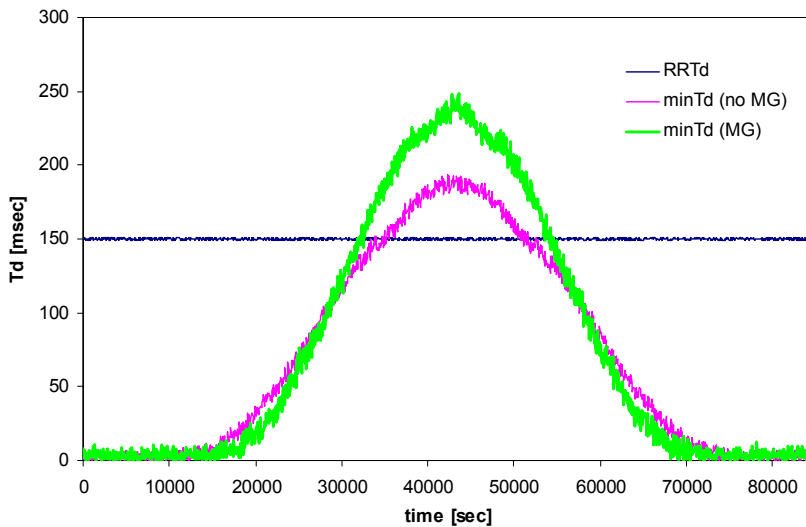


Fig. 10. Measured delay.

During the simulations we have used the following parameter values: $N_r = 10$, $N_{all} = 150$, $q = 1$, $t_{frm} = 10\text{ms}$.

From the results we can conclude that the contention based bandwidth request procedure which uses multicast grouping almost always outperforms the case where no multicasting groups are used. This is the case if our comparison criterion is based on the transmission slots utilization. The same conclusion is not entirely valid if the criterion is based on the average bandwidth access delay. The round-robin mode of operation, as expected,

outperforms the other modes of operation when the number of active users is close to the number of registered users.

5. Conclusion

The performance of a communication system (or protocol) is the major indicator for the successfulness of the standard that specifies that system. During the development process of new communication technologies, the aim is to increase the overall protocol performance, which automatically means to produce a better standard. The formal-based performance evaluation method described in this chapter, which uses SDL network prototyping, provides the most relevant results of the system performance without the need for its early and prematurely hardware implementation. This is crucial for the production of competitive communication products which will be free of hidden flows during the development process. The aim of this chapter was to emphasize the importance of conducting an early performance evaluation of the communication protocols and systems, and to suggest an appropriate solution for carrying out such an activity.

We have illustrated the proposed framework through an actual case study which targets the WiMAX bandwidth allocation methods. Three schemes were investigated: contention based bandwidth requesting without multicasting, contention based procedure with multicast grouping, and round-robin polling based bandwidth allocation scheme. With the help of the SDL performance model, we have found that the preferred scheme should be selected based on the network working conditions (i.e. number of active subscriber stations) and according to the performance criterion (transmission slots utilization or average bandwidth access delay).

The proposed new protocol entities which are product of the communication protocol engineering process are simple and accurate, and can be easily implemented at the BS side in order to optimize the performance of the WiMAX bandwidth request procedure.

The presented results are only a hint to the possible evaluation outcomes from the network emulations created using the proposed methodology.

6. Acknowledgment

This research is sponsored by the Faculty of Electrical Engineering and Information Technologies - Skopje, Ss. Cyril and Methodius University in Skopje, Macedonia, through the MOBIKS (Modeling and Optimization of Wireless Information-Communications Systems) project. The authors want to express gratitude to all participants involved in this project.

7. References

- ASN.1 (1993). ITU-T, Specification of abstract syntax notation one (ASN.1), *ITU- T Recommendation X.208, Technical Report, Telecommunication Standardization Sector of ITU*, March 1993.
- Bluetooth (2005). IEEE Std 802.15.1, Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Wireless Personal Area Networks (WPANs)
- Booch G. et al. (2000). OMG Unified Modeling Language Specification, *Version 1.3 First Edition: March 2000*.
- Chen Hui et al., (2010). Formal specification and verification of reconfigurable protocol stack for networked control system, *Proceedings of 2010 International Conference on*

- Networking, Sensing and Control (ICNSC)*, pp. 441 – 446, ISBN: 978-1-4244-6450-0, Chicago, USA, 10-12 April 2010
- Estelle (1989). Information Processing Systems - OSI: Estelle, A Formal Description Technique Based on an Extended State Transition Model, *International Standard 9074*, June 1989 ETSI. Available from <http://www.etsi.org>
- Gotzhein R., et al. (2009). Energy-Aware System Design with SDL, *Proceedings of the 14th international SDL conference on Design for motes and mobiles - SDL'09*, pp. 19-33, ISBN:3-642-04553-7 978-3-642-04553-0, September 22-24, Bochum, Germany
- Hall A. (1990), Seven Myths of Formal Methods, *IEEE Software*, Vol. 7, No. 5, Sept. 1990, pp. 11–19, September 1990 IEEE. Available from <http://www.ieee.org>
- Latkoski P. et al. (2010). Modeling and optimization of bandwidth request procedure in IEEE 802.16 networks, *Proceedings of the IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), 2010*, pp. 1469 – 1474, ISBN: 978-1-4244-8017-3, September 26-30, 2010, Istanbul, Turkey.
- Larsen K. G. et al. (1997). UPPAAL in a Nutshell, *International Journal on Software Tools for Technology Transfer (STTT)*, Volume 1, Numbers 1-2, pp.134-152
- Lopez J. et al., (2005). Security Protocols Analysis: A SDL-based Approach, *Computer Standards & Interfaces*, Vol. 27, No. 3, pp. 489-499, ISSN: 0920-5489
- LOTOS (2000). Information technology Enhancements to LOTOS (E-LOTOS), *SO/IEC JTC1/SC7, International Standard 15437*, July 2000
- Mitschele-Thiel A. (2001). *Systems Engineering with SDL: Developing Performance-Critical Communication Systems*, Wiley, ISBN: 978-0-471-49875-9, New York, USA
- MSC (2001). Series Z: Languages and General Software Aspects for Telecommunication Systems, Message Sequence Chart, *ITU-T Recommendation Z.120*, Geneva, Switzerland, 2001
- Petri C. A. (1996), Nets, Time and Space, *Theoretical Computer Science, Special Volume on Petri Nets*, Vol. 153, No. 1-2, pp. 3-48
- SDL (2011). Series Z: Languages and General Software Aspects for Telecommunication Systems, Specification and Description Language, *ITU-T Recommendation Z.100*, Geneva, Switzerland, Latest edition 2011
- Sherif M. H. & Sparrell D. K. (1992), Standards and Innovations in Telecommunications, *IEEE Communication Magazine*, Vol. 30, No. 7, July 1992, pp. 22–29, ISSN 0163-6804
- Sherif M. H. (2001). A Framework for Standardization in Telecommunications and Information Technology, *IEEE Communications Magazine*, No.4, April 2001, pp. 94-100, ISSN 0163-6804
- Showk A., et al. (2009). Modeling LTE protocol for mobile terminals using a formal description technique, *Proceedings of the 14th international SDL conference on Design for motes and mobiles - SDL'09*, pp. 222-238, ISBN:3-642-04553-7 978-3-642-04553-0, September 22-24, Bochum, Germany SPIN. Available from <http://netlib.sandia.gov/spin/index.html>
- TTCN (2006). ITU-T, Recommendation Z.140 Tree and Tabular Combined Notation (TTCN), March 2006.
- Wing J. M. (1990), A Specifier's Introduction to Formal Methods, *IEEE Computer*, Vol. 23, No. 9, pp. 8-24, September 1990
- WiMAX (2010). IEEE Std 802.16-2004, IEEE Standard for Local and metropolitan area networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems
- WiFi (2007) IEEE Std IEEE 802.11, IEEE Standard for Wireless LAN Medium Access Control and Physical Layer Specification. 3GPP. Available from <http://www.3gpp.org>

Cell Dwell Time and Channel Holding Time Relationship in Mobile Cellular Networks

Anum L. Enlil Corral-Ruiz¹, Felipe A. Cruz-Pérez¹
and Genaro Hernández-Valdez²

¹*Electrical Engineering Department, CINVESTAV-IPN*

²*Electronics Department, UAM-A
Mexico*

1. Introduction

Channel holding time (CHT) is of paramount importance for the analysis and performance evaluation of mobile cellular networks. This time variable allows one to derive other key system parameters such as channel occupancy time, new call blocking probability, and handoff call dropping probability. CHT depends on cellular shape, cell size, user's mobility patterns, used handoff scheme, and traffic flow characteristics. Traffic flow characteristics are associated with unencumbered service time (UST), while the overall effects of cellular shape, users' mobility, and handoff scheme are related to cell dwell time (CDT).

For convenience and analytical/computational tractability, the teletraffic analysis of mobile cellular networks has been commonly performed under the unrealistic assumption that CDT and/or CHT follow the negative exponential distribution (Lin et al., 1994; Hong & Rappaport, 1986). However, a plenty of evidences showed that these assumptions are not longer valid (Wang & Fan, 2007; Christensen et al., 2004; Fang, 2001, 2005; Orlik & Rappaport, 1998; Fang & Chlamtac, 1999; Fang et al., 1999; Alfa & Li, 2002; Rahman & Alfa, 2009; Soong & Barria, 2000; Yeo & Jun, 2002; Pattaramalai, et al., 2007). Recent papers have concluded that in order to capture the overall effects of users' mobility, one needs suitable models for CDT distribution (Lin, 1994; Hong & Rappaport, 1986). In specific, the use of general distributions for modeling this time variable has been highlighted. In this research direction, some authors have used Erlang, gamma, uniform, deterministic, hyper-Erlang, sum of hyper-exponentials, log-normal, Pareto, and Weibull distributions to model the pdf of CDT; see (Wang & Fan, 2007; Fang, 2001, 2005; Orlik & Rappaport, 1998; Fang & Chlamtac, 1999; Fang et al., 1999; Rahman & Alfa, 2009; Pattaramalai et al., 2007, 2009; Hidata et al., 2002; Thajchayapong & Toguz, 2005; Khan & Zeghlache, 1997; Zeng et al. 2002; Kim & Choi, 2009) and the references therein. Fang in (Fang, 2001)) emphasizes the use of phase-type (PH) distributions for modeling CDT. The reason is twofold. First, PH distributions provide accurate description of the distributions of different time variables in wireless cellular networks, while retaining the underlying Markovian properties of the distribution. Markovian properties are essential in generating tractable queuing models for cellular networks. Second, there have been major advances in fitting PH distributions to real data. Among the PH probability distributions, the use of either Coxian or Hyper-Erlang distributions are of

particular interest because their universality property (i.e, they can be used to approximate any non negative distribution arbitrarily close) (Soong & Barria, 2000; Fang, 2001).

Due to the discrepancy and the wide variety of proposed models, it appears mandatory to investigate the implications of the cell dwell time distribution on channel holding time characteristics in mobile wireless networks. This is the topic of research of the present chapter. Let us describe the related work reported in this research direction.

1.1 Previously related work

In (Fang, 2001; Zeng et al. 2002), it is observed that, depending on the variance of CDT, the mean channel holding time for new calls (CHT_n) can be greater than the mean channel holding time for handoff calls (CHTh). However, in these works, it is neither explained nor discussed the physical reasons for this observed behavior. This phenomenon (which is addressed in Section 3.1) and the lack of related published numerical results have motivated the present chapter.

Most of the previously published papers that have developed mathematical models for the performance analysis of mobile cellular systems considering general probability distribution for cell dwell time have either only presented numerical results for the Erlang (Wang & Fan, 2007; Fang et al., 1999; Rahman & Alfa, 2009; Kim & Choi, 2009) or Gamma distributions with shape parameter greater than one¹ (Yeo & Jun, 2002; Fang, 2005), and/or only for the CHTh² (Fang, 2001; Fang & Chlamtac, 1999), or have not presented numerical results at all (Fang, 2005; Alfa & Li, 2002; Soong & Barria, 2000). Thus, numerical results both for values of the coefficient of variation (CoV) of CDT greater than one and/or for the CHT_n have been largely ignored. Exceptions of this are the papers (Orlik & Rappaport, 1998; Fang et al., b, 1997; Pattaramalai, et al., 2009).

On the other hand, probability distribution of CHT has been determined under the assumption of the staged distributions sum of hyper-exponentials, Erlang, and hyper-Erlang for the CDT (Orlik & Rappaport, 1998; Soong & Barria, 2000). However, to the best of the authors' knowledge, probability distribution of CHT in mobile cellular networks with neither hyper-exponential nor Coxian distributed CDT has been previously reported in the literature.

In this Chapter, the statistical relationships among residual cell dwell time (CDTr), CDT, and CHT for new and handoff calls are revisited and discussed. In particular, under the assumption that UST is exponentially distributed and CDT is phase-type distributed, a novel algebraic set of general equations that examine the relationships both between CDT and CDTr and between CDT and channel holding times are obtained. Also, the condition upon which the mean CHT_n is greater than the mean CHTh is derived. Additionally, novel mathematical expressions for determining the parameters of the resulting CHT distribution as functions of the parameters of the CDT distribution are derived for hyper-exponentially or Coxian distributed CDT.

¹ For the Erlang distribution and for the Gamma distribution with shape parameter greater than one, the coefficient of variation of its associated random variable is smaller than one.

² Also referred as handoff call channel occupancy time.

2. System model

A homogeneous multi-cellular system with omni-directional antennas located at the centre of each cell is assumed; that is, the underlying processes and parameters for all cells within the cellular network are the same, so that all cells are statistically identical. As mobile user moves through the coverage area of a cellular network, several variables can be defined: cell dwell time, residual cell dwell time, channel holding time, among others. These time variables are defined in the next section.

2.1 Definition of time interval variables

In this section the different time interval variables involved in the analytical model of a mobile cellular network are defined.

First, the *unencumbered service time* per call x_s (also known as the *requested call holding time* (Alfa and Li, 2002) or *call holding duration* (Rahman & Alfa, 2009)) is the amount of time that the call would remain in progress if it experiences no forced termination. It has been widely accepted in the literature that the unencumbered service time can adequately be modeled by a negative exponentially distributed random variable (RV) (Lin et al., 1994; Hong & Rappaport, 1986). The RV used to represent this time is \mathbf{X}_s and its mean value is $E\{\mathbf{X}_s\} = 1/\mu$.

Now, *cell dwell time* or *cell residence time* $x_d^{(j)}$ is defined as the time interval that a mobile station (MS) spends in the j -th (for $j = 0, 1, \dots$) handed off cell irrespective of whether it is engaged in a call (or session) or not. The random variables (RVs) used to represent this time are $\mathbf{X}_d^{(j)}$ (for $j = 0, 1, \dots$) and are assumed to be independent and identically generally phase-type distributed. For homogeneous cellular systems, this assumption has been widely accepted in the literature (Lin et al., 1994; Hong & Rappaport, 1986; Orlik & Rappaport, 1998; Fang & Chlamtac, 1999; Alfa & Li, 2002; Rahman & Alfa, 2009).

In this Chapter, cell dwell time is modeled as a general phase-type distributed RV with the probability distribution function (pdf) $f_{\mathbf{X}_d}(t)$, the cumulative distribution function (CDF) $F_{\mathbf{X}_d}(t)$, and the mean $E\{\mathbf{X}_d\} = 1/\eta$.

The *residual cell dwell time* x_r is defined as the time interval between the instant that a new call is initiated and the instant that the user is handed off to another cell. Notice that residual cell dwell time is only defined for new calls. The RV used to represent this time is \mathbf{X}_r . Thus, the probability density function (pdf) of \mathbf{X}_r , $f_{\mathbf{X}_r}(t)$, can be calculated in terms of \mathbf{X}_d using the excess life theorem (Lin et al., 1994)

$$f_{\mathbf{X}_r}(t) = \frac{1}{E[\mathbf{X}_d]} [1 - F_{\mathbf{X}_d}(t)] \quad (1)$$

where $E[\mathbf{X}_d]$ and $F_{\mathbf{X}_d}(t)$ are, respectively, the mean value and cumulative probability distribution function (CDF) of \mathbf{X}_d .

Finally, we define *channel holding time* as the amount of time that a call holds a channel in a particular cell. In this Chapter we distinguish between channel holding times for handed off (CHTh) and channel holding time for new calls (CHTn). CHTh (CHTn) is represented by the random variable $\mathbf{X}_c^{(h)}$ ($\mathbf{X}_c^{(N)}$).

3. Mathematical analysis

3.1 Relationship between X_d and X_r

The relationship between the probability distributions of CDT and CDTr is determined by the residual life theorem. In Table I some particular typically considered CDT distributions and the corresponding CDTr distributions obtained by applying the residual life theorem are shown.

Probability density function of cell dwell time or its Laplace transform.	Probability density function of residual cell dwell time or its Laplace transform.	Parameters of $f_{X_r}(t)$ as a function of the parameters of $f_{X_d}(t)$
Negative Exponential $\eta e^{-\eta t}$	Negative Exponential $\eta e^{-\eta t}$	
Erlang of k order $\frac{\eta^k t^{k-1}}{(k-1)!} e^{-\eta t}$	Hyper-Erlang with k stages of 1, 2, ... and k phases $\sum_{j=1}^k P_j^{(N)} \frac{\eta^j (t)^{j-1}}{(j-1)!} e^{-\eta t}$	$P_j^{(N)} = \frac{1}{k}$
Hyper-exponential of n order $\sum_{i=1}^n P_i \lambda_i e^{-\lambda_i t}$	Hyper-exponential of n order $\sum_{i=1}^n P_i^{(N)} \lambda_i e^{-\lambda_i t}$	$P_i^{(N)} = \frac{P_i \prod_{j=1, j \neq i}^n \lambda_j}{\sum_{i=1}^n P_i \prod_{j=1, j \neq i}^n \lambda_j}$
Hypo-exponential ³ of m order $f_{X_d}^*(s) = \prod_{i=1}^m \frac{\eta_i}{s + \eta_i}$	Generalized Coxian of m order $f_{X_r}^*(s) = \sum_{i=1}^m P_i^{(N)} \prod_{j=i}^m \frac{\eta_j}{s + \eta_j}$	$P_i^{(N)} = \frac{1}{\sum_{j=1}^m \frac{1}{\eta_j}}$
Hyper-Erlang of common order (n, m) $\sum_{i=1}^n P_i \frac{\eta_i^m t^{m-1}}{(m-1)!} e^{-\eta_i t}$	Hyper-Erlang of non common order $\sum_{i=1}^{nm} P_i^{(N)} \frac{\left(\eta_{\lfloor \frac{i-1}{m} \rfloor + 1}\right)^z t^{z-1}}{(z-1)!} e^{-\eta_{\lfloor \frac{i-1}{m} \rfloor + 1} t}$ $z = \text{mod}\left(\frac{i-1}{m}\right) + 1$	$P_i^{(N)} = \frac{P_i \prod_{l=1, l \neq i}^n \eta_l}{\sum_{k=1}^n P_k m \prod_{l=1, l \neq k}^n \eta_l}$ $z = \text{mod}\left(\frac{i-1}{m}\right) + 1$
Constant $\{\delta(t - E\{X_d\}) ; t = E\{X_d\}$ $\{ 0 ; \text{otherwise}$	Uniform $\frac{1}{E\{X_d\}} ; 0 \leq t \leq E\{X_d\}$	
Coxian of m order $f_{X_d}^*(s)$ $= \sum_{i=1}^m P_i \prod_{j=1}^i \frac{\eta_j}{(s + \eta_j)}$	Generalized Coxian of m order $f_{X_r}^*(s)$ $= \sum_{j=1}^{\frac{m(m+1)}{2}} P_j^{(N)} \left(\prod_{k=h(j)}^{f(j)} \frac{\eta_k}{s + \eta_k} \right)$	$P_j^{(N)}$ $= \frac{P_{f(j)} \prod_{k=1, k \neq h(j)}^m \eta_k}{\sum_{i=1}^m \left[\left(\prod_{k=1, k \neq i}^m \eta_k \right) \left(\sum_{l=i}^m P_l \right) \right]}$

³ Also known as Generalized Erlang.

Probability density function of cell dwell time or its Laplace transform.	Probability density function of residual cell dwell time or its Laplace transform.	Parameters of $f_{X_r}(t)$ as a function of the parameters of $f_{X_d}(t)$
Generalized Coxian of m order $f_{X_d}^*(s) = \sum_{j=1}^{\frac{m(m+1)}{2}} P_j \left(\prod_{k=h(j)}^{f(j)} \frac{\eta_k}{s + \eta_k} \right)$	Generalized Coxian of m order $f_{X_r}^*(s) = \sum_{j=1}^{\frac{m(m+1)}{2}} P_j^{(N)} \left(\prod_{k=h(j)}^{f(j)} \frac{\eta_k}{s + \eta_k} \right)$	$P_j^{(N)} = \frac{\prod_{k=1}^m \eta_k \left(\sum_{n=j-h(j)+1}^j P_n \right)}{k \neq h(j)}$ $; A = \sum_{i=1}^m \left[\prod_{\substack{j=1 \\ j \neq i}}^m \eta_j \left(\sum_{k=i}^m P_{\frac{i^2-i+2}{2}} + \sum_{l=1}^{i-1} P_{\frac{i^2-i+2}{2}+l} \right) \right]$
Gamma $\frac{x^{k-1} e^{-\frac{x}{\theta}}}{\Gamma(k)\theta^k}$	$\frac{1}{k\theta} \left[1 - P\left(k, \frac{x}{\theta}\right) \right]$	
Weibull $\frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}$	$\frac{1}{\lambda \Gamma\left(1 + \frac{1}{k}\right)} e^{-\left(\frac{x}{\lambda}\right)^k}$	
Pareto $\frac{\alpha X_m^\alpha}{t^{\alpha+1}} ; t > X_m$	$\begin{cases} \frac{\alpha - 1}{\alpha X_m} \left[\left(\frac{X_m}{t}\right)^\alpha \right] ; t > X_m \\ \frac{\alpha - 1}{\alpha X_m} ; 0 \leq t \leq X_m \end{cases}$	

Table I. Examples of corresponding distribution for X_r given the distribution of X_d .

The functional relationship between the moments of the residual cell dwell time and the cell residual time was obtained in (Kleinrock, 1975) applying the Laplace transform to the residual life theorem. That is,

$$\mathcal{L}\{f_{X_r}(t)\} = \mathcal{L}\left\{\frac{1}{E\{X_d}\}\right\} - \mathcal{L}\left\{\frac{1}{E\{X_d}\} F_{X_d}(t)\right\} \tag{2}$$

This equation can be rewritten as

$$f_{X_r}^*(s) = \left[\frac{1}{s}\right] \frac{1}{E\{X_d}\} [1 - f_{X_d}^*(s)] \tag{3}$$

The n -th moment of the residual cell dwell time in terms of the moments of the cell dwell time can be obtained by deriving n times equation (3) with negative argument and substituting $s=0$. Then (Kleinrock, 1975),

$$E\{\mathbf{X}_r\}^n = \frac{E\{\mathbf{X}_d\}^{n+1}}{(n+1)E\{\mathbf{X}_d\}} \quad (4)$$

The mean residual cell dwell time as function of the moments of cell dwell time can be obtained as (Kleinrock, 1975)

$$E\{\mathbf{X}_r\} = \frac{E\{\mathbf{X}_d\}}{2} + \frac{VAR(\mathbf{X}_d)}{2E\{\mathbf{X}_d\}} \quad (5)$$

$E\{\mathbf{X}_d\}$ and $VAR(\mathbf{X}_d)$ represent the mean and variance of CDT, respectively. Considering this equation and that $CoV\{\mathbf{X}_d\}$ represents the coefficient of variation of CDT, the condition for which the mean CDT_r is greater than the mean CDT ($E\{\mathbf{X}_r\} > E\{\mathbf{X}_d\}$) is given by

$$\frac{E\{\mathbf{X}_d\}}{2} + \frac{VAR\{\mathbf{X}_d\}}{2E\{\mathbf{X}_d\}} > E\{\mathbf{X}_d\} \quad (6)$$

$$CoV\{\mathbf{X}_d\} > 1$$

In this way, the relationship between mean CDT and mean CDT_r only depends on the value of the *CoV* of CDT. Thus, the mean CDT_r is greater than the mean CDT (i.e., $E\{\mathbf{X}_r\} > E\{\mathbf{X}_d\}$) when the *CoV* of CDT is greater than one. This behavior (i.e., $E\{\mathbf{X}_r\} > E\{\mathbf{X}_d\}$) may seem to be counterintuitive due to the fact that, for a particular realization and by definition, CDT_r cannot be greater than CDT⁴. This occurs because in such conditions there is a high variability on the cell dwell times in different cells and it is more probable to start new calls on cells where users spent more time. Then, residual cell dwell times tend to be greater than the mean CDT. This phenomenon that may seem to be counterintuitive is now explained and mathematically formulated in this Chapter.

3.2 Channel holding time distribution for handed off and new calls

Channel holding times for handed off and new calls (denoted by $X_c^{(h)}$ and $X_c^{(N)}$, respectively) are given by the minimum between UST and CDT or CDT_r, respectively. The CDF of the CHTh and CHTn are, respectively, given by

$$F_{X_c^{(h)}}(t) = 1 - [1 - F_{X_s}(t)][1 - F_{X_d}(t)] \quad (7)$$

$$F_{X_c^{(N)}}(t) = 1 - [1 - F_{X_s}(t)][1 - F_{X_r}(t)] \quad (8)$$

Due to the fact that the Laplace transform of the pdf of both UST and CDT_r are rational functions, the Laplace transform of the pdf of CHTn can be obtained using the Residue Theorem as follows (Wang & Fan, 2007)

$$f_{X_c^{(N)}}^*(s) = f_{X_s}^*(s) + s \sum_{p \in \Omega_{X_s}} \underset{\xi = p + s}{Res} \frac{f_{X_r}^*(\xi) f_{X_s}^*(s-\xi)}{\xi - s-\xi} \quad (9)$$

where $p \in \Omega_{X_s}$ is the set of poles of $f_{X_s}^*(s)$, and $f_X^*(s)$ is the Laplace transform of pdf $f_X(t)$. A similar expression can be obtained for the Laplace transform of the pdf of the channel holding time for handed off calls by replacing residual cell dwell time (X_r) by cell dwell time (X_d).

⁴ Note that the beginning of CDT_r is randomly chosen within the CDT interval.

Under the condition that UST is general phase type (PH) distributed, the authors of (Alfa & Li, 2002) prove that the CDT is PH distributed if and only if the CHTn is PH distributed or the CHTh is PH distributed.

The probability distributions of CHTn and CHTh for different staged probability distributions of CDT assuming that the UST is exponentially distributed are shown in Table II. The first entry of this table is a well known result⁵. In (Soong & Barria, 2000), it was shown that when CDT has Erlang or hyper-Erlang distribution, channel holding times have the uniform Coxian and hyper-uniform Coxian distribution, respectively. Uniform Coxian is a special case of the Coxian distribution where all the phases have the same parameter (Perros & Khoshgoftaar, 1989). The hyper-uniform Coxian distribution is a mixture of uniform Coxian distributions.

pdf of cell dwell time.	pdf of channel holding time for new calls or its Laplace Transform.	pdf of channel holding time for handed off calls or its Laplace Transform.
Exponential (Lin et al., 1994)	Exponential $(\mu + \eta)e^{-(\mu+\eta)t}$	Exponential $(\mu + \eta)e^{-(\mu+\eta)t}$
Erlang of k -th order (Soong & Barria, 2000)	Uniform Coxian of k -th order $f_{X_c^*}^*(s) = \sum_{i=1}^k P_i^{O(N)} \prod_{j=1}^i \frac{\mu + \eta}{s + \mu + \eta}$	Uniform Coxian of k -th order $f_{X_c^*(h)}^*(s) = \sum_{i=1}^k P_i^{O(h)} \prod_{j=1}^i \frac{\mu + \eta}{s + \mu + \eta}$
Hyper-Erlang of common order (n, m) (Soong & Barria, 2000)	Hyper-Uniform Coxian $f_{X_c^*}^*(s) = \sum_{i=1}^k P_i^{O(N)} \prod_{j=1}^z \frac{\mu + \eta_l}{s + \mu + \eta_l}$ where $z = \text{mod}\left(\frac{i-1}{m}\right) + 1$ $l = \left\lfloor \frac{i-1}{m} \right\rfloor + 1$	Hyper-Uniform Coxian $f_{X_c^*(h)}^*(s) = \sum_{i=1}^k P_i^{O(h)} \prod_{j=1}^z \frac{\mu + \eta_l}{s + \mu + \eta_l}$ where $z = \text{mod}\left(\frac{i-1}{m}\right) + 1$ $l = \left\lfloor \frac{i-1}{m} \right\rfloor + 1$
Hyper-exponential	Hyper-exponential $\sum_{i=1}^n P_i^{(N)} (\mu + \eta_i) e^{-(\mu+\eta_i)t}$ where $P_i^{(N)} = \frac{P_i \prod_{j=1}^n \eta_j}{\sum_{i=1}^n P_i \prod_{j=1}^n \eta_j}$ $j \neq i$	Hyper-exponential $\sum_{i=1}^n P_i (\mu + \eta_i) e^{-(\mu+\eta_i)t}$

⁵ Authors in (Lin et al., 1994) give a condition under which the channel holding time is exponentially distributed, that is, the cell residence time needs to be exponentially distributed.

pdf of cell dwell time.	pdf of channel holding time for new calls or its Laplace Transform.	pdf of channel holding time for handed off calls or its Laplace Transform.
Coxian	Generalized Coxian $f_{X_c^{(N)}}^*(s) = \sum_{j=1}^{\frac{m(m+1)}{2}} P_j^{O(N)} \left(\prod_{k=h(j)}^{f(j)} \frac{\eta_k}{s + \mu + \eta_k} \right)$ where $P_j^{O(N)} = \left[\prod_{i=h(j)}^{f(j)-1} \frac{\eta_i}{\mu + \eta_i} \right] \left[P_j^{(N)} + \sum_{k=f(j)+1}^m P_{\frac{k^2-k+2}{2}}^{(N)} \left(\frac{\mu}{\mu + \eta_{f(j)}} \right) \right]$	Coxian $f_{X_c^{(h)}}^*(s) = \sum_{j=1}^m P_j^{O(h)} \prod_{i=1}^j \frac{\eta_i}{(s + \mu + \eta_i)}$ where $P_j^{O(h)} = \left[\prod_{i=1}^{j-1} \frac{\eta_i}{\mu + \eta_i} \right] \left[P_j + \sum_{k=j+1}^m P_k \left(\frac{\mu}{\mu + \eta_j} \right) \right]$
Generalized Coxian (Corral-Ruiz et al., a, 2010)	Generalized Coxian $f_{X_c^{(N)}}^*(s) = \sum_{j=1}^{\frac{m(m+1)}{2}} P_j^{O(N)} \left(\prod_{k=h(j)}^{f(j)} \frac{\eta_k}{s + \mu + \eta_k} \right)$ where $P_j^{O(N)} = \left[\prod_{i=h(j)}^{f(j)-1} \frac{\eta_i}{\mu + \eta_i} \right] \left[P_j^{(N)} + \sum_{k=f(j)+1}^m P_{\frac{k^2-k+2}{2}}^{(N)} \left(\frac{\mu}{\mu + \eta_{f(j)}} \right) \right]$	Generalized Coxian $f_{X_c^{(h)}}^*(s) = \sum_{j=1}^{\frac{m(m+1)}{2}} P_j^{O(h)} \left(\prod_{k=h(j)}^{f(j)} \frac{\eta_k}{s + \mu + \eta_k} \right)$ where $P_j^{O(h)} = \left[\prod_{i=h(j)}^{f(j)-1} \frac{\eta_i}{\mu + \eta_i} \right] \left[P_j + \sum_{k=f(j)+1}^m P_{\frac{k^2-k+2}{2}} \left(\frac{\mu}{\mu + \eta_{f(j)}} \right) \right]$

Table II. Examples of corresponding distributions for $X_c^{(N)}$ and $X_c^{(h)}$.

Next, it is shown that when the UST is exponentially distributed and CDT has hyper-exponential distribution of order n , the distribution of CHTh has also a hyper-exponential distribution of order n . Similarly, when CDT has Coxian distribution of order n , the distribution of CHTn has also a Coxian distribution of order n .

3.2.1 Case 1: Hyper-exponentially distributed cell dwell time

Considering that CDT has a hyper-exponential pdf of order n given by

$$f_{X_d}(t) = \sum_{j=1}^n P_j \eta_j e^{-\eta_j t} \tag{10}$$

For exponentially distributed UST and using (4), the CDF of the CHTh can be expressed as follows

$$F_{X_c^{(h)}}(t) = 1 - [e^{-\mu t}] \left[\sum_{i=1}^n P_i e^{-\eta_i t} \right]$$

$$F_{X_c^{(h)}}(t) = 1 - \sum_{i=1}^n P_i e^{-(\mu+\eta_i)t}$$
(11)

This expression corresponds to a hyper-exponential distribution of order n with phase parameters $\mu + \eta_i$ and probabilities P_i of choosing stage i (for $i = 1, \dots, n$).

As the CDTr is hyper-exponentially distributed when CDT has hyper-exponential distribution, the CHTn is also hyper-exponentially distributed. In this case, the probability of choosing stage i (for $i = 1, \dots, n$) is given by

$$P_i^{(N)} = \frac{P_i \prod_{j=1, j \neq i}^n \eta_j}{\sum_{j=1}^n P_j \prod_{k=1, k \neq j}^n \eta_k}$$
(12)

3.2.2 Case 2: Coxian distributed cell dwell time

Considering that cell dwell time has an m -th order Coxian distribution (which diagram of phases is shown in Fig. 1) with Laplace transform of its pdf given by

$$f_{X_d}^*(s) = \sum_{j=1}^m P_j \prod_{i=1}^j \frac{\eta_i}{(s+\eta_i)}$$
(13)

where

$$P_j = \alpha_j \prod_{i=1}^{j-1} (1 - \alpha_i)$$
(14)

$(1-\alpha_i)$ represents the probability of passing from the i -th phase to the $(i+1)$ -th one.

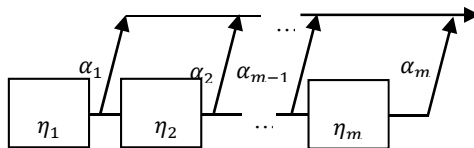


Fig. 1. Diagram of phases of the considered Coxian distribution of order m for modeling cell dwell time.

For exponentially distributed UST and using (9), the Laplace transforms of the pdf of CHTh and CHTn are given by

$$f_{X_c^{(h)}}^*(s) = \frac{\mu}{s+\mu} + \frac{s}{s+\mu} [f_{X_d}^*(s + \mu)]$$
(15)

$$f_{X_c^{(N)}}^*(s) = \frac{\mu}{s+\mu} + \frac{s}{s+\mu} [f_{X_r}^*(s + \mu)]$$
(16)

Replacing (13) into (15), it can be written as

$$f_{\mathbf{X}_c^{(h)}}^*(s) = \sum_{j=1}^m P_j^{O(h)} \prod_{i=1}^j \frac{\eta_i}{(s+\mu+\eta_i)} \tag{17}$$

where

$$P_j^{O(h)} = \left[\prod_{i=1}^{j-1} \frac{\eta_i}{\mu+\eta_i} \right] \left[P_j + \sum_{k=j+1}^m P_k \left(\frac{\mu}{\mu+\eta_j} \right) \right] \tag{18}$$

for $i = 1, \dots, m$. Then, CHTh has also a Coxian distribution of order m but with parameters $(\mu + \eta_i)$, for $i = 1, \dots, m$.

On the other hand, the Laplace transform of the residual cell dwell time can be shown to be given by

$$f_{\mathbf{X}_r}^*(s) = \sum_{j=1}^{\frac{m(m+1)}{2}} P_j^{(N)} \left(\prod_{k=h(j)}^{f(j)} \frac{\eta_k}{s+\eta_k} \right) \tag{19}$$

where

$$P_j^{(N)} = \frac{P_{f(j)} \prod_{\substack{k=1 \\ k \neq h(j)}}^m \eta_k}{\sum_{i=1}^m \left[\left(\prod_{\substack{k=1 \\ k \neq i}}^m \eta_k \right) (\sum_{l=i}^m P_l) \right]} \tag{20}$$

$$f(j) = \left\lfloor \frac{1 \pm \sqrt{1+8(j-1)}}{2} \right\rfloor \tag{21}$$

$$h(j) = j - \frac{f(j)(f(j)-1)}{2} \tag{22}$$

for $j = 1, \dots, m(m+1)/2$. Substituting (19) into (16), Laplace transform of CHTn can be written as

$$f_{\mathbf{X}_c^{(N)}}^*(s) = \sum_{j=1}^{\frac{m(m+1)}{2}} P_j^{O(N)} \left(\prod_{k=h(j)}^{f(j)} \frac{\eta_k}{s+\mu+\eta_k} \right) \tag{23}$$

where

$$P_j^{O(N)} = \left[\prod_{i=h(j)}^{f(j)-1} \frac{\eta_i}{\mu+\eta_i} \right] \left[P_j^{(N)} + \sum_{k=f(j)+1}^m P_{\frac{k^2-k+2}{2}}^{(N)} \left(\frac{\mu}{\mu+\eta_{f(j)}} \right) \right] \tag{24}$$

Equation (23) corresponds to the Laplace transform of a generalized Coxian pdf.

The above analytical results show that CHTh (CHTn) has the same probability distribution as CDT (CDTr) but with different parameters of the phases, probabilities of reaching the absorbing state after each phase, and probabilities of choosing each stage. The detailed derivation of the last entry of Tables I and II (i.e., when cell dwell time has generalized Coxian distribution) is addressed in (Corral-Ruiz et al., a, 2010).

3.3 Relationship between $\mathbf{X}_c^{(h)}$ and $\mathbf{X}_c^{(N)}$

Using (15) and (16) it is straightforward to show that the mean values of CHTn and CHTh are, respectively, given by

$$E\{\mathbf{X}_c^{(N)}\} = \frac{1}{\mu} \left[1 - \frac{\eta}{\mu} [1 - f_{\mathbf{X}_d}^*(\mu)] \right] \quad (25)$$

$$E\{\mathbf{X}_c^{(h)}\} = \frac{1}{\mu} [1 - f_{\mathbf{X}_d}^*(\mu)] \quad (26)$$

At this point, it is important to mention that authors in (Fang, 2001; Zeng et al., 2002) stated that, depending on the variance of CDT, the mean CHTn can be greater than the mean CHTh. However, it was neither explained nor discussed the physical reasons for this observed behavior. This behavior occurs because the residual cell dwell times tend to increase as the variance of cell dwell time increases, as it was explained above.

Using (25) and (26), the condition for which the mean CHTn is greater than the mean CHTh, that is,

$$E\{\mathbf{X}_c^{(N)}\} > E\{\mathbf{X}_c^{(h)}\} \quad (27)$$

can be easily found. This condition is given by

$$f_{\mathbf{X}_d}^*(\mu) > \frac{\eta}{\mu + \eta} \quad (28)$$

Thus, the relationship between the mean new and handoff call channel holding times is determined by the mean values of both CDT and UST (μ) and by the Laplace transform of the pdf of CDT evaluated at the inverse of the mean UST.

Finally, in a similar way, the squared coefficient of variation for CHTn and CHTh can be shown to be given, respectively, by

$$CoV^2(\mathbf{X}_c^{(N)}) = \frac{-4\eta E\{\mathbf{X}_c^{(h)}\} + 2 \left[1 - \eta \frac{d f_{\mathbf{X}_d}^*(\mu)}{d\mu} \right]}{[E\{\mathbf{X}_c^{(N)}\} \mu]^2} - 1 \quad (29)$$

$$CoV^2(\mathbf{X}_c^{(h)}) = \frac{2}{(E\{\mathbf{X}_c^{(h)}\})^2 \mu} \left[\frac{d f_{\mathbf{X}_d}^*(\mu)}{d\mu} + E\{\mathbf{X}_c^{(h)}\} \right] - 1 \quad (30)$$

It can be shown that the n -th moments for new and handoff call channel holding times are given, respectively, by

$$E\{(\mathbf{X}_c^{(N)})^n\} = \frac{1}{\mu} \left[n E\{(\mathbf{X}_c^{(N)})^{n-1}\} - \eta E\{(\mathbf{X}_c^{(h)})^n\} \right] \quad (31)$$

$$E\{(\mathbf{X}_c^{(h)})^n\} = \frac{n}{\mu} \left((-1)^n \frac{d^n [f_{\mathbf{X}_d}^*(\mu)]}{d\mu^n} + E\{(\mathbf{X}_c^{(h)})^{n-1}\} \right) \quad (32)$$

4. Numerical results and discussion

In this section, numerical results on how the distribution of cell dwell time (CDT) affects the characteristics of channel holding time (CHT) are presented. We use different distributions to model CDT, say, negative-exponential, constant (deterministic), Pareto with shape parameter α in the range (1, 2] (i.e., when infinite variance is considered), Pareto with $\alpha > 2$

(i.e., when finite variance is considered), log-normal, gamma, hyper-Erlang of order (2,2), hyper-exponential of order 2, and Coxian of order 2. Three different mobility scenarios for the numerical evaluation are assumed: $E\{X_d\}=5\cdot E\{X_s\}$ (low mobility), $E\{X_d\}=E\{X_s\}$ (moderate mobility), and $E\{X_d\}=0.2\cdot E\{X_s\}$ (high mobility). In the plots of this section we use $E\{X_s\}=180$ s. In our numerical results, the effect of CoV and skewness of CDT on CHT characteristics is investigated. In the plots presented in this section, "HC" and "NC" stand for channel holding time for handoff calls (CHTh) and channel holding time for new calls (CHTn), respectively.

4.1 Cell dwell time distribution completely characterized by its mean value

Fig. 2 plots the mean value of both CHTn and CHTh versus the mean value of CDT when it is modeled by negative-exponential (EX), constant, and Pareto with $1 < \alpha \leq 2$ distributions. It is important to remark that all of these distributions are completely characterized by their respective mean values. As expected, Fig. 2 shows that, for the case when CDT is exponentially distributed, mean CHTn is equal to mean CHTh. An interesting observation on the results shown in Fig. 2 is that, irrespective of the mean value of CDT, there exists a significant difference between the mean value of CHTn when CDT is modeled as exponential distributed RV and the corresponding case when it is modeled by a heavy-tailed Pareto distributed RV (this behavior is especially true for the case when $\alpha=1.1$). Notice, however, that this difference is negligible for the case when $\alpha=2$ and high mobility scenarios (say, $E\{X_d\} < 50$ s) are considered. Similar behaviors are observed if mean CHTh is considered. Consequently, for high mobility scenarios where CDT can be statistical characterized by a Pareto distribution with shape parameter close to 2, the exponential distribution represents a suitable model for the CDT distribution. Fig. 2 also shows that, for

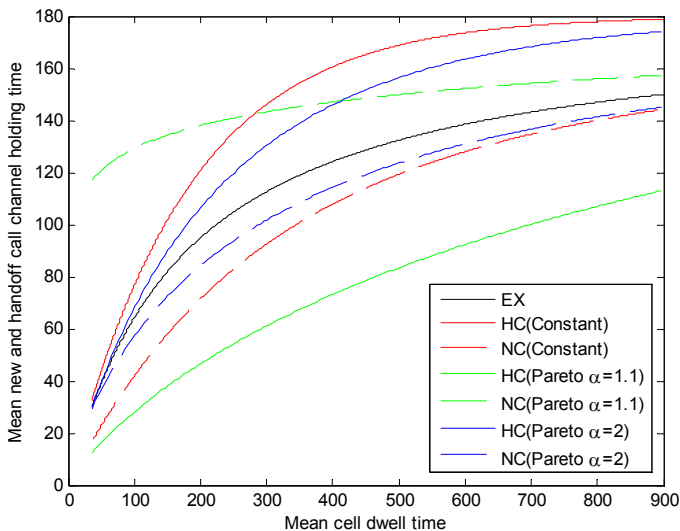


Fig. 2. Mean new and handoff call channel holding time for deterministic, negative exponentially, and Pareto distributed CDT against the mean CDT.

a given value of the mean CDT and considering the case when CDT is Pareto distributed with $\alpha=1.1$ ($\alpha=2$), mean CHTn always is greater (lower) than mean CHTh. This behavior can be explained by the combined effect of the following two facts. First, as α comes closer to 1 (2), the probability that CDT takes higher values increases (decreases). This fact contributes to increase (reduce) the mean CHTh. Second, in general, new calls are more probable to start on cells where users spent more time and, as α comes closer to 1, this probability increases. This fact contributes to increase mean CHTn relative to the mean CHTh. Then, the combined effect is dominated by the first (second) fact as α comes closer to 2 (1). This leads us to the behavior explained above and illustrated in Fig.2. It may be interesting to derive the condition upon which the mean CHTn is greater than the mean CHTh when CDT is heavy-tailed Pareto distributed. This represents a topic of our current research.

4.2 Cell dwell time distribution completely characterized by its first two moments

Fig. 3 plots the mean value of both CHTn and CHTh versus the CoV of CDT when it is modeled by Pareto with shape parameter $\alpha>2$, lognormal, and Gamma distributions; all of them with mean value equal to 180 s. It is important to remark that all of these distributions are completely characterized by their respective first two moments. Fig. 3 shows that both mean CHTn and mean CHTh are highly sensitive to the type of distribution of CDT; this fact is especially true for $CoV>2$. Notice that, for the particular case when $CoV=0$, the mean values of both CHTn and CHTh are identical to the corresponding values for the case when CDT is deterministic with mean value equals 180 s, as expected. Fig. 3 also shows that, for values of CoV of CDT greater than 1 (1.2), mean CHTn is greater that mean CHTh when CDT is Gamma (log-normal) distributed. On the other hand, when CDT is Pareto distributed and irrespective of the value of its CoV, CHTh always is greater that mean CHTn. This behavior is mainly due to the heavy-tailed characteristics of the Pareto distribution.

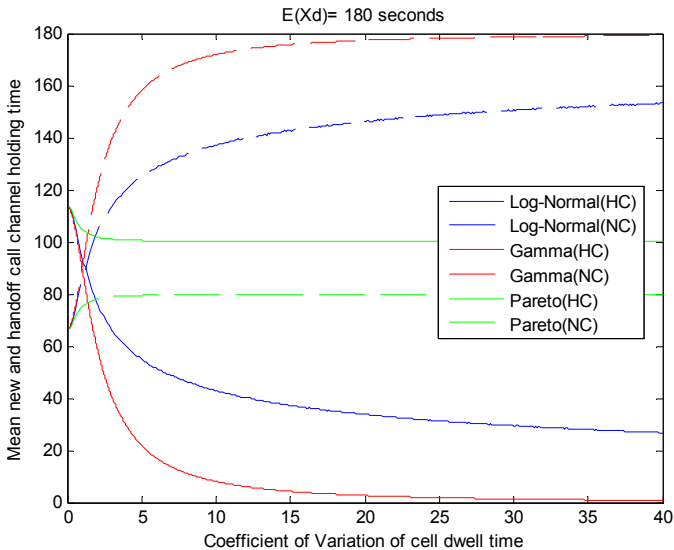


Fig. 3. Mean new and handoff call channel holding time for gamma, log-normal, and Pareto distributed cell dwell time versus CoV of cell dwell time.

4.3 Cell dwell time distribution completely characterized by its first three moments

Figs. 4, 5, and 6 (7, 8, and 9) plot the mean value (CoV) of both CHTn and CHTh versus both the CoV and skewness of CDT when it is modeled by hyper-Erlang (2,2), hyper-exponential of order 2, and Coxian of order 2 distributions, respectively. It is important to remark that all of these distributions are completely characterized by their respective first three moments. Results of (Johnson & Taaffe, 1989; Telek & Heindl, 2003) are used to calculate the parameters of these distributions as function of their first three moments. In Figs. 4 to 9, two different values for the mean CDT are considered: 36 s (high mobility scenario) and 900 s (low mobility scenario). From Figs. 2, 5 and 6 the following interesting observation can be extracted. Notice that, for the case when CDT is modeled by either hyper-exponential or Coxian distributions and irrespective of the mean value of CDT, the particular scenario where skewness and CoV of CDT are, respectively, equal to 2 and 1, corresponds to the case when CDT is exponential distributed (in the exponential case mean CHTn and mean CHTh are identical).

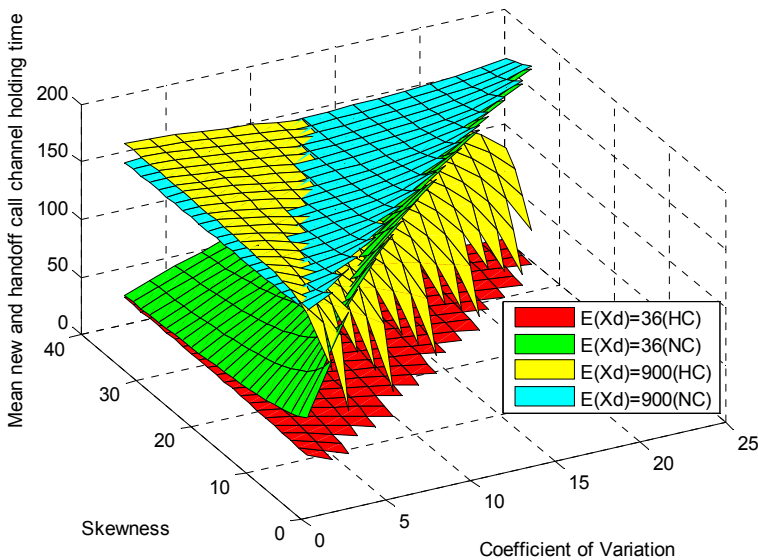


Fig. 4. Mean CHTn and mean CHTh for hyper-Erlang distributed CDT versus CoV and skewness of CDT, with the mean CDT as parameter.

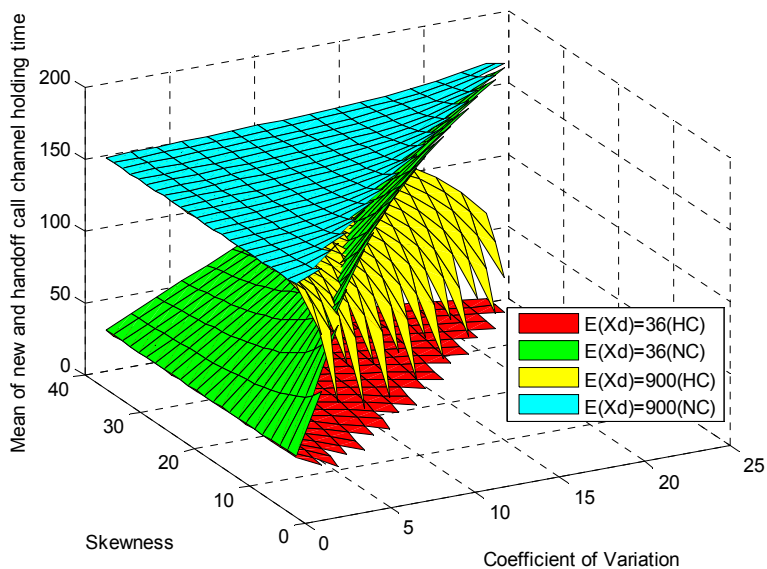


Fig. 5. Mean CHTn and mean CHTh for hyper-exponentially distributed CDT versus CoV and skewness of CDT, with the mean CDT as parameter.

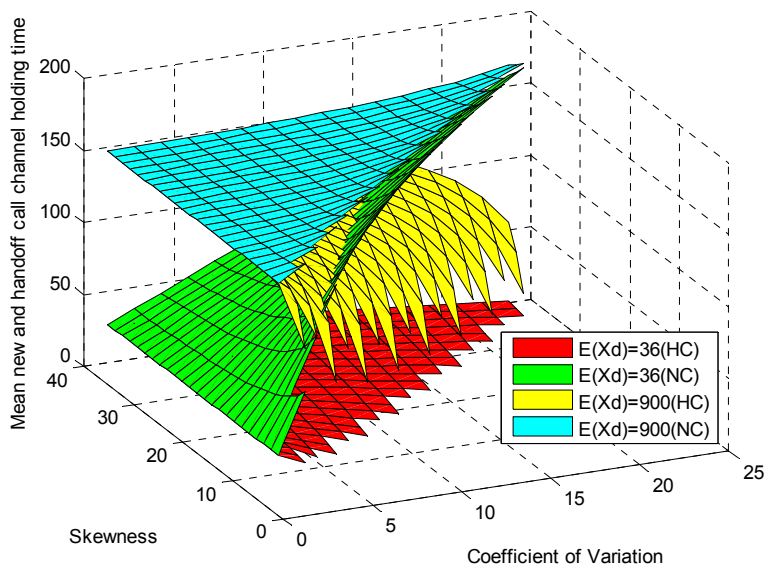


Fig. 6. Mean CHTn and mean CHTh for Coxian distributed cell dwell time versus CoV and skewness of cell dwell time, with the mean CDT as parameter.

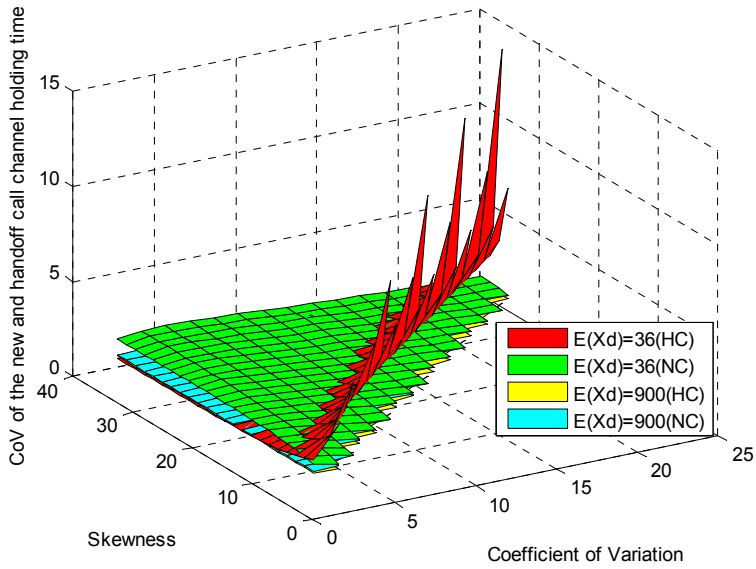


Fig. 7. CoV of CHT_n and CHT_h for hyper-Erlang distributed CDT versus CoV and skewness of CDT, with the mean CDT as parameter.

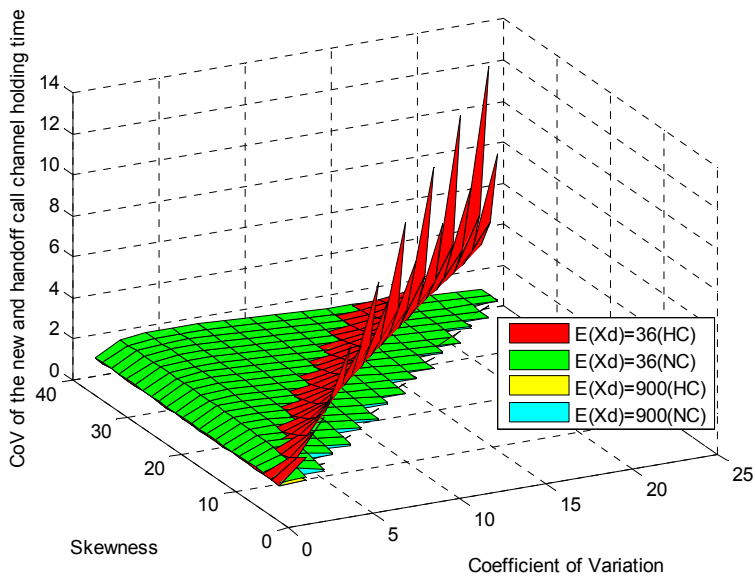


Fig. 8. CoV of CHT_n and CHT_h for hyper-exponential distributed CDT versus CoV and skewness of CDT, with the mean CDT as parameter.

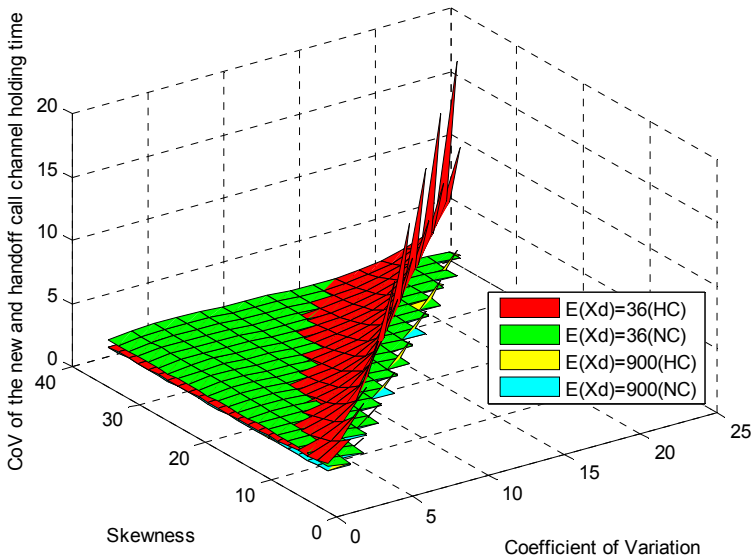


Fig. 9. CoV of CHTn and CHTh for Coxian distributed CDT versus CoV and skewness of cell dwell time, with the mean cell dwell time as parameter.

On the other hand, Fig. 4 shows that the case when hyper-Erlang distribution with skewness equals 2 and CoV equals 1 is used to model CDT does not strictly correspond to the exponential distribution; however, the exponential model represents a suitable approximation for CDT in this particular case. From Figs. 4 to 9, it is observed that the qualitative behavior of mean and CoV of both CHTn and CHTh is very similar for all the phase-type distributions under study. The small quantitative difference among them is due to moments higher than the third one. Analyzing the impact of moments of CDT higher than the third one on channel holding time characteristics represents a topic of our current research.

From Fig. 10 is observed that the difference among the mean values of CHTn and CHTh is strongly sensitive to the CoV of the CDT, while it is practically insensitive to the skewness of the CDT. This difference is higher for the case when the CDT is modeled as hyper-exponential distributed RV compared with the case when it is modeled as hyper-Erlang distributed RV. Also, it is observed that this difference remains almost constant for the entire range of values of the CoV of the CDT.

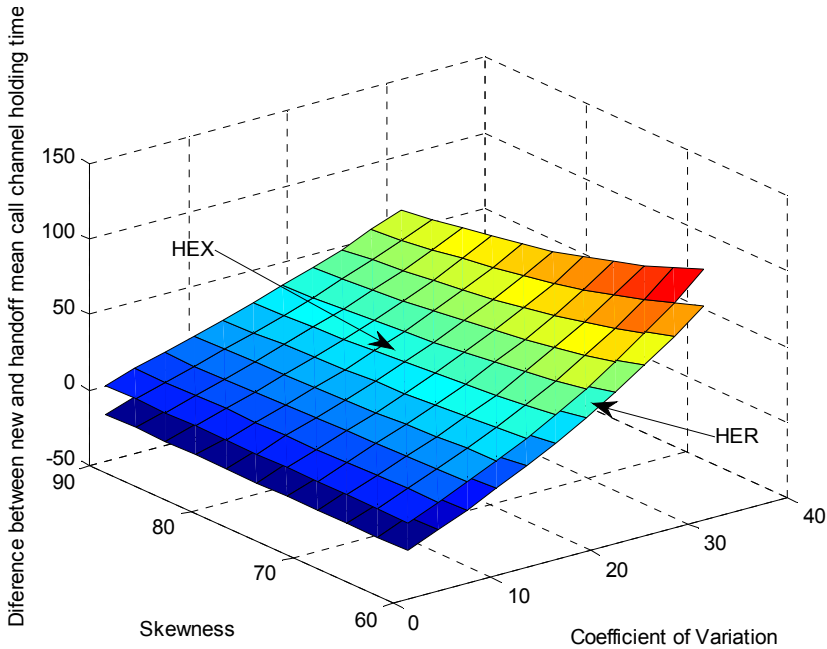


Fig. 10. Difference among the mean values of new and handoff call channel holding times for hyper-Erlang and hyper-exponential distributed cell dwell time versus CoV and skewness of cell dwell time, for the moderate-mobility scenario.

Finally, in Fig. 11 the mean channel holding time for new and handoff calls considering the gamma, hyper-Erlang (2,2), hyper-exponential of order 2, and Coxian of order 2 distributions for the cell dwell time are shown for different values of the coefficient of variation. The numerical results shown in Fig. 11 are obtained by equaling the first three moments of the different distributions to those of the gamma distribution. From Fig. 11, it is observed that for the hyper-exponential and Coxian distributions practically the same results are obtained for the mean channel holding time for both new and handoff calls. The differences among the other distributions are due to the fact that they differ on the higher order moments. To show this, the fourth standardized moment (i.e., excess kurtosis) of the different distributions is shown in Fig. 12 for different values of the coefficient of variation, equaling the first three moments of the different distributions to those of the gamma distribution. From Fig. 12, it is observed that the hyper-exponential and Coxian distributions practically have the same value of excess kurtosis but this differs for that of the gamma and hyper-Erlang distributions. The gamma distribution shows the more different value of the excess kurtosis and, therefore, for this distribution the more different values of the mean channel holding times in Fig. 11 are obtained. Then, it could be necessary to capture more than three moments, even though the lower order moments dominate in importance. Similar conclusion was drawn in (Gross & Juttijudata, 1997).

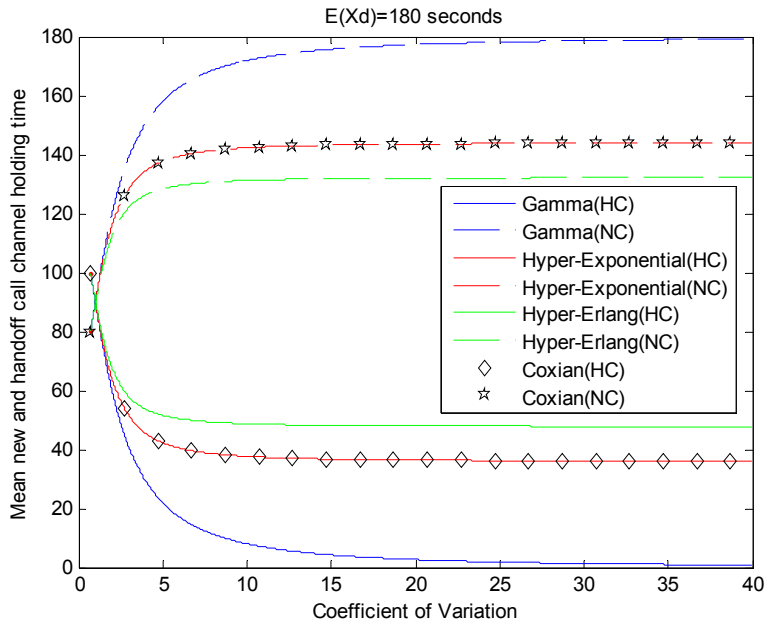


Fig. 11. Mean new and handoff call channel holding time for gamma, hyper-exponential (2), hyper-Erlang (2,2) and Coxian (2) distributed cell dwell time versus CoV of cell dwell time.

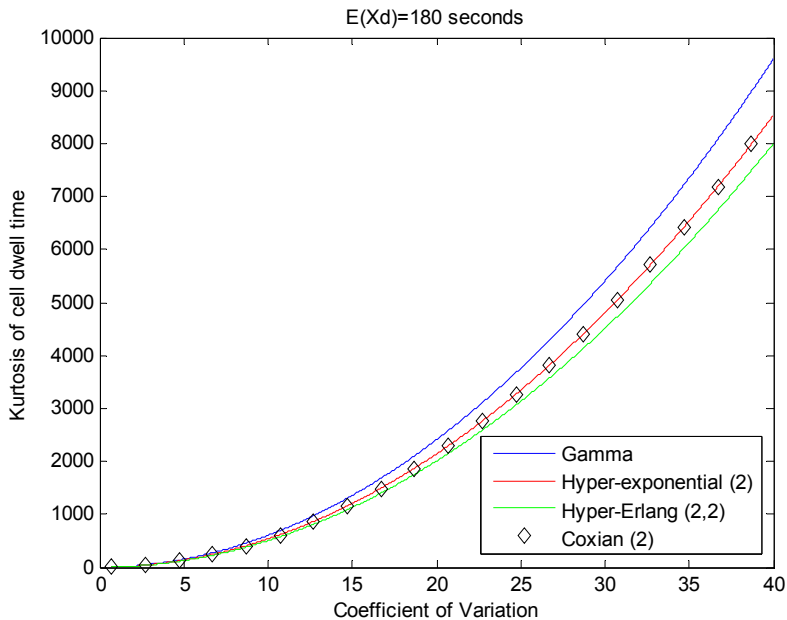


Fig. 12. Kurtosis of cell dwell time for gamma, hyper-exponential (2), Coxian (2) and hyper-Erlang (2,2) distributed cell dwell time versus CoV of cell dwell time.

5. Conclusions

In this Chapter, under the assumption that unencumbered service time is exponentially distributed, a set of novel general-algebraic equations that examines the relationships between cell dwell time and residual cell dwell time as well as between cell dwell time and new and handoff channel holding times was derived. This work includes relevant new analytical results and insights into the dependence of channel holding time characteristics on the cell dwell time probability distribution. For instance, we found that when cell dwell time is Coxian or hyper-exponentially distributed, channel holding times are also Coxian or hyper-exponentially distributed, respectively. Also, our analytical results showed that the mean and coefficient of variation of the new and handoff call channel times depend on Laplace transform and first derivative of the Laplace transform of the probability density function of cell dwell time evaluated at the inverse of the mean unencumbered service time as well as on the mean of both cell dwell time and unencumbered service time. Additionally, we derive the condition upon which the mean new call channel holding time is greater than the mean handoff call channel holding time. Similarly, the condition upon which the mean residual cell dwell time is greater than the mean cell dwell time was also derived. To the best authors' knowledge, this phenomenon that may seem to be counterintuitive has been explained and mathematically formulated in this Chapter. We believe that the study presented here is important for planning, designing, dimensioning, and optimizing of mobile cellular networks.

6. References

- Alfa A.S. and Li W., "A homogeneous PCS network with Markov call arrival process and phase type cell dwell time," *Wirel. Net.*, vol. 8, no. 6, pp. 597-605, 2002.
- Christensen T.K., Nielsen B.F., and Iversen V.B., "Phase-type models of channel-holding times in cellular communication systems," *IEEE Trans. Veh. Technol.*, vol. 53, no. 3, pp. 725-733, May 2004.
- Corral-Ruiz A.L.E., Cruz-Pérez F.A., and Hernández-Valdez G., a, "Channel holding time in mobile cellular networks with generalized Coxian distributed cell dwell time," *IEEE PIMRC'2010*, Istanbul, Turkey, Sep. 2010.
- Corral-Ruiz A.L.E., Rico-Páez Andrés, Cruz-Pérez F.A., and Hernández-Valdez G., b, "On the Functional Relationship between Channel Holding Time and Cell Dwell Time in Mobile Cellular Networks," *IEEE GLOBECOM'2010*, Miami, Florida, USA, Dec. 2010.
- Fang Y., "Hyper-Erlang distribution model and its application in wireless mobile networks," *Wirel. Networks*, vol. 7, no. 3, pp. 211-219, May. 2001.
- Fang Y., a, "Performance evaluation of wireless cellular networks under more realistic assumptions," *Wirel. Commun. Mob. Comp.*, vol. 5, no. 8, pp. 867-885, Dec. 2005.
- Fang Y., b, "Modeling and performance analysis for wireless mobile networks: a new analytical approach," *IEEE Trans. Networking*, vol. 13, no. 5, pp. 989-1002, Oct. 2005.
- Fang Y. and Chlamtac I., a, "Teletraffic analysis and mobility modeling of PCS networks," *IEEE Trans. Commun.*, vol 47, no. 7, pp. 1062-1072, July 1999.

- Fang Y., Chlamtac I., and Lin Y.-B., b, "Channel occupancy times and handoff rate for mobile computing and PCS networks," *IEEE Trans. Computers*, vol 47, no. 6, pp. 679-692, 1999.
- Fang Y., Chlamtac I., and Lin Y.-B., a, "Modeling PCS networks under general call holding time and cell residence time distributions," *IEEE/ACM Trans. Networking*, vol. 5, no. 6, pp. 893-906, Dec. 1997.
- Fang Y., Chlamtac I., and Lin Y.-B., b, "Call performance for a PCS network," *IEEE J. Select. Areas Commun.*, vol. 15, no. 8, pp. 1568-1581, Oct. 1997.
- Gross D. and Juttijudata M., "Sensitivity of output performance measures to input distributions in queueing," in *Proc. Winter Simulation Conference (WSC'97)*, Atlanta, GA, Dec. 1997.
- Hidaka H., Saitoh K., Shinagawa N., and Kobayashi T., "Self similarity in cell dwell time caused by terminal motion and its effects on teletraffic of cellular communication networks," *IEICE Trans. Fund.*, vol. E85-A, no. 7, pp. 1445-1453, 2002.
- Hong D. and Rappaport S. S., "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans. Veh. Technol.*, vol. 35, no. 3, pp. 77-92, Aug. 1986.
- Johnson M.A. and Taaffe M.R.. "Matching moments to phase distributions: mixture of Erlang distributions of common order," *Stochastic Models*, vol. 5, no. 4, pp. 711-743, 1989.
- Khan F. and Zeghlache D., "Effect of cell residence time distribution on the performance of cellular mobile networks," *Proc. IEEE VTC'97*, Phoenix, AZ, May 1997, pp. 949-953.
- Kim K. and Choi H., "A mobility model and performance analysis in wireless cellular network with general distribution and multi-cell model," *Wirel. Pers. Commun.*, published on line: 10 March 2009.
- Kleinrock L. *Queueing Systems*. John Wiley and Sons: New York, NY, 1975.
- Lin Y.-B., Mohan S. and Noerpel A., "Queueing priority channel assignment strategies for PCS and handoff initial access," *IEEE Trans. Veh. Technol.*, vol. 43. no. 3, pp. 704-712, Aug. 1994.
- Orlik P.V. and Rappaport S.S., a, "A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions," *IEEE J. Select. Areas Commun.*, vol. 16, no. 5, pp. 788-803, June 1998.
- Orlik P.V. and Rappaport S.S., b, "Traffic performance and mobility modeling of cellular communications with mixed platforms and highly variable mobility," *Proc. of the IEEE*, vol. 86, no. 7, pp. 1464-1479, July 1998.
- Pattaramalai S., Aalo V.A., and Efthymoglou G.P., "Evaluation of call performance in cellular networks with generalized cell dwell time and call-holding time distributions in the presence of channel fading," *IEEE Trans. Veh. Technol.*, vol. 58, no. 6, pp. 3002-3013, July 2009.
- Pattaramalai S., Aalo V.A., and Efthymoglou G.P., "Call completion probability with Weibull distributed call holding time and cell dwell time," in *Proc. IEEE Globecom'2007*, Washington, DC, Nov. 2007, pp. 2634-2638.
- Perros H. and Khoshgoftaar T., "Approximating general distributions by a uniform Coxian distribution," in *Proc. 20th Annual Pittsburgh Conf. on Modeling and Simulation*, Instrument Society of America, 1989, 325-333.

- Rahman M.M. and Alfa A.S., "Computationally efficient method for analyzing guard channel schemes," *Telecomm. Systems*, vol. 41, pp. 1-11, published on line: 22 April 2009.
- Sinclair B., "Coxian Distributions," *Connexions* module: m10854. Available on line: <http://cnx.org/content/m10854/latest/>
- Soong B.H. and Barria J.A., "A Coxian model for channel holding time distribution for teletraffic mobility modeling," *IEEE Commun. Letters*, vol. 4, no. 12, pp. 402-404, Dec. 2000.
- Telek M. and Heindl A., "Matching moments for acyclic discrete and continuous phase-type distributions of second order," *International Journal of Simulation*, vol. 3, no. 3-4, pp. 47-57, 2003.
- Thajchayapong S. and Tonguz O. K., "Performance implications of Pareto-distributed cell residual time in distributed admission control scheme (DACs)," in *Proc. IEEE WCNC'05*, New Orleans, LA, Mar. 2005, pp 2387-2392.
- Wang X. and Fan P., "Channel holding time in wireless cellular communications with general distributed session time and dwell time," *IEEE Commun. Letters*, vol. 11, no. 2, Feb. 2007.
- Yeo K. and Jun C.-H., "Teletraffic analysis of cellular communication systems with general mobility based on hyper-Erlang characterization," *Computer & Industrial Engineering*, vol. 42, pp. 507-520, 2002.
- Zeng H., Fang Y., and Chlamtac I., "Call blocking performance study for PCS Networks under more realistic mobility assumptions," *Telecomm. Systems*, vol. 19, no. 2, pp. 125-146, 2002.

Part 5

Next Generation Wireless Communication Technologies

Automatic Modulation Classification for Adaptive Wireless OFDM Systems

Lars Häring
*Department of Communication Systems, University of Duisburg-Essen
Germany*

1. Introduction

The flexible adaption of the transmission scheme to the current channel state becomes more and more a key issue in future communication systems. One efficient solution in multicarrier systems like Orthogonal Frequency Division Multiplexing (OFDM) has been proven to be adaptive modulation (AM) where the modulation scheme is selected on a subcarrier-basis or group of subcarriers. A lot of research has been carried out on AM or bit loading algorithms (Campello, 1998; Chow et al., 1995; Czylik, 1996; Fischer & Huber, 1996; Hughes-Hartogs, 1987).

A basic disadvantage, however, is that the receiver requires the knowledge about the selection of the modulation schemes to decode the transmitted data. The conventional measure is to transmit the so-called bit allocation table (BAT) via a signaling channel.¹ This leads to a considerable reduction of the effective data rate. In contrast to wired communication links like the digital subscriber line (DSL) in which AM is already well-established, the time-variance of mobile radio channels usually necessitates a continuous and fast update of the BAT. Even sophisticated signaling schemes using state-dependent source coding of signaling bits reduce the throughput by 3 – 4% for short packets (Chen et al., 2009). If the channel statistics are not known, the signaling overhead is significantly larger.

In order to lower the amount of the signaling overhead and to obtain more flexibility, the BAT can be automatically detected at the receiver side. Such automatic modulation classification (AMC) algorithms have already been explored intensively since several decades, primarily for military applications but not for civil radio communication systems. The classifiers can be categorized into two types: likelihood-based (Boiteau & Martret, 1998; Long et al., 1994; Polydoros & Kim, 1990; Sills, 1999; Wei & Mendel, 2000) and feature-based methods (Dobre et al., 2004; Hsue & Soliman, 1989; Nandi & Azzouz, 1998; Swami & Sadler, 2000). While likelihood-based approaches arise from a defined optimality criterion, feature-based methods are usually heuristically motivated using e.g. higher-order moments. On the other hand, likelihood algorithms tend to require a higher complexity. A comprehensive overview of existing AMC algorithms is given in (Dobre et al., 2007).

In this book chapter, we will highlight the classification of digital quadrature amplitude modulation (QAM) schemes in wireless adaptive OFDM systems using the likelihood principle (Edinger et al., 2007; Huang et al., 2007; Lampe, 2004). We particularly focus on

¹ The BAT contains the information about the modulation schemes on each subcarrier.

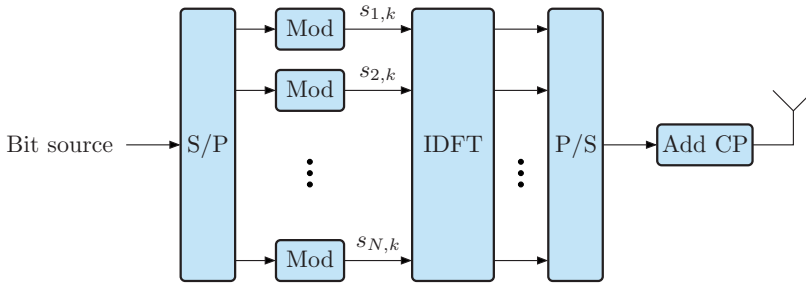


Fig. 1. Block diagram of an OFDM transmitter

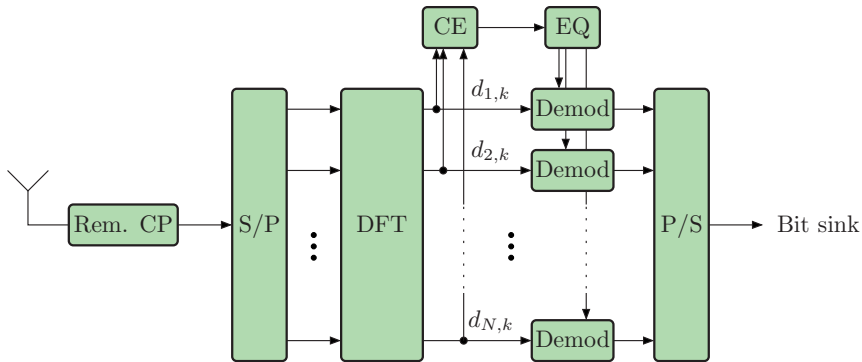


Fig. 2. Block diagram of an OFDM receiver

time-division duplex (TDD) systems in which the channel can be regarded as reciprocal. In contrast to other research work, a lot of new constraints are taken into account. Namely, many parameters are known by the receiver that can be utilized to enhance the classification reliability (Håring et al., 2010a; Håring et al., 2010b; 2011).

2. System model

2.1 Signal model

In Fig. 1 and 2, the baseband models of an OFDM transmitter and receiver are depicted. In OFDM, information data are transmitted blockwise. A sequence of bits is split into blocks, fed to different subcarriers and modulated. For the k -th block, an inverse discrete Fourier transform (IDFT) of length N on the symbols of all carriers is carried out. Subsequently, in order to combat interblock interference, a cyclic prefix of sufficient length N_g is preceded before transmission via the frequency-selective radio channel.

At the receiver side, the cyclic prefix is removed. In order to decode in OFDM, a discrete Fourier transform (DFT) is carried out. In a perfectly synchronized OFDM system, the received symbol $d_{n,k}$ on the n -th subcarrier ($1 \leq n \leq N$) of the k -th OFDM block ($1 \leq k \leq K$)

can be modeled by

$$d_{n,k} = H_n \cdot s_{n,k} + v_{n,k} , \tag{1}$$

where $s_{n,k}$ and H_n denote the transmitted data symbol and the transfer function value on the n -th subcarrier of the k -th OFDM block, respectively. We consider a propagation scenario with slowly time-variant channels, typical for indoor communications. Thus the channel transfer function does not change significantly during one transmission frame, i. e. it holds: $H_{n,k} = H_n$. The additive white noise exhibits a complex Gaussian distribution: $v_{n,k} \sim \mathcal{CN}(0, \sigma_v^2)$. Due to the multicarrier principle, low-data rate signals are transmitted via flat-fading subchannels. This enables a simple frequency domain channel estimation (CE) and equalization (EQ) shown in Fig. 2.

In OFDM systems using adaptive modulation, symbols on different subcarriers can emanate from different symbol alphabets. Without loss of generality, we restrict ourselves to the digital modulation schemes with maximum bandwidth efficiencies 6 bit/symbol according to Table 1. In Fig. 3, the respective signal constellations are depicted.

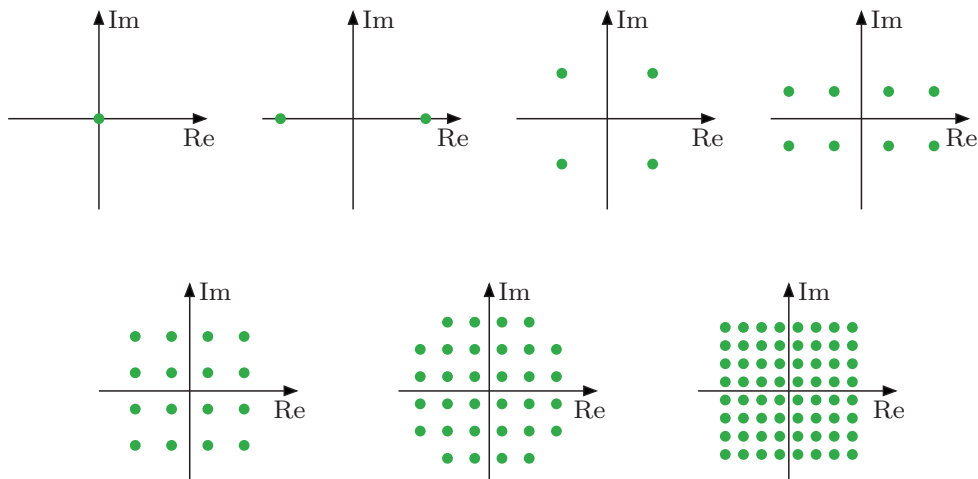


Fig. 3. QAM signal constellations: no modulation, BPSK, 4QAM, 8QAM, 16QAM, 32QAM, 64QAM

bandwidth efficiency [bit/symbol]	0	1	2	3	4	5	6
modulation type	null	BPSK	4QAM	8QAM	16QAM	32QAM	64QAM

Table 1. Considered digital modulation types

2.2 Adaptive modulation

Due to the frequency-selective nature of the radio propagation channel, some subcarriers exhibit good channel conditions whereas others suffer from a low signal-to-noise power ratio

(SNR). The overall system performance in terms of the raw bit-error ratio is dominated by the poor subcarriers.

The idea of adaptive modulation is to distribute the total amount of data bits among all subcarriers in an optimal way. If the subcarrier SNR is high, more bits than the average are loaded and higher-order modulation schemes are used. If the subcarrier SNR is low, less or even no bits are loaded such that the bit-error ratios on different subcarriers are evened out.

Using this principle, either the average bit-error ratio can be decreased at the same data rate or the data rate can be increased at the same target bit-error ratio. Since the knowledge about the data rate turns out to be an important feature of the AMC, the first approach with a fixed data rate is investigated here.

A huge amount of research on adaptive modulation algorithms has been carried out during the last twenty years (Campello, 1998; Chow et al., 1995; Czylik, 1996; Fischer & Huber, 1996; Hughes-Hartogs, 1987). In the following, we focus on algorithms that utilize the bit metric:

$$b_n = \log_2 \left(1 + \frac{\gamma_n}{k \cdot \gamma} \right) \quad \text{s.t.} \quad \sum_{n=1}^N b_n = N_b, \quad (2)$$

where γ and γ_n denote the average signal-to-noise power ratio and the SNR of the n -th subcarrier. This bit metric b_n is motivated by the channel capacity formula which takes the SNR gap (Starr et al., 1999) into account. As an example of adaptive modulation, the magnitude of the channel transfer function $|H_n|$ in a typical indoor propagation scenario (dashed line) and the corresponding bandwidth efficiencies (solid line) are shown in Fig. 4. There are two challenges involved in the application of AM in practical systems:

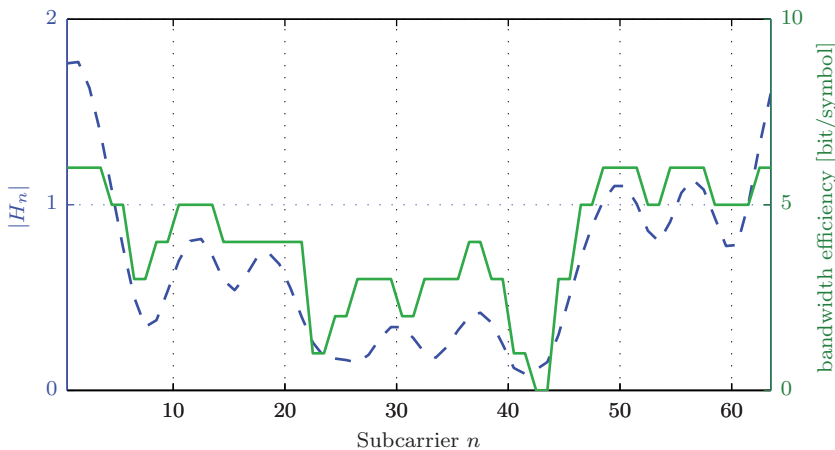


Fig. 4. Example of adaptive modulation

- *Channel knowledge at transmitter side*

In order to be able to apply AM, the transmitter must know about the subcarrier SNRs. There are two ways to obtain this knowledge: 1) via feedback from the receiver or 2) using reciprocity in time-division duplex systems. Here, the focus is on TDD systems.

In our analysis, the channel transfer factors H_n are therefore obtained by a preamble-based channel estimation in the receive mode.

- *BAT knowledge at receiver side*

In order to be able to decode the transmitted information, the receiver must know about the bit allocation table which includes the assignment of modulation schemes to subcarriers. Either this information is transmitted via a signaling channel or it is automatically classified.

2.3 Problem formulation

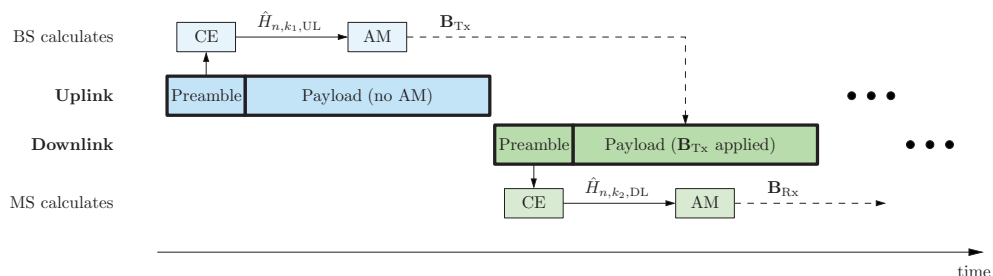


Fig. 5. TDD system structure

Fig. 5 shows the signal flow at the initiation of adaptive modulation in downlink transmission of the considered time-division duplex system model:

- 1a) In the uplink (UL), the mobile station (MS) transmits a frame consisting of training information and payload. When establishing the link, the MS does not know the channel state. Hence, for the first frame transmitted, no AM can be applied.
- b) Based on the received preamble, the base station (BS) estimates the channel $\hat{H}_{n,k_1,UL}$ and calculates the optimal bit allocation table B_{TX} using an AM algorithm.
- 2a) In the downlink (DL), the BS transmits a frame composed of training symbols and payload according to B_{TX} .
- b) Based on the received preamble, the MS estimates the channel $\hat{H}_{n,k_2,DL}$ and calculates the optimal bit allocation table B_{RX} using the same AM algorithm as the BS.
- ...

In order to decode the payload that has been sent by the base station, the mobile station requires the knowledge about B_{TX} . Since we assume that the BAT information is not signaled, the receiver must *automatically* classify the modulation schemes on each subcarrier solely based upon the received signal. In this analysis it is shown that utilization of side information that is typically available in wireless communication systems can significantly boost the classification reliability. More specifically, the AMC algorithms can exploit:

- channel correlation in frequency direction (e. g. subcarrier grouping)
- channel correlation in time direction (fixed modulation order on subcarriers and/or subgroups for entire frame)

- channel reciprocity in TDD mode
- knowledge about overall data rate (total number of loaded bits)

3. Automatic modulation classifier

In this section, automatic modulation classifiers that are based on different levels of knowledge are introduced. Denote the group of M possible digital QAM schemes by $(1 \leq m \leq M)$

$$\mathcal{I}^{(m)} = \{S_1^{(m)}, S_2^{(m)}, \dots, S_{L^{(m)}}^{(m)}\}, \quad (3)$$

where $L^{(m)}$ is the constellation size (number of constellation points) of the m -th modulation scheme and $S_i^{(m)}$ denote the complex constellation symbols.

After collecting the received symbols $\mathbf{d}_n^T = [d_{n,1}, d_{n,2}, \dots, d_{n,K}]$, where K is the data frame length, the following M hypotheses are tested for each subcarrier n :

$$\mathcal{H}_n^{(m)} \triangleq \text{the used modulation scheme of the received data symbols } \mathbf{d}_n \text{ was } \mathcal{I}^{(m)}.$$

Based upon these symbols, the underlying modulation scheme is to be detected.

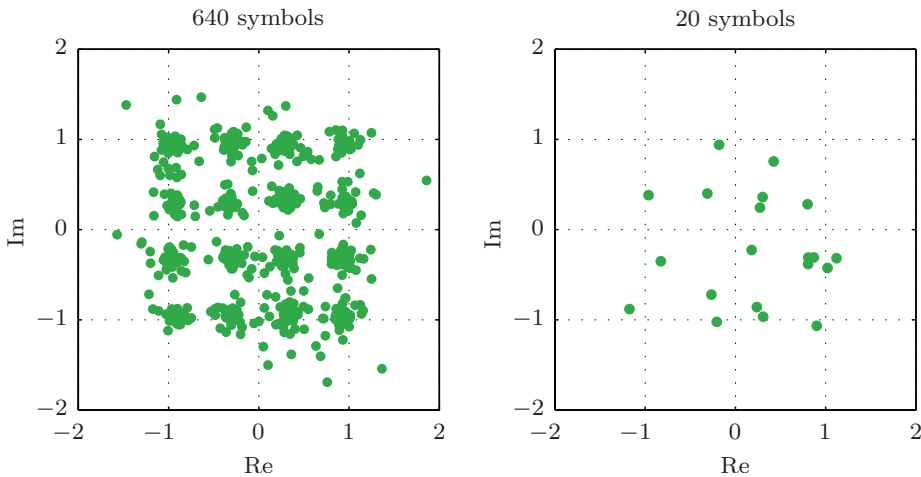


Fig. 6. Constellation diagrams of equalized 16QAM symbols at an average SNR of $\gamma = 20$ dB

As a motivating example, random signal constellations of received (and equalized) 16QAM symbols $d_{n,k}/H_n$ at an average SNR of $\gamma = 20$ dB are shown in Fig. 6. Whereas it seems obvious that the symbols emanate from a 16QAM scheme if a large number of symbols is available, it becomes much more difficult to classify the underlying modulation scheme if only a small number of symbols is available. Hence, the exploitation of additional information turns out to be a key aspect for a robust and reliable classification.

3.1 Maximum-likelihood (ML)

The maximum-likelihood (ML) approach chooses the hypothesis whose likelihood-function is maximum:

$$\hat{\mathcal{H}}_{n,\text{ML}} = \arg \max_{\mathcal{H}_n^{(m)}} p(\mathbf{d}_n | \mathcal{H}_n^{(m)}). \quad (4)$$

The probability density function of the received symbols under the condition that the m -th modulation scheme was used (hypothesis $\mathcal{H}_n^{(m)}$) is

$$p(d_{n,k} | \mathcal{H}_n^{(m)}) = \sum_{i=1}^{L^{(m)}} p(d_{n,k} | S_i^{(m)}) \cdot p(S_i^{(m)} | \mathcal{I}^{(m)}). \quad (5)$$

Each symbol within its constellation is equiprobable, i. e. $p(S_i^{(m)} | \mathcal{I}^{(m)}) = \frac{1}{L^{(m)}}$. Since v_n is assumed to be Gaussian distributed, it holds:

$$p(d_{n,k} | \mathcal{H}_n^{(m)}) = \frac{1}{L^{(m)}} \sum_{i=1}^{L^{(m)}} \frac{1}{\pi \sigma_v^2} \cdot \exp\left(-\frac{|d_{n,k} - H_n S_i^{(m)}|^2}{\sigma_v^2}\right). \quad (6)$$

If symbols of different OFDM blocks $1 \leq k \leq K$ are statistically independent, the joint probability density function $p(\mathbf{d}_n | \mathcal{H}_n^{(m)})$ is given by:

$$p(\mathbf{d}_n | \mathcal{H}_n^{(m)}) = \prod_{k=1}^K p(d_{n,k} | \mathcal{H}_n^{(m)}). \quad (7)$$

The log-likelihood function is:

$$\ln p(\mathbf{d}_n | \mathcal{H}_n^{(m)}) = \sum_{k=1}^K \ln p(d_{n,k} | \mathcal{H}_n^{(m)}) \quad (8)$$

$$= -K \cdot \ln L^{(m)} + c + \sum_{k=1}^K \ln \left(\sum_{i=1}^{L^{(m)}} \exp\left(-\gamma \cdot |d_{n,k} - H_n S_i^{(m)}|^2\right) \right) \quad (9)$$

with the average SNR $\gamma = E\{|S_i^{(m)}|^2\} / \sigma_v^2$ (= average symbol energy to noise spectral density E_S / N_0), $E_S = E\{|S_i^{(m)}|^2\} = 1$ and $E\{|H_n|^2\} = 1$. Neglecting irrelevant terms for the maximization, the ML-based classifier can be formulated as:

$$\hat{\mathcal{H}}_{n,\text{ML}} = \arg \max_{\mathcal{H}_n^{(m)}} J_{\text{ML}}(\mathbf{d}_n, \mathcal{H}_n^{(m)}) \quad \text{with}$$

$$J_{\text{ML}}(\mathbf{d}_n, \mathcal{H}_n^{(m)}) = \sum_{k=1}^K \ln \left(\sum_{i=1}^{L^{(m)}} \exp\left(-\gamma \cdot |d_{n,k} - H_n S_i^{(m)}|^2\right) \right) - K \cdot \ln L^{(m)}. \quad (10)$$

An example for the probability of correct classifications is given in Fig. 7. At the transmitter, the bandwidth efficiencies from 0 to 6 bit/symbol have been loaded by the AM algorithm in (Chow et al., 1995).

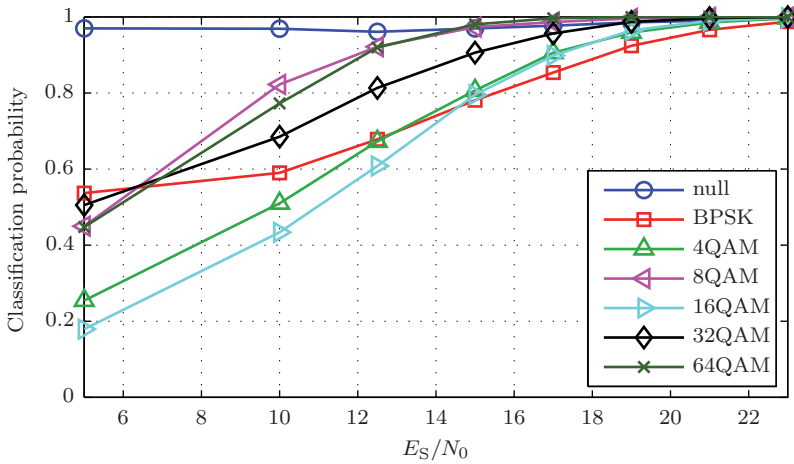


Fig. 7. Classification probability of the ML algorithm versus E_S/N_0 , frame length $K = 10$

3.2 Maximum-a-posteriori (MAP)

One major drawback of the ML algorithm is that it cannot take into account that the hypotheses under test are not equally likely. Depending on the current channel status, however, some modulation schemes will be used more frequently than others. The ML algorithm is therefore not suitable for OFDM-based systems with adaptive modulation.

A first step to improve the performance is to maximize the a-posteriori probability $p(\mathcal{H}_n^{(m)}|\mathbf{d}_n)$ instead of $p(\mathbf{d}_n|\mathcal{H}_n^{(m)})$. The main difference can be seen by using the Bayes theorem:

$$p(\mathcal{H}_n^{(m)}|\mathbf{d}_n) = \frac{p(\mathbf{d}_n|\mathcal{H}_n^{(m)}) \cdot p(\mathcal{H}_n^{(m)})}{p(\mathbf{d}_n)}. \quad (11)$$

Since $p(\mathbf{d}_n)$ is irrelevant for the maximization of $p(\mathcal{H}_n^{(m)}|\mathbf{d}_n)$, the MAP classifier can be formulated as:

$$\hat{\mathcal{H}}_{n,\text{MAP}} = \arg \max_{\mathcal{H}_n^{(m)}} J_{\text{MAP}}(\mathbf{d}_n, \mathcal{H}_n^{(m)}) \quad \text{with} \quad (12)$$

$$J_{\text{MAP}}(\mathbf{d}_n, \mathcal{H}_n^{(m)}) = \sum_{k=1}^K \ln \left(\sum_{i=1}^{L^{(m)}} \exp(-\gamma \cdot |d_{n,k} - H_n S_i^{(m)}|^2) \right) - K \cdot \ln L^{(m)} + \ln p(\mathcal{H}_n^{(m)}).$$

The a-priori information about the occurrence probabilities $p(\mathcal{H}_n^{(m)})$ is utilized.² But still, the MAP algorithm in its current form is not able to sufficiently consider the specific characteristics of adaptive modulation.

² If these probabilities are equal, i.e. it holds: $p(\mathcal{H}_n^{(m)}) = 1/M$, then the MAP approach reduces to the ML algorithm.

3.3 MAP algorithms exploiting channel reciprocity (MAP-R)

One key feature to increase the classification reliability is the utilization of channel reciprocity in TDD systems. Under ideal conditions (perfect channel reciprocity, channel time-invariance and channel state information (CSI)), the receiver can perfectly reconstruct the transmit BAT by applying the same AM algorithm based on the propagation channel in the receive direction. In that case, the bit allocation table \mathbf{B}_{Rx} equals \mathbf{B}_{Tx} . To be more realistic, we assume that the channel is time-variant and CSI is taken from a data-aided channel estimation. This causes \mathbf{B}_{Tx} and \mathbf{B}_{Rx} to be different but still correlated.

Fig. 8 shows an illustrating example of the magnitudes of the channel transfer function at transmitter and receiver side and the corresponding BATs \mathbf{B}_{Tx} and \mathbf{B}_{Rx} at an SNR of $\gamma = 10$ dB, a frame duration of $T_{\text{fr}} = 0.1$ ms and a Doppler frequency of $f_{\text{dop}} = 15$ Hz. A typical preamble-based zero-forcing method to estimate the channels has been applied.

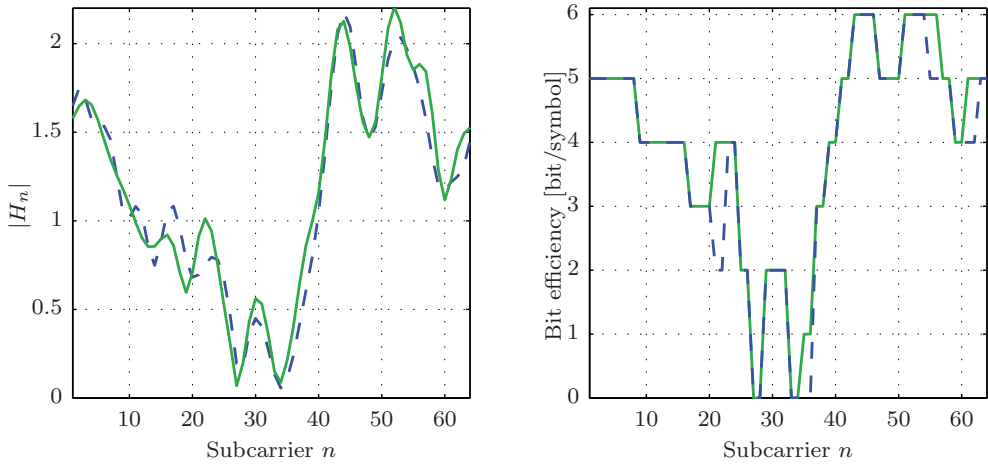


Fig. 8. Example channel transfer functions (left) and corresponding BATs (right) at transmitter (dashed lines) and receiver side (solid lines)

Both the channel transfer functions as well as the BATs differ for the transmitter and receiver, respectively. However, they are very similar which an AMC algorithm can take advantage of.

Two different concepts to benefit from this channel reciprocity are discussed now.

3.3.1 Receive bit allocation table (MAP-RQ)

Let $\hat{\mathcal{H}}_{n,\text{Rx}}$ be the modulation scheme for the n -th subcarrier in \mathbf{B}_{Rx} which was computed by the AM algorithm. The method in (11) is extended by the knowledge about $\hat{\mathcal{H}}_{n,\text{Rx}}$. The modified approach which exploits channel reciprocity in terms of the quantized information $\hat{\mathcal{H}}_{n,\text{Rx}}$ can be written as:

$$\hat{\mathcal{H}}_{n,\text{MAP-RQ}} = \arg \max_{\mathcal{H}_n^{(m)}} p(\mathcal{H}_n^{(m)} | \mathbf{d}_n, \hat{\mathcal{H}}_{n,\text{Rx}}). \quad (13)$$

By applying the Bayes theorem and then neglecting irrelevant terms, we obtain:

$$\hat{\mathcal{H}}_{n,\text{MAP-RQ}} = \arg \max_{\mathcal{H}_n^{(m)}} \frac{p(\mathbf{d}_n, \hat{\mathcal{H}}_{n,\text{Rx}} | \mathcal{H}_n^{(m)}) p(\mathcal{H}_n^{(m)})}{p(\mathbf{d}_n, \hat{\mathcal{H}}_{n,\text{Rx}})} \quad (14)$$

$$= \arg \max_{\mathcal{H}_n^{(m)}} p(\mathbf{d}_n, \hat{\mathcal{H}}_{n,\text{Rx}} | \mathcal{H}_n^{(m)}) p(\mathcal{H}_n^{(m)}) . \quad (15)$$

Since $p((A, B) | C) = p(A | (B, C)) \cdot p(B | C)$, it holds:

$$p((\mathbf{d}_n, \hat{\mathcal{H}}_{n,\text{Rx}}) | \mathcal{H}_n^{(m)}) = p(\mathbf{d}_n | (\hat{\mathcal{H}}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})) p(\hat{\mathcal{H}}_{n,\text{Rx}} | \mathcal{H}_n^{(m)}) . \quad (16)$$

The simplification $p(\mathbf{d}_n | (\hat{\mathcal{H}}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})) \approx p(\mathbf{d}_n | \mathcal{H}_n^{(m)})$ turns out to be reasonable for increasing correlation between \mathbf{B}_{Tx} and \mathbf{B}_{Rx} . We set:

$$\hat{\mathcal{H}}_{n,\text{MAP-RQ}} = \arg \max_{\mathcal{H}_n^{(m)}} \left\{ p(\mathbf{d}_n | \mathcal{H}_n^{(m)}) p(\hat{\mathcal{H}}_{n,\text{Rx}} | \mathcal{H}_n^{(m)}) p(\mathcal{H}_n^{(m)}) \right\} . \quad (17)$$

Taking the logarithm and using $p(\hat{\mathcal{H}}_{n,\text{Rx}}, \mathcal{H}_n^{(m)}) = p(\hat{\mathcal{H}}_{n,\text{Rx}} | \mathcal{H}_n^{(m)}) p(\mathcal{H}_n^{(m)})$, the MAP classifier based on the receive BAT is:

$$\begin{aligned} \hat{\mathcal{H}}_{n,\text{MAP-RQ}} &= \arg \max_{\mathcal{H}_n^{(m)}} J_{\text{MAP-RQ}}(\mathbf{d}_n, \mathcal{H}_n^{(m)}) \quad \text{with} \\ J_{\text{MAP-RQ}}(\mathbf{d}_n, \mathcal{H}_n^{(m)}) &= \sum_{k=1}^K \ln \left(\sum_{i=1}^{L^{(m)}} \exp \left(-\gamma \cdot |d_{n,k} - H_n S_i^{(m)}|^2 \right) \right) - K \cdot \ln L^{(m)} \\ &\quad + \ln p(\hat{\mathcal{H}}_{n,\text{Rx}}, \mathcal{H}_n^{(m)}) . \end{aligned} \quad (18)$$

The joint probability $p(\hat{\mathcal{H}}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$ could be numerically determined in advance and stored in look-up tables. Simulation results have shown that a coarse knowledge of these values is sufficient to achieve a significant performance improvement compared to the conventional ML algorithm. However, the probabilities $p(\hat{\mathcal{H}}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$ are usually not available in practice and, moreover, strongly depend on the transmission system and the propagation scenario. In order to overcome this disadvantage we present a heuristic approach to obtain these values. The classification performance is similar to the optimal one.

Approximation of joint probabilities

Suppose that $\hat{\mathcal{H}}_{n,\text{Rx}} = \mathcal{I}^{(\mu)}$ is the modulation scheme for the n -th subcarrier in \mathbf{B}_{Rx} . Due to the correlation between transmit and receive BAT, $\mathcal{I}^{(\mu)}$ is said to be more likely than the other $M - 1$ possible modulation schemes. Consequently, we set

$$p(\hat{\mathcal{H}}_{n,\text{Rx}} = \mathcal{I}^{(\mu)}, \mathcal{H}_n^{(m)}) = \alpha \cdot \begin{cases} w & , m = \mu \\ \frac{1-w}{M-1} & , m \neq \mu \end{cases} \quad (19)$$

with $\frac{1}{M} \leq w \leq 1$. The proportional factor α is irrelevant for the metric maximization. The probabilities of all other hypotheses $m \neq \mu$ have been set equally to $\frac{1-w}{M-1}$. Further numerical investigations have shown that distributing the "residual" probability $1 - w$ in a more sophisticated way does not lead to a significant advantage.

The optimal values of the weighting factors w depend on a multitude of effects: channel quality, channel estimation method, adaptive modulation algorithm etc. The more correlated

\mathbf{B}_{Tx} and \mathbf{B}_{Rx} are, the higher the value of w should be chosen. Unfortunately, the analytical search for the optimum seems to be intractable.

The simplicity of this heuristic approach appears attractive for a practical implementation. The receiver needs to calculate \mathbf{B}_{Rx} anyway for the application of adaptive modulation in the next transmission frame.

A general drawback of using the BAT in order to exploit channel reciprocity is the quantization of bandwidth efficiencies.

3.3.2 Receive channel state information (MAP-RS)

An even better way to incorporate the channel correlation in transmit and receive direction into the AMC algorithm is described now. We assume that the AM algorithm at transmitter and receiver side is based upon bit loading according to the widely used criterion (Chow et al., 1995) for the estimated bandwidth efficiency:

$$\hat{b}_{n,\text{Rx}} = \log_2 \left(1 + \frac{\gamma_n}{k \cdot \gamma} \right) = \log_2 \left(1 + \frac{|H_n|^2}{k} \right), \quad (20)$$

in which k is adapted such that the target bit rate is achieved. The classification method that utilizes the *soft* channel information $\hat{b}_{n,\text{Rx}}$ instead of the *hard* BAT information $\hat{\mathcal{H}}_{n,\text{Rx}}$ can be formulated as:

$$\hat{\mathcal{H}}_{n,\text{MAP-RS}} = \arg \max_{\mathcal{H}_n^{(m)}} \left\{ p(\mathcal{H}_n^{(m)} | \mathbf{d}_n, \hat{b}_{n,\text{Rx}}) \right\}. \quad (21)$$

Following the same steps as in the previous section, the solution of (21) is:

$$\begin{aligned} \hat{\mathcal{H}}_{n,\text{MAP-RS}} &= \arg \max_{\mathcal{H}_n^{(m)}} J_{\text{MAP-RS}}(\mathbf{d}_n, \mathcal{H}_n^{(m)}, \hat{b}_{n,\text{Rx}}) \quad \text{with} \\ J_{\text{MAP-RS}} &= \sum_{k=1}^K \ln \left(\sum_{i=1}^{L^{(m)}} \exp \left(-\gamma \cdot |d_{n,k} - H_n S_i^{(m)}|^2 \right) \right) - K \cdot \ln L^{(m)} \\ &\quad + \ln p(\hat{b}_{n,\text{Rx}}, \mathcal{H}_n^{(m)}). \end{aligned} \quad (22)$$

Whereas the AMC algorithm based on the receive BAT in section 3.3.1 requires the knowledge about $p(\hat{\mathcal{H}}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$, here the function $p(\hat{b}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$ must be known. Fig. 9 depicts simulation examples of the joint probability density function $p(\hat{b}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$ and the probabilities $p(\hat{\mathcal{H}}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$.

Approximation of joint probability density functions

It is unrealistic to assume that the joint probabilities $p(\hat{b}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$ are available in practical systems. Again, we find approximations based on a heuristic approach: Suppose that the AM algorithm at the receiver has computed $\hat{b}_{n,\text{Rx}} = b_0$ for subcarrier n . Then it is obvious that those hypotheses which are "closer" to b_0 are more likely than others. We use the heuristic measure

$$p(\hat{b}_{n,\text{Rx}} = b_0, \mathcal{H}_n^{(m)}) = \beta \cdot \exp \left(- \left(\frac{b(\mathcal{H}_n^{(m)}) - b_0}{\sqrt{2}\sigma} \right)^2 \right) \quad (23)$$

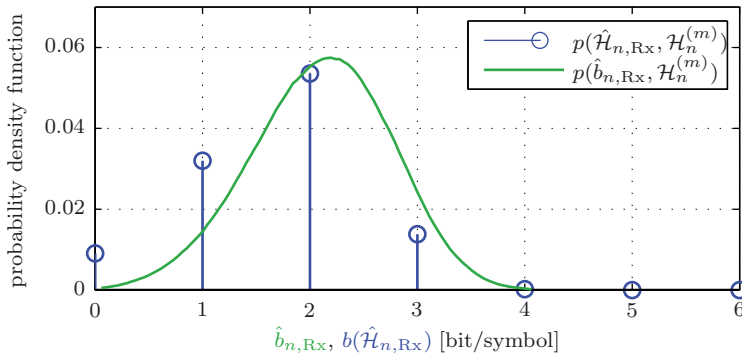


Fig. 9. Example of simulated probability density functions $p(\hat{b}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$ and $p(\hat{\mathcal{H}}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$ for $b(\mathcal{H}_n^{(m)}) = 2$ bit/symbol and $\gamma = 10$ dB; the average bandwidth efficiency is 4 bit/symbol

parameter	value
sampling period T	50 ns (20 MHz bandwidth)
FFT length N	64
cyclic prefix length N_{cp}	16
Channel model	(Medbo & Schramm, 1998) (indoor)
Delay spread τ_{ds}	100 ns
Doppler frequency f_{dop}	15 Hz (Jakes spectrum)

Table 2. Simulation set-up

thus assuming a Gaussian distributed deviation with the design parameter σ . Here, the operator $b(\mathcal{H})$ denotes the number of bits transmitted for hypothesis \mathcal{H} ; β is a proportional factor that is irrelevant for the maximization.

The more correlated the channels at transmitter and receiver are, the smaller σ should be chosen. Due to the complex influence of the AM algorithm, it seems to be intractable to find the optimal value σ analytically.

However, compared to the quantized information that is used in the MAP-RQ algorithm, the MAP-RS method benefits from the more reliable soft information leading to a higher classification performance.

3.3.3 Discussion

In Fig. 10, the classification error probability $P_{\text{ce}} = \Pr\{\hat{\mathbf{B}}_{\text{Tx}} \neq \mathbf{B}_{\text{Tx}}\}$ versus E_S/N_0 for various AMC methods is shown. If already one modulation scheme in \mathbf{B}_{Tx} is incorrectly classified, the whole packet would get lost. Throughout the entire contribution, the performance discussion is based upon a common simulation scenario: The main parameters concordant with a WLAN IEEE 802.11a/n system (IEEE, 2005) are summarized in Table 2. The bandwidth efficiencies, however, vary between 0 – 6 bit/symbol. In average, 4 bit/symbol are loaded by the AM algorithm proposed in (Chow et al., 1995). Due to the fact that the subcarrier spacing is small compared to the channel coherence bandwidth, 2 neighbouring subcarriers are grouped without sacrificing significant system performance. The AMC algorithms utilize this

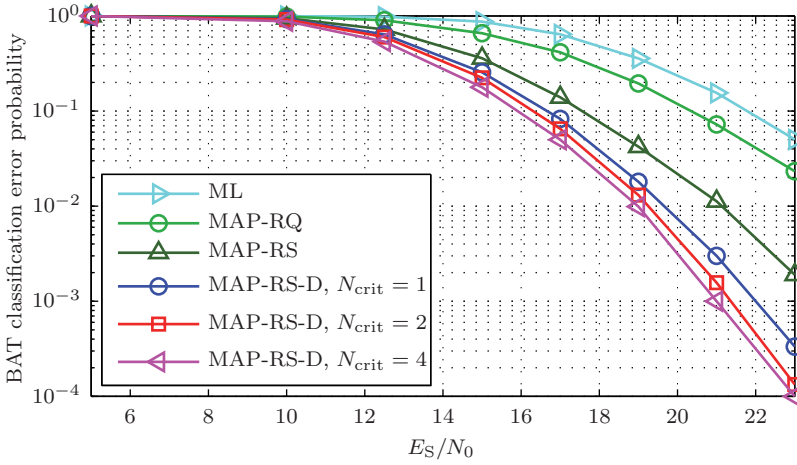


Fig. 10. BAT classification error probability P_{ce} of AMC algorithms versus E_S/N_0 , frame length $K = 10$

subcarrier grouping by additively combining the corresponding metric values. A zero-forcing channel estimation is carried out on 2 training OFDM blocks with a total length of 160 samples. It is assumed that the receiver has perfect knowledge about the total data rate since this single information can be protected against errors well by suitable channel coding. The design parameters for the reciprocity-based classifiers are set to $w = 0.8$ and $\sigma = 0.3$.

First, we refer only to the ML algorithm (section 3.1) and the reciprocity-based methods MAP-RQ and MAP-RS: Due to the additional reciprocity information used, the reciprocity-based MAP algorithms outperform the classical ML algorithm significantly. With the soft information used in MAP-RS, the number of modulation classification errors can even be further reduced compared to MAP-RQ. Another step towards a higher reliability is to include even more side information in the modulation classification.

3.4 MAP-R algorithms exploiting the knowledge about the data rate (MAP-R-D)

In communication systems, the overall data rate is typically signaled to the receiver via a control channel. Therefore, we can incorporate the available information about the total number of transmitted bits N_b per OFDM block into the MAP-R algorithms.³

As an example, we describe the extension for the algorithm MAP-RS. The corresponding modifications in MAP-RQ are analogous. The scheme that jointly classifies the bandwidth efficiencies on all subcarriers can be formulated as:

$$\begin{aligned} \hat{\mathcal{H}}_{n,\text{joint}} &= \arg \max_{\mathcal{H}_n^{(m)}} \left\{ p(\mathcal{H}_n^{(m)} | \mathbf{d}_n, \hat{b}_{n,\text{Rx}}) \right\}, \quad 1 \leq n \leq N, \\ \text{s.t. } & \sum_{n=1}^N b(\mathcal{H}_n^{(m)}) = N_b. \end{aligned} \quad (24)$$

³ Note that the BAT is fixed for the entire transmission burst.

With the abbreviations

$$\vec{\mathcal{H}} = [\mathcal{H}_1^{(m)}, \dots, \mathcal{H}_N^{(m)}] \quad (25)$$

$$\hat{\mathbf{b}}_{\text{Rx}} = [\hat{b}_{1,\text{Rx}}, \dots, \hat{b}_{N,\text{Rx}}] \quad (26)$$

$$\hat{\mathcal{H}}_{\text{joint}} = [\hat{\mathcal{H}}_{1,\text{joint}}, \dots, \hat{\mathcal{H}}_{N,\text{joint}}] \quad (27)$$

$$\vec{\mathbf{d}} = [\mathbf{d}_1, \dots, \mathbf{d}_N] \quad (28)$$

for the hypotheses under test, the bandwidth efficiencies computed at the receiver, the classified hypotheses and the collected received data symbols on all subcarriers, we reformulate (24):

$$\begin{aligned} \hat{\mathcal{H}}_{\text{joint}} &= \arg \max_{\vec{\mathcal{H}}} \left\{ p(\vec{\mathcal{H}} | \vec{\mathbf{d}}, \hat{\mathbf{b}}_{\text{Rx}}) \right\} \\ \text{s.t. } & b(\vec{\mathcal{H}}) = N_b . \end{aligned} \quad (29)$$

Since the modulation schemes and data symbols on different subcarriers are independent from each other, it holds:

$$p(\vec{\mathcal{H}} | \vec{\mathbf{d}}, \hat{\mathbf{b}}_{\text{Rx}}) \approx \prod_{n=1}^N p(\mathcal{H}_n^{(m)} | \mathbf{d}_n, \hat{b}_{n,\text{Rx}}) , \quad (30)$$

and hence:

$$\vec{\mathcal{H}}_{\text{joint}} = \arg \max_{\vec{\mathcal{H}}} J_{\text{joint}}(\vec{\mathbf{d}}, \vec{\mathcal{H}}) \quad \text{s.t. } b(\vec{\mathcal{H}}) = N_b \quad \text{with} \quad (31)$$

$$\begin{aligned} J_{\text{joint}}(\vec{\mathbf{d}}, \vec{\mathcal{H}}) &= \sum_{n=1}^N \left[\sum_{k=1}^K \ln \left(\sum_{i=1}^{L^{(m)}} \exp \left(-\gamma \cdot |d_{n,k} - H_n S_i^{(m)}|^2 \right) \right) - K \cdot \ln L^{(m)} \right. \\ &\quad \left. + \ln p(\hat{b}_{n,\text{Rx}}, \mathcal{H}_n^{(m)}) \right] . \end{aligned} \quad (32)$$

The maximum search must be carried out over a set of all possible hypothesis *candidate combinations* $\vec{\mathcal{H}}$. Due to its high complexity, such a joint search is not feasible in practice. We investigate a trade-off between this joint algorithm and the subcarrier-independent methods instead.

First, we split the set of all subcarriers $\mathcal{S} = \{1, \dots, N\}$ into two subsets:

- $\mathcal{S}_{\text{rel}} := \{n \mid \text{reliable decision}\}$ including subcarriers for which *reliable* decisions are possible, and
- $\mathcal{S}_{\text{crit}} := \{n \mid \text{critical decision}\} = \mathcal{S} \setminus \mathcal{S}_{\text{rel}}$ including subcarriers with *critical* decisions.

The number of elements in $\mathcal{S}_{\text{crit}}$ is denoted as N_{crit} . The distinction is based upon an ordering procedure according to the absolute distance between the largest and second largest metric value of the subcarrier-independent metrics (18) or (22), respectively. Subcarriers with the largest absolute distances are inserted in \mathcal{S}_{rel} ; the remaining N_{crit} subcarriers are included in $\mathcal{S}_{\text{crit}}$. In other words, decisions are said to be critical if the two largest metric values are similar.

Surprisingly, ordering by subcarrier SNR values turns out to be unfavourable. Due to adaptive modulation in the considered OFDM systems, channel gains and modulation orders

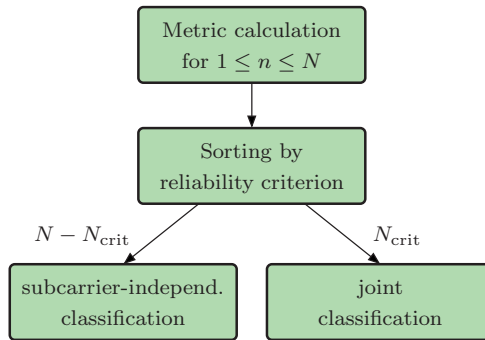


Fig. 11. Signal flow of joint algorithm

are mutually coupled – bandwidth efficiencies increase with increasing SNR. High-order modulation schemes like 64QAM are, however, more difficult to classify than low-order modulation schemes like BPSK.

The automatic modulation classification is performed for the two subsets in a different way:

- For set \mathcal{S}_{rel} , we apply the subcarrier-wise MAP-R algorithms described in section 3.3, e. g.:

$$\hat{\mathcal{H}}_{n,\text{MAP-RS-D}} = \arg \max_{\mathcal{H}_n^{(m)}} J_{\text{MAP-RS}}(\mathbf{d}_n, \mathcal{H}_n^{(m)}) \quad \forall n \in \mathcal{S}_{\text{rel}}. \quad (33)$$

- For set $\mathcal{S}_{\text{crit}}$, we apply the joint MAP algorithm following (31):

$$\vec{\hat{\mathcal{H}}}_{\text{MAP-RS-D}} = \arg \max_{\vec{\mathcal{H}}_{\mathcal{S}_{\text{crit}}}} J_{\text{joint}}(\vec{\mathbf{d}}, \vec{\mathcal{H}}_{\mathcal{S}_{\text{crit}}}) \quad \text{s.t. } b(\vec{\mathcal{H}}_{\mathcal{S}_{\text{crit}}}) = N_b - b(\vec{\mathcal{H}}_{\mathcal{S}_{\text{rel}}}) \quad (34)$$

with $\vec{\mathcal{H}}_{\mathcal{S}_{\text{crit}}}$ and $\vec{\mathcal{H}}_{\mathcal{S}_{\text{rel}}}$ denoting the vectors that contain the hypotheses of all subcarriers in set $\mathcal{S}_{\text{crit}}$ and \mathcal{S}_{rel} , respectively.

The choice of the design parameter N_{crit} balances performance and complexity.

3.4.1 Discussion

We refer to Fig. 10 again. It shows, among others, the classification reliability of the MAP-RS-D algorithm for different values of N_{crit} . A significant increase of the reliability can be seen, independent of the value of N_{crit} . For $N_{\text{crit}} > 2$, the performance improvement saturates whereas the complexity grows rapidly. In numerous cases at high SNR, only very few decisions are ambiguous. By incorporating additional information, the reliability of these vague decisions can be considerably improved. Against the background of a practical design, small values of N_{crit} are an appropriate choice.

3.5 MAP algorithm with reduced complexity (MinMAP)

The complexity of all presented MAP methods is still rather high and may be prohibitive for real-time applications. In order to reduce the complexity, we take a closer look e. g. at the

classification metric (22) illustrated in Fig. 12 which is based on the expression

$$\sum_{k=1}^K \ln \left(\sum_{i=1}^{L^{(m)}} \exp \left(-\gamma \cdot |d_{n,k} - H_n S_i^{(m)}|^2 \right) \right) - K \cdot \ln L^{(m)} + \ln p(\hat{b}_{n,Rx}, \mathcal{H}_n^{(m)}) . \quad (35)$$

Both logarithmic and exponential functions in the metric are intensive operations which are

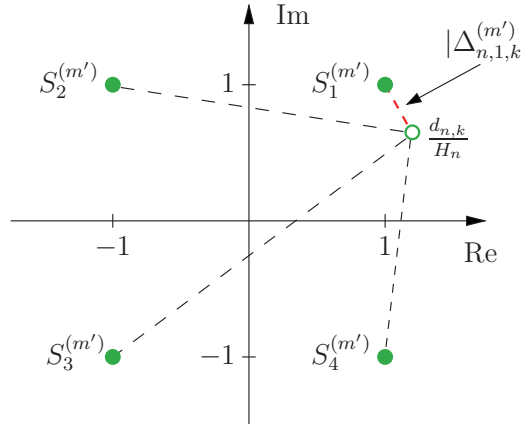


Fig. 12. Example of metric calculation for 4QAM transmission: Dominant contribution term $\exp(-\gamma_n |\Delta_{n,i,k}^{(m)}|^2)$ for fixed n and k arises from closest constellation point

not suitable for implementations in hardware structures. The metric evaluates weighted and squared distances $\Delta_{n,i,k}^{(m)}$

$$\gamma \cdot |d_{n,k} - H_n S_i^{(m)}|^2 = \gamma_n \cdot \left| \frac{d_{n,k}}{H_n} - S_i^{(m)} \right|^2 = \gamma_n \cdot |\Delta_{n,i,k}^{(m)}|^2 \quad (36)$$

between the received equalized symbol and all possible constellation symbols of the modulation scheme under test.

Let us consider the high SNR region obtaining a received equalized symbol $d_{n,k}/H_n$ located close to a possible constellation point. Thanks to the fast decrease of the exponential function for decreasing arguments, only the term $\exp(-\gamma_n \cdot \min\{|\Delta_{n,i,k}^{(m)}|^2\})$ will significantly contribute to the inner sum in (22):

$$\sum_{i=1}^{L^{(m)}} \exp \left(-\gamma_n |\Delta_{n,i,k}^{(m)}|^2 \right) \approx \max_{S_i^{(m)} \in \mathcal{I}^{(m)}} \left\{ \exp \left(-\gamma_n |\Delta_{n,i,k}^{(m)}|^2 \right) \right\} \quad (37)$$

$$= \exp \left(- \min_{S_i^{(m)} \in \mathcal{I}^{(m)}} \left\{ \gamma_n |\Delta_{n,i,k}^{(m)}|^2 \right\} \right) . \quad (38)$$

The resulting simplified metric is denoted by MinMAP as the operations $\log(\cdot)$ and $\exp(\cdot)$ are replaced by a simple minimum search. As an example, the MinMAP-RS metric is:

$$\begin{aligned}
\hat{\mathcal{H}}_{n,\text{MinMAP-RS}}^{(m)} &= \arg \max_{\mathcal{H}_n^{(m)}} J_{\text{MinMAP-RS}}(\mathbf{d}_n, \mathcal{H}_n^{(m)}) \quad \text{with} \\
J_{\text{MinMAP-RS}} &= \sum_{k=1}^K \ln \left(\exp \left(- \min \left\{ \gamma_n |\Delta_{n,i,k}^{(m)}|^2 \right\} \right) \right) - K \ln L^{(m)} + \ln p(\hat{b}_{n,\text{Rx}}, \mathcal{H}_n^{(m)}) \\
&= -\gamma_n \sum_{k=1}^K \min_{S_i^{(m)} \in \mathcal{Z}^{(m)}} \left\{ |\Delta_{n,i,k}^{(m)}|^2 \right\} - K \ln L^{(m)} + \ln p(\hat{b}_{n,\text{Rx}}, \mathcal{H}_n^{(m)}). \quad (39)
\end{aligned}$$

No exponential and almost no logarithm operations are needed in the classifier. This leads to an essentially decreased complexity. The algorithms MAP-RS-D, MAP-RQ and MAP-RQ-D are modified alike, i. e. denoted as MinMAP-RS-D, MinMAP-RQ and MinMAP-RQ-D.

Metric analysis for high SNR:

It will be shown that the error caused by the simplification (37) tends to zero (under mild conditions) if $\gamma \rightarrow \infty$. For simplicity reasons, we neglect all subcarrier, modulation and block indices. First, we analyze the sum:

$$\ln \left(\sum_{i=1}^L e^{-\gamma |\Delta_i|^2} \right) = \ln \left(e^{-\gamma |\Delta_1|^2} + e^{-\gamma |\Delta_2|^2} + \dots + e^{-\gamma |\Delta_L|^2} \right). \quad (40)$$

Here, we have sorted $|\Delta_i|$ in an ascending order with $|\Delta_1|$ being the minimum and $|\Delta_L|$ being the maximum distance. Now, let us define the error ε between the optimal MAP and the simplified MAP metrics ($K = 1$):

$$\varepsilon = \ln \left(\sum_{i=1}^L e^{-\gamma |\Delta_i|^2} \right) - \ln \left(e^{-\gamma |\Delta_1|^2} \right) \quad (41)$$

$$= \ln \left(\frac{\sum_{i=1}^L e^{-\gamma |\Delta_i|^2}}{e^{-\gamma |\Delta_1|^2}} \right) \quad (42)$$

$$= \ln \left(\sum_{i=1}^L e^{-\gamma (|\Delta_i|^2 - |\Delta_1|^2)} \right) \quad (43)$$

$$= \ln \left(1 + \sum_{i=2}^L e^{-\gamma (|\Delta_i|^2 - |\Delta_1|^2)} \right) \geq 0. \quad (44)$$

Under the condition $|\Delta_i| \neq |\Delta_1|$ for $i \neq 1$, the error in the metric function for $\gamma \rightarrow \infty$ is:

$$\lim_{\gamma \rightarrow \infty} \varepsilon = \ln(1 + 0) = 0, \quad (45)$$

since $|\Delta_i|^2 - |\Delta_1|^2 > 0 \quad \forall \quad 2 \leq i \leq L$. Numerical results have shown moderate deviations between the simplified and the optimal metrics already for practical SNR ranges.

Assume that the received equalized symbols converge to their constellation points for high SNR. Then e. g. the special case $|\Delta_2| = |\Delta_1|$ or $|\Delta_2| \approx |\Delta_1|$ that we excluded in our consideration so far can only occur if the hypothesis under test is incorrect. Since the neglect of terms in (37) lowers the metric value of this incorrect hypothesis, this condition can even have a favourable effect on the discrimination of the modulation schemes.

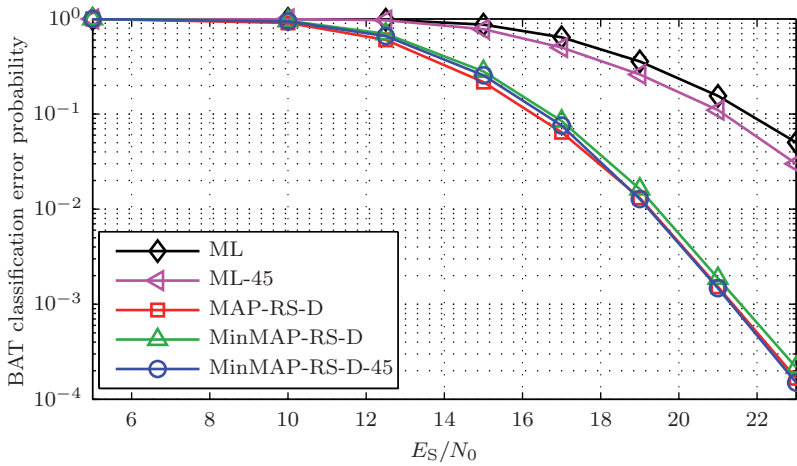


Fig. 13. BAT classification error probability P_{ce} of AMC algorithms versus E_S/N_0 , frame length $K = 10$, $N_{crit} = 2$

An overview of the presented metrics is appended at the chapter end in Table 3.

3.5.1 Discussion

Fig. 13 shows the classification performance of the algorithm MAP-RS-D with and without the previously described metric simplification. It indicates that the influence of the metric simplification on the classification performance is minor. Only a small performance degradation is visible which will even decrease with increasing SNR. Obviously, the MinMAP-RS-D approach seems to be a proper tradeoff between performance and complexity.

3.6 QAM symbol rotation

Apart from the algorithm design of the receiver, also the transmitter can be factored into the modulation classification for performance improvements. A simple example is given here: Similar to pattern recognition, automatic classification becomes effective if the objects to discriminate are *as different as possible* and can therefore be easily separated.

However, especially the constellation sets of the higher-order modulation schemes are very *similar*. A large number of received symbols must be observed to judge safely from which set the symbols have been generated. As a first measure to achieve a better separation of the constellation sets we rotate all 16QAM symbols by 45° as shown in Fig. 14 (right). Note that the optimal constellation modifications would take the rotation of all modulation schemes into account.

3.6.1 Discussion

The algorithms using the constellation modification at the transmitter side are denoted as ML-45 and MinMAP-RS-D-45, respectively. A slight performance improvement due to the

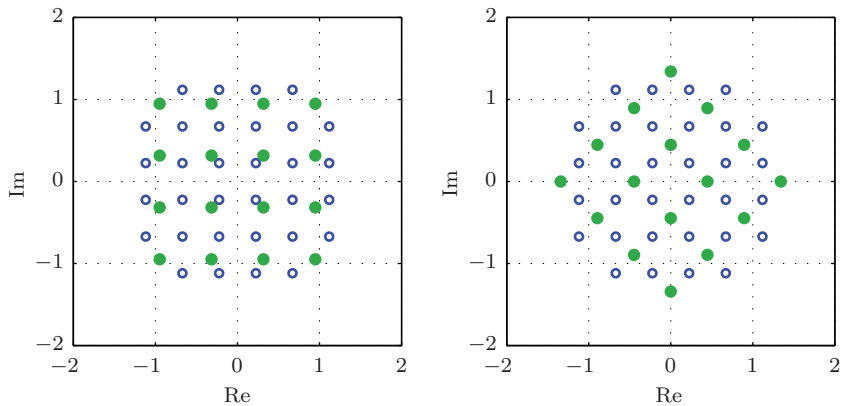


Fig. 14. Constellation diagrams of 32QAM and 16QAM (left) and 32QAM and rotated 16QAM (right)

symbol rotation by 45° can be seen in Fig. 13, especially in case of the ML algorithm. For the advanced techniques, the effect of the symbol rotation becomes less significant.

However, we expect further potential in jointly optimizing the adaptive modulation at the transmitter and the automatic modulation classification at the receiver side.

4. Overall system performance

Finally we analyze the overall system performance of a typical adaptive OFDM-based transmission system which applies AMC. We are primarily interested in the influence of errors caused by imperfect AMC on the packet error ratio (PER). A packet error is observed here if either the BAT or the payload data is detected erroneously. For these PER simulations, a hard-decision Viterbi algorithm decodes the information bits that have been encoded with a convolutional code of rate $R_C = 1/2$. Since each frame consists of 10 data OFDM blocks, the payload size amounts to $10 \text{ blocks} \cdot 64 \text{ subcarriers/block} \cdot 4 \text{ bit/subcarrier} \cdot R_C = 1280 \text{ bit} = 160 \text{ bytes}$.

Fig. 15 depicts the PER versus E_S/N_0 for the following four scenarios:

- non-adaptive: The transmitter uses the same modulation scheme on all subcarriers.
- adaptive, ML: The transmitter applies AM; the receiver detects the transmit BAT automatically using the ML algorithm.
- adaptive, MinMAP-RS-D: The transmitter applies AM; the receiver detects the transmit BAT automatically using the MinMAP-RS-D algorithm.
- adaptive, BAT known: The transmitter applies AM; the receiver has perfect knowledge of B_{Tx} (reference).

For the frequency-selective indoor propagation channel considered here, the transmission system benefits significantly from adaptive modulation; the number of packet errors is lowest in the adaptive case with perfect knowledge of B_{Tx} . Clearly, AMC degrades the system performance due to BAT classification errors. Whereas the degradation in case of the ML

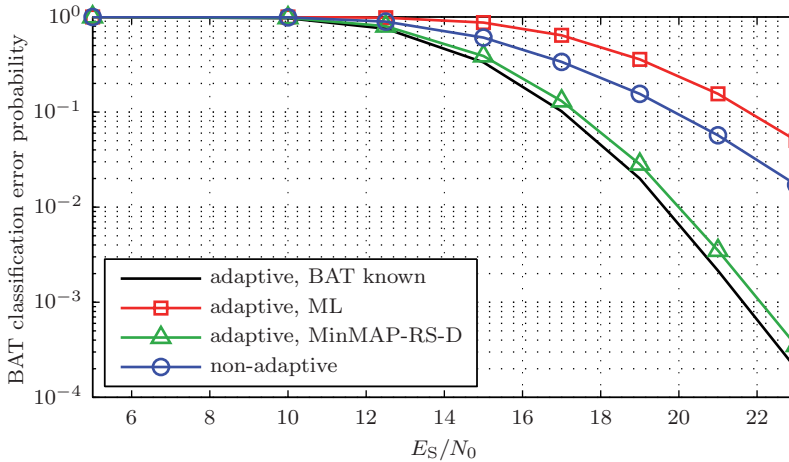


Fig. 15. BAT classification error probability P_{ce} of the MinMAP algorithm versus E_S/N_0 , frame length $K = 10$, $N_{crit} = 2$

algorithm is considerable and even overcompensates the benefit from adaptive modulation, the overall PER with perfect knowledge and classified BAT according to MinMAP-RS-D is very similar. The longer the frames, the smaller the influence classification errors on the PER will be. On the one hand, AMC benefits from more symbols to average. On the other hand, payload detection errors will occur more frequently with increasing frame lengths.

By using the presented AMC approach, we can fully gain from adaptive modulation (at costs of a higher complexity) without loss in the effective data rate due to signaling of the BAT, also in wireless communication scenarios.

5. Summary

In this contribution, a framework of likelihood-based automatic modulation classification algorithms for wireless orthogonal frequency division duplex systems with adaptive modulation has been presented. Instead of signaling the bit allocation table to the receiver, the bit allocation table can be efficiently detected solely based upon the received signal and side information which is available in time-division duplex communication systems.

It has been shown that the well-known maximum-likelihood algorithm does not offer a sufficiently high classification reliability in typical wireless communication scenarios. Therefore, an improved maximum-a-posteriori technique has been presented that utilizes additional information, i.e. a fixed bit allocation table per frame, channel reciprocity in time-division duplex systems and the information about the overall data rate. A metric simplification is possible which reduces the computational burden considerably without sacrificing much performance.

By using these advanced classifiers, there is almost no performance loss in terms of the packet error ratio compared to the case with perfect knowledge of the bit allocation table. Due to the moderate computational complexity and high classification reliability even for short

packets, the application of automatic modulation classification can be an attractive alternative to conventional signaling schemes.

The automatic modulation classification could be even further improved if the transmitter and receiver signaling processing is considered jointly. A brief example has been given by rotating the symbols of the conventional 16QAM scheme. However, this sophisticated topic of a joint transmitter and receiver design will be part of future research.

Algorithm	Metric
ML	$\sum_{k=1}^K \ln \left(\sum_{i=1}^{L^{(m)}} \exp \left(-\gamma_n \cdot \Delta_{n,i,k}^{(m)} ^2 \right) \right) - K \cdot \ln L^{(m)}$
MAP	$J_{\text{ML}}(\mathbf{d}_n, \mathcal{H}_n^{(m)}) + \ln p(\mathcal{H}_n^{(m)})$
MAP-RQ	$J_{\text{ML}}(\mathbf{d}_n, \mathcal{H}_n^{(m)}) + \ln p(\hat{\mathcal{H}}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$
MAP-RS	$J_{\text{ML}}(\mathbf{d}_n, \mathcal{H}_n^{(m)}) + \ln p(\hat{b}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$
MAP-RQ-D	two-step procedure in section 3.4 based on MAP-RQ
MAP-RS-D	two-step procedure in section 3.4 based on MAP-RS
MinMAP-RQ	$-\gamma_n \cdot \sum_{k=1}^K \min_{S_i^{(m)} \in \mathcal{I}^{(m)}} \left\{ \Delta_{n,i,k}^{(m)} ^2 \right\} - K \ln L^{(m)} + \ln p(\hat{\mathcal{H}}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$
MinMAP-RS	$-\gamma_n \cdot \sum_{k=1}^K \min_{S_i^{(m)} \in \mathcal{I}^{(m)}} \left\{ \Delta_{n,i,k}^{(m)} ^2 \right\} - K \ln L^{(m)} + \ln p(\hat{b}_{n,\text{Rx}}, \mathcal{H}_n^{(m)})$
MinMAP-RQ-D	two-step procedure in section 3.4 based on MinMAP-RQ
MinMAP-RS-D	two-step procedure in section 3.4 based on MinMAP-RS

Table 3. Metrics overview

6. References

- Boiteau, D. & Martret, C. L. (1998). A generalized maximum likelihood framework for modulation classification, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2165–2168.
- Campello, J. (1998). Optimal discrete bit loading for multicarrier modulation systems, *Proc. IEEE International Symposium on Information Theory*, p. 193.
- Chen, Y., Häring, L. & Czylwik, A. (2009). Reduction of AM-induced signaling overhead in WLAN-based OFDM systems, *Proc. of the 14th International OFDM-Workshop (InOWo)*, Hamburg, Germany, pp. 30–34.
- Chow, P., Cioffi, J. & Bingham, J. (1995). A practical discrete multitone transceiver loading algorithm for data transmission over spectrally shaped channels, *IEEE Transactions on Communications* 43(2/3/4): 773–775.
- Czylwik, A. (1996). Adaptive OFDM for wideband radio channels, *Proc. of Global Telecommunications Conference GLOBECOM '96*, pp. 713–718.
- Dobre, O. A., Abdi, A., Bar-Ness, Y. & Su, W. (2007). Survey of automatic modulation classification techniques: classical approaches and new trends, *IET Communications* 1(2): 137–156.

- Dobre, O., Bar-Ness, Y. & Su, W. (2004). Robust QAM modulation classification algorithm using cyclic cumulants, *Proc. of IEEE Wireless Communication and Networking Conference (WCNC)*, Vol. 2, pp. 745–748 Vol.2.
- Edinger, S., Gaida, M. & Fliege, N. J. (2007). Classification of QAM signals for multicarrier systems, *Proc. of the European Signal Processing Conference (EUSIPCO)*, pp. 227–230.
- Fischer, R. F. H. & Huber, J. B. (1996). A new loading algorithm for discrete multitone transmission, *Proc. of IEEE Global Telecommunications Conference GLOBECOM'96*, pp. 724–728.
- Häring, L., Chen, Y. & Czylwik, A. (2010a). Automatic modulation classification methods for wireless OFDM systems in TDD mode, *IEEE Trans. on Communications* (9): 2480 – 2485.
- Häring, L., Chen, Y. & Czylwik, A. (2010b). Efficient modulation classification for adaptive wireless OFDM systems in TDD mode, *Proc. of the Wireless Communications and Networking Conference*, Sydney, Australia, pp. 1–6.
- Häring, L., Chen, Y. & Czylwik, A. (2011). Utilizing side information in modulation classification for wireless OFDM systems with adaptive modulation, *Proc. of the IEEE Vehicular Technology Conference 2011-Fall*, San Francisco, USA.
- Hsue, S. Z. & Soliman, S. S. (1989). Automatic modulation recognition of digitally modulated signals, *Proc. of IEEE MILCOM*, pp. 645–649.
- Huang, Q.-S., Peng, Q.-C. & Shao, H.-Z. (2007). Blind modulation classification algorithm for adaptive OFDM systems, *IEICE Trans. Commun.* E.90-B No. 2: 296–301.
- Hughes-Hartogs, D. (1987). Ensemble modem structure for imperfect transmission media, *U.S. Patent 4,679,227*.
- IEEE (2005). IEEE 802.11n, *Technical report*, <http://grouper.ieee.org/groups/802/11>.
- Lampe, M. (2004). *Adaptive Techniques for Modulation and Channel Coding in OFDM Communication Systems*, PhD thesis.
- Long, C., Chugg, K. & Polydoros, A. (1994). Further results in likelihood classification of QAM signals, *Proc. of IEEE MILCOM*, pp. 57–61.
- Medbo, J. & Schramm, P. (1998). Channel models for HiperLAN/2 in different indoor scenarios, ETSI/BRAN document no. 3ERI085B.
- Nandi, A. & Azzouz, E. (1998). Algorithms for automatic modulation recognition of communication signals, *IEEE Trans. on Communications* 46(4): 431–436.
- Polydoros, A. & Kim, K. (1990). On the detection and classification of quadrature digital modulations in broad-band noise, *IEEE Trans. on Communications* 38(8): 1199–1211.
- Sills, J. A. (1999). Maximum-likelihood modulation classification for PSK/QAM, *Proc. of IEEE MILCOM*, pp. 57–61.
- Starr, T., Cioffi, J. M. & Silverman, P. J. (1999). *Understanding Digital Subscriber Line Technology*, Prentice Hall.
- Swami, A. & Sadler, B. M. (2000). Hierarchical digital modulation classification using cumulants, *IEEE Trans. on Communications* 48: 416–429.
- Wei, W. & Mendel, J. (2000). Maximum-likelihood classification for digital amplitude-phase modulations, *IEEE Trans. on Communications* 48(2): 189–193.

User Oriented Quality of Service Framework for WiMAX

Niharika Kumar, Siddu P. Algur and Amitkeerti M. Lagare
*RNSIT
BVB College of Engineering
Motorola Mobility
India*

1. Introduction

IEEE 802.16 provides last mile broadband wireless access. Also called as WiMAX, IEEE 802.16 is rapidly being adopted as the technology for Wireless Metropolitan area networking (MAN). WiMAX operates at the microwave frequency and each WiMAX cell can have coverage area anywhere between 5 to 15 kilometers and provide data rates upto 70Mbps.

IEEE 802.16m has been submitted to ITU as a candidate for 4G. With data rates of 100Mbps for mobile users and 1Gbps for fixed users, IEEE 802.16m holds a lot of promise as a true 4G broadband wireless technology.

This chapter introduces a user based framework in WiMAX. In section 2, user based bandwidth allocation algorithms are introduced. In section 3, user based packet classification mechanism is explored. In section 4 user based call admission control algorithm is explored.

2. User based bandwidth allocation

IEEE 802.16 (WiMAX) provides differentiated Quality of Service (QoS) (IEEE 802.16 2004) (IEEE 802.16e 2005) (Vaughan-Nichols 2004). This is achieved by having five different types of service classes. Each of these service classes caters to specific type of data. Unsolicited Grant Services (UGS) supports real time data streams that generate fixed size packets at periodic intervals. For example Voice over IP without silence suppression, T1/E1. Extended Real Time Polling Services (eRTPS) is designed to support real-time service flows that generate variable sized data packets on periodic basis, like VoIP with silence suppression. Real Time Polling Services (RTPS) supports real time data streams that generate variable size packets on periodic basis. For example Multimedia formats like an MPEG video. Non Real Time Polling Services (nRTPS) supports delay tolerant data streams generating variable size data packets, like FTP. Best Effort(BE) supports data streams which do not require any service level. Ex Web browsing, Email etc.

User keeps generating the data. This data gets queued into one of the five service classes based on the type of data and the quality of service requirements for the data. Once the data

gets queued, the device needs to request for bandwidth so that the data packets can be transmitted. Classically, the widely used bandwidth allocation algorithms have followed contention based logic. The device contends for the wireless medium. If no other device is contending for the bandwidth then the device transmits the data. Algorithms like ALOHA, Slotted ALOHA, CSMA, CSMA-CD use contention based bandwidth allocation. Even IEEE 802.11 (Wi-Fi) uses contention based bandwidth allocation mechanism called CSMA-CA.

WiMAX supports demand based bandwidth allocation mechanism. Each Mobile Station (MS) is allocated small amount of bandwidth that is used by the MS to request for additional bandwidth. Based on the availability of bandwidth and the type of service requesting for bandwidth, the Base Station (BS) allocates bandwidth. MS requests bandwidth on a per service class basis and the BS allocates bandwidth on a per-SS basis. Various types of contention based bandwidth request/allocation mechanisms have been proposed in WiMAX. Aggregate bandwidth request mechanism is proposed in (Tao & Gani, 2009). Instead of sending separate bandwidth request for each service class, a single request is sent. Service class bandwidth allocation is proposed in (Wee & Lee, 2009). Delay intolerant service classes are provided bandwidth on priority. Subsequently delay tolerant service classes are allocated bandwidth. Adaptive bandwidth request scheme is proposed in (Liu & chen, 2008). Contention free bandwidth request opportunities are provided within the contention based request opportunities for some SS. Predictive bandwidth allocation algorithm is proposed in (Peng et. al, 2007). Based on the current arrival pattern, bandwidth is requested beforehand for future packets. Channel aware bandwidth allocation algorithm is proposed in (Lin et. al, 2008). Another form of adaptive bandwidth allocation algorithm is proposed in (Chiang et. al, 2007). The TDD frame is dynamically adjusted based on the amount of uplink and downlink data. In (Park, 2009) bandwidth request algorithm is proposed that takes both the current size of the queue and the deadline assigned to each packet. CDMA bandwidth request code based bandwidth allocation mechanism is proposed in (Lee et. al, 2010). The CDMA bandwidth request code is chosen randomly, but in (Lee et. al, 2010) the bandwidth request code is intelligently chosen so that the code itself indicates the amount of bandwidth needed by the MS. This reduces the number of control message transactions between the MS and SS. In (Rong et al, 2007) two algorithms are proposed namely adaptive power allocation (APA) and call admission control (CAC). The two algorithms work in tandem to allocate bandwidth to the MS.

All the algorithms proposed above are service class based bandwidth request/allocation algorithms. MS shall send bandwidth request for all its service classes. Bandwidth is then allocated based on the service class. All UGS service classes from different users are allocated bandwidth first then the RTPS service flows are allocated bandwidth followed by eRTPS. Next, the delay tolerant service class nRTPS is allocated bandwidth. Finally BE service class is allocated bandwidth. This method of bandwidth allocation treats all MS alike. If there are 10 MS in the network and if all of them are generating BE traffic then all the BE service classes are allocated bandwidth on a first come first serve basis. Of these 10 users, there may be some users who may wish to pay more if their BE traffic is treated on priority. So, users can be segregated into different groups and bandwidth can be allotted to the users based on the group to which they belong to. In this section we shall explore three user based bandwidth allocation algorithms. Fig. 1 shows service class based bandwidth allocation mechanism.

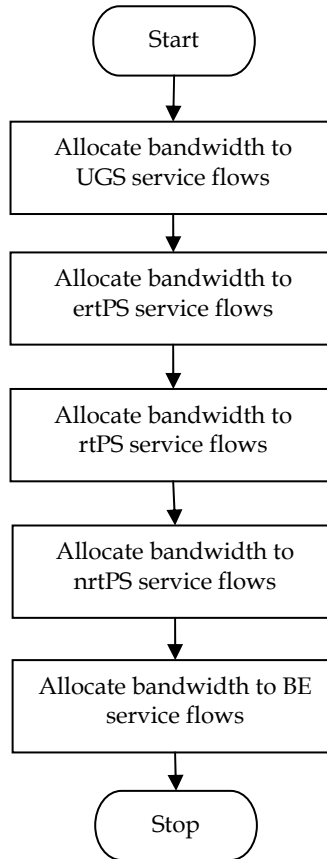


Fig. 1. Service class based bandwidth allocation mechanism.

2.1 Differentiated Bandwidth Allocation Mechanism (DBAM)

There shall be three different categories of users/MS as listed in Table 1.

User Category	Priority Value	Description
High-Priority User/MS/SS	1	Users who will receive higher priority for their traffic within each of the WiMAX service class. High-Priority users could be those users who are ready to pay more to enjoy higher QoS.
Low-Priority User/MS/SS	2	Users who will receive lower priority for their traffic for each of the WiMAX service class. Low-Priority users could be those users who wish to pay less and settle for lower quality of service.
Regular User/MS/SS	0	Users who fall in-between High-Priority and Low-Priority users.

Table 1. Classification of Users into three different categories.

Bandwidth allocation is done for all the service class for the three types of users as per fig. 2

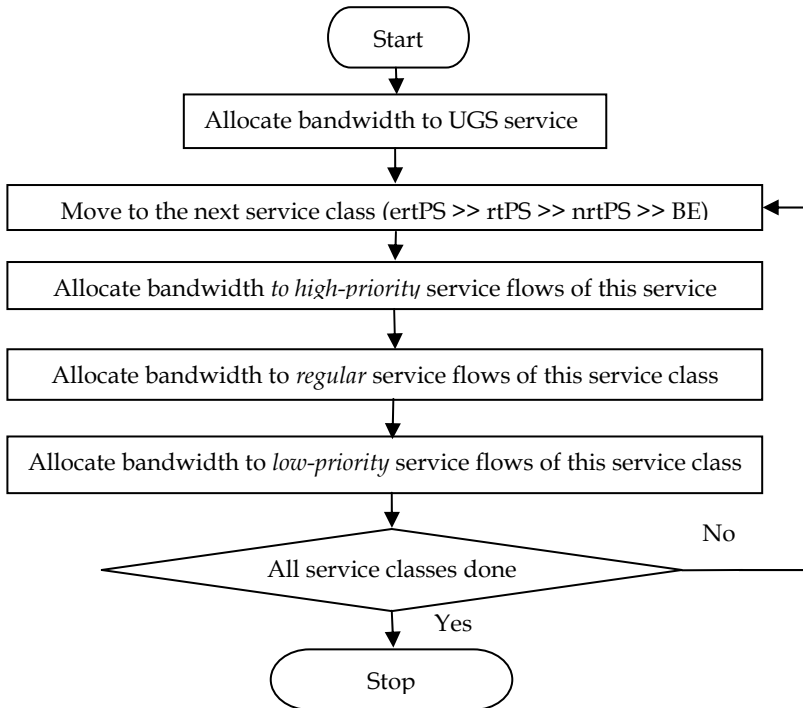


Fig. 2. DBAM Algorithm.

From the algorithm in fig.2, we see that when the BS receives bandwidth requests for BE traffic from High-Priority, Regular and Low-Priority users, BS shall allocate bandwidth first to the high-priority user then the regular user and finally to the low-priority user (Kumar et al. 2011a).

2.1.1 Implementation of BDAM

The WiMAX Network Reference architecture is given in the fig. 3.

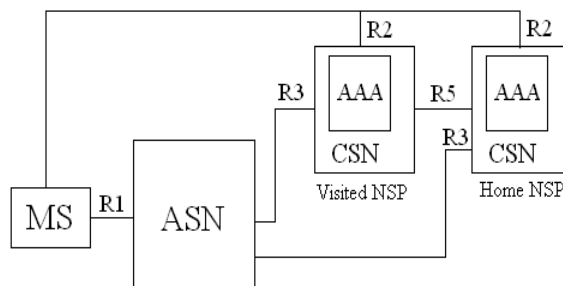


Fig. 3. WiMAX Network Reference Architecture.

Each network service provider (NSP) has a Authentication, Authorization and Accounting (AAA) server.. This server maintains the information about the users. The Access Service Network (ASN) interacts with the AAA server to obtain the information about the user.

The AAA server shall maintain a table of MAC address for the users and the priority value associated with the user. A sample state of the table could be as shown in Table-2

MAC Address	Priority Value
12:34:56:78:9a:bc	2
bc:9a:78:56:34:12	1
11:11:11:11:11:11	0
22:22:22:22:22:22	0
33:33:33:33:33:33	2
88:77:66:11:22:44	1
.....

Table 2. Sample state table of priority values for users.

When the MS initiates the ranging process, it sends the Ranging Request (RNG-REQ). Upon receiving the ranging request, BS shall query the AAA server to obtain the priority value associated with the user. BS shall store the priority value for the user in its local cache.

Subsequently, when the MS makes bandwidth request for any of its service flows, BS shall check the priority of the MS. Based on the priority value, bandwidth shall be allotted to the service flow.

2.1.2 Analytical modeling

Table 3 lists the notations used for analytical modeling.

Symbol	Description
$ertps_pri_bw_req(p)$	Bandwidth needs of pth ertPS service flow of priority SS.
$ertps_reg_bw_req(p)$	Bandwidth needs of pth ertPS service flow of regular SS.
$ertps_npr_bw_req(p)$	Bandwidth needs of pth ertPS service flow of low-priority SS
$ertps_pri_bw_allot(p)$	Bandwidth allotted to the pth ertPS service flow of priority SS.
$ertps_reg_bw_allot(p)$	Bandwidth allotted to the pth ertPS service flow of regular SS.
$ertps_npr_bw_allot(p)$	Bandwidth allotted to the pth ertPS service flow of low-priority SS.
tot_bw	Total bandwidth available on the uplink for the current frame
tr	Minimum Reserved traffic rate
avl_bw	Amount of unallocated bandwidth available in the frame.
m	Number of high-priority ertPS service flows
n	Number of regular ertPS service flows
o	Number of low-priority ertPS service flows

Table 3. Notations used in Analytical Modeling.

Throughput modeling is described below. For the purpose of brevity bandwidth allocation is explained for the three types of users for eRTPS service flow. Similar equations can be derived for the other service flows.

BS allots bandwidth to the high-priority eRTPS service flows as per eqn 1.

$$ertps_pri_bw_allot(p) = \begin{cases} ertps_pri_bw_req(p) & \text{if } ertps_pri_bw_req(p) < tr \\ & \text{and } ertps_pri_bw_req(p) < avl_bw \\ tr & \text{if } tr \leq ertps_pri_bw_req(p) \\ & \text{and } ertps_pri_bw_req(p) < avl_bw \\ avl_bw & \text{if } avl_bw \leq ertps_pri_bw_req(p) \\ & \text{and } avl_bw \leq tr \\ tr & \text{otherwise} \end{cases} \quad (1)$$

Once bandwidth is allotted to a high-priority eRTPS service flow, the leftover bandwidth is calculated as per eqn 2.

$$avl_bw = tot_bw - \left(\sum_{j=1}^x ertps_pri_bw_allot(j) \right) \quad (2)$$

$x \leq m,$

After all the high-priority eRTPS service flows are allotted bandwidth, bandwidth is allotted to the regular eRTPS service flows as per eqn 3.

$$ertps_reg_bw_allot(p) = \begin{cases} ertps_reg_bw_req(p) & \text{if } ertps_reg_bw_req(p) < tr \\ & \text{and } ertps_reg_bw_req(p) < avl_bw \\ tr & \text{if } tr \leq ertps_reg_bw_req(p) \\ & \text{and } ertps_reg_bw_req(p) < avl_bw \\ avl_bw & \text{if } avl_bw \leq ertps_reg_bw_req(p) \\ & \text{and } avl_bw \leq tr \\ tr & \text{otherwise} \end{cases} \quad (3)$$

After allocating bandwidth to a regular eRTPS service flow, leftover bandwidth is calculated as per eqn. 4

$$avl_bw = avl_bw - \left(\sum_{j=1}^x ertps_reg_bw_allot(j) \right) \quad (4)$$

$x \leq n,$

Once we are through with the regular eRTPS service flows, bandwidth is allotted to the low-priority eRTPS service flows as per eqn. 5.

$$ertps_npr_bw_allot(p) = \begin{cases} ertps_npr_bw_req(p) & \text{if } ertps_npr_bw_req(p) < tr \\ & \text{and } ertps_npr_bw_req(p) < avl_bw \\ tr & \text{if } tr \leq ertps_npr_bw_req(p) \\ & \text{and } ertps_npr_bw_req(p) < avl_bw \\ avl_bw & \text{if } avl_bw \leq ertps_npr_bw_req(p) \\ & \text{and } avl_bw \leq tr \\ tr & \text{otherwise} \end{cases} \quad (5)$$

After allotting bandwidth to the j th low-priority eRTPS service flow, leftover bandwidth is calculated as per eqn. 6

$$avl_bw = avl_bw - \left(\sum_{j=1}^x ertps_npr_bw_allot(j) \right) \quad (6)$$

$x \leq o,$

At this point bandwidth has been allotted to all the eRTPS connections. The above method of bandwidth allocation is repeated for RTPS, nRTPS and BE. This ensures that for each service flow, bandwidth is allotted to high-priority users first followed by regular users and finally the low-priority users.

2.1.3 Simulation results

In order to evaluate DBAM, simulations were carried out on NS-2. Light WiMAX module (LWX) (Chen 2008) was used to simulate the WiMAX environment in NS-2. Strict priority bandwidth allocation algorithm of LWX was modified to accommodate DBAM algorithm. Simulations were carried out with the parameters from table 4.

Parameter	Value
Uplink data rate	10 Mbps
OFDMA Frame Duration	5 ms
OFDMA symbol time	100.94 μ s
eRTPS data arrival rate	1 Mbps

Table 4. Simulation parameters for DBAM.

Simulation network was setup such that at any point in time, 33% of the SS are priority SS, next 33% are regular SS and the final 1/3rd are low-priority SS. Each SS generates only eRTPS traffic. Uplink data is generated at the rate of 1Mbps. Downlink ftp traffic was also added. Downlink data is generated at the rate of 1Mbps.

Simulation results for throughput are shown in Fig. 4.

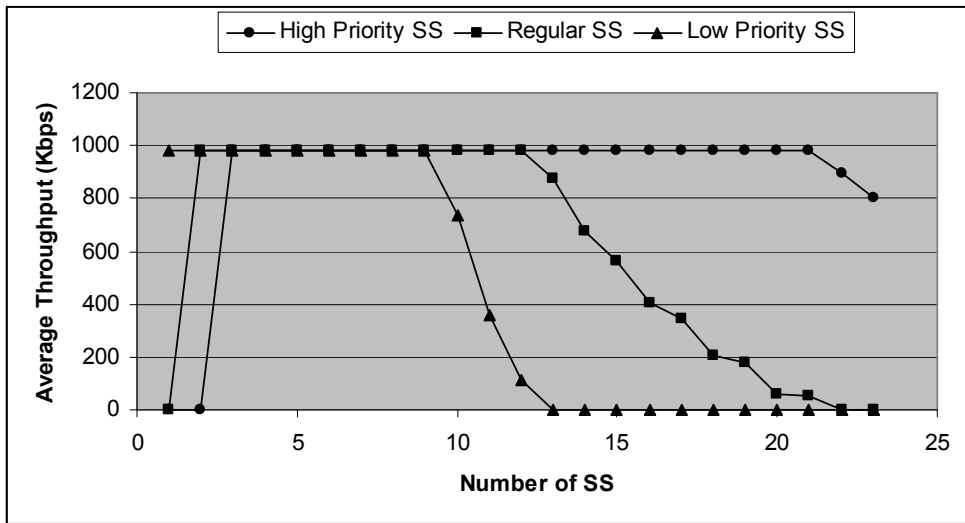


Fig. 4. Simulation results for throughput for the three types of SS.

When the number of MS is 9 each MS has sufficient bandwidth to transmit its data. But, when the number of SS is more than 9, there isn't sufficient bandwidth to support all SS. DBAM provides bandwidth to high-priority SS first then regular SS and the leftover bandwidth is shared by low-priority SS. When the number of SS crosses 13, bandwidth for regular SS keeps reducing. Theoretical Results are shown in Figure 5.

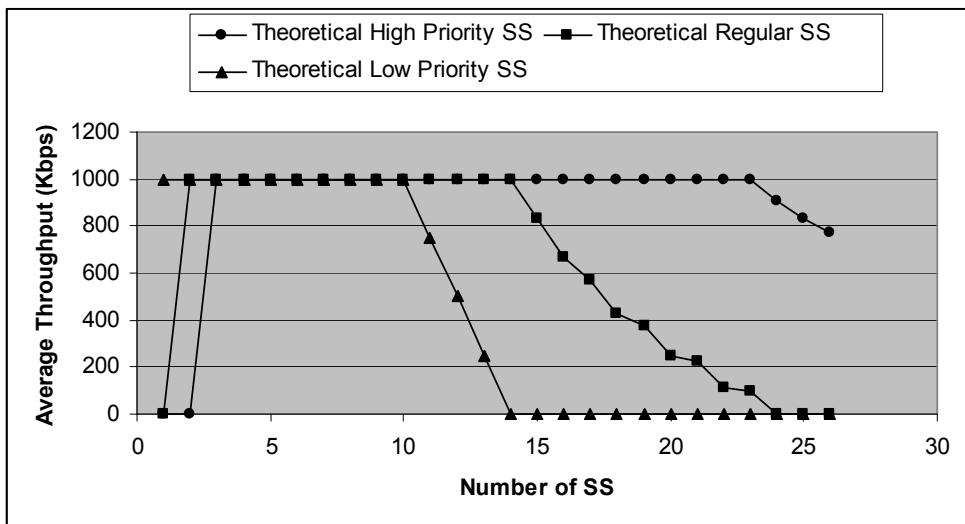


Fig. 5. Theoretical results for DBAM.

Comparing figure 4 and figure 5 we see that the simulation results closely follow the theoretical results.

By introducing DBAM we can provide graded quality of service to the users. This is a win-win situation for both users and operators. The users win because their data gets prioritized and hence they get a better quality of service. The service providers stand to gain because they get higher revenue for the same amount of data being transmitted. Its just that the order of bandwidth allocation is modified.

2.2 Enhanced Differentiated Bandwidth Allocation Mechanism (eDBAM)

In case of DBAM, the order of bandwidth allocation follows the below sequence:

High-priority RTPS > Regular priority RTPS > Low-priority RTPS > High-priority nRTPS > Regular priority nRTPS > Low-priority nRTPS > High-priority BE > Regular priority BE > Low-priority BE

Basically DBAM ensured that the order of service class is maintained and within the service class we can have graded users. However there is scope for further optimization. We can have seven different ways in which the bandwidth can be allotted. Table 5 and Table 6 list the seven different ways in which bandwidth can be allotted. Each column in the table represents a unique way of bandwidth allotment. The order of allotment is from top to bottom (Kumar et. al, 2011b).

DBAM	eDBAM Method 1	eDBAM Method 2	eDBAM Method 3
High-priority RTPS	High-Priority RTPS	High-priority RTPS	High-priority RTPS
Regular priority RTPS	High-priority nRTPS	Regular priority RTPS	High-priority nRTPS
Low-priority RTPS	High-priority BE	Low-priority RTPS	Regular priority RTPS
High-priority nRTPS	Regular priority RTPS	High-priority nRTPS	Low-priority RTPS
Regular priority nRTPS	Low-priority RTPS	High-priority BE	Regular priority nRTPS
Low-priority nRTPS	Regular priority nRTPS	Regular priority nRTPS	Low-priority nRTPS
High-priority BE	Low-priority nRTPS	Low-priority nRTPS	High-priority BE
Regular priority BE	Regular priority BE	Regular priority BE	Regular priority BE
Low-priority BE	Low-priority BE	Low-priority BE	Low-priority BE

Table 5. Method 1 to Method 3 of eDBAM.

eDBAM Method 4	eDBAM Method 5	eDBAM Method 6	eDBAM Method 7
High-priority RTPS	High-priority RTPS	High-priority RTPS	High-priority RTPS
High-priority nRTPS	Regular priority RTPS	Regular priority RTPS	Regular priority RTPS
Regular priority RTPS	High-priority nRTPS	High-priority nRTPS	Low-priority RTPS
Low-priority RTPS	Regular priority nRTPS	Regular priority nRTPS	High-priority nRTPS
High-priority BE	Low-priority RTPS	High-priority BE	Regular priority nRTPS
Regular priority nRTPS	Low-priority nRTPS	Regular priority BE	High-priority BE
Low-priority nRTPS	High-priority BE	Low-priority RTPS	Regular priority BE
Regular priority BE	Regular priority BE	Low-priority nRTPS	Low-priority nRTPS
Low-priority BE	Low-priority BE	Low-priority BE	Low-priority BE

Table 6. Method 4 to Method 7 of eDBAM.

In eDBAM (for example Method 2), low priority service class of high priority user (ex: Low-Priority BE) can be allocated bandwidth ahead of high-priority service class of regular/low-priority user (Regular/Low priority nRTPS). This out of turn allocation of bandwidth improves the throughput for even low priority service class (BE) for high-priority users.

2.2.1 Implementation

Implementation of eDBAM is similar to DBAM. The AAA server shall maintain a mapping of MAC address to the priority value associated with the MAC address. When a MS sends RNG-REQ to BS, BS shall obtain the priority value associated with the MS and allocated bandwidth based on one of the seven methods proposed for eDBAM. BS does not switch between the seven different methods of eDBAM. Each BS shall implement one of the seven methods and stick to that method throughout its operation.

2.2.2 Analytical modeling

Throughput modeling follows similar patterns as that of DBAM. Only the order of bandwidth allocation shall change. Delay modeling is explained in this section. The notations used for delay modeling are given in Table 7.

Symbol	Description
λ	Mean arrival rate
μ	Mean service rate
ρ	Service utilization
L	Mean number of packets of a service flow for a particular SS in the system.
W	Mean end-to-end delay for secure packets of a particular service flow for a particular SS.

Table 7. Delay modeling parameters.

For BE packets, Packet arrivals are assumed to have a Poisson arrival.

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \quad (7)$$

We know that, Service Utilization = Mean arrival rate / Mean service rate. i.e.

$$\rho = \frac{\lambda}{\mu} \quad (8)$$

For BE traffic (exponential distribution), mean number of packets for a service flow for a particular-SS is given in (9)

$$L = \frac{\rho}{1-\rho} \quad (9)$$

Queuing delay for a service flow for a particular SS is given in (10)

$$W = \frac{L}{\lambda} = \frac{\frac{\rho}{1-\rho}}{\lambda} \quad (10)$$

For RTPS and nRTPS we assume constant arrival pattern. So mean number of packets for a service flow for a particular SS is given in (11)

$$L = \frac{\rho(2-\rho)}{2(1-\rho)} \quad (11)$$

Hence the queuing delay for packets that have constant arrival pattern is:

$$W = \frac{L}{\lambda} = \frac{\frac{\rho(2-\rho)}{2(1-\rho)}}{\lambda} \quad (12)$$

2.2.3 Simulation of eDBAM

Simulation was carried out using NS 2.29. LWX was used to simulate wimax on top of ns2. Simulations were carried out for method-2 for eDBAM. Simulation parameters used, are given in Table-8

Parameter	Value
Data rate	10 Mbps
OFDMA Frame Duration	5 ms
OFDMA symbol time	100.94 μ s
RTPS data arrival rate	333 Kbps
nRTPS data arrival rate	333 Kbps
BE data arrival rate	333 Kbps

Table 8. Simulation parameters for eDBAM.

Simulation setup was done such that at any given time the network consists of 1/3rd High-priority SS, 1/3rd Regular-SS and 1/3rd low-priority SS. Each SS is assumed to have RTPS, nRTPS and BE traffic. Downlink ftp traffic at 1 Mbps was introduced.

2.2.3.1 Throughput results

Simulation was done to compare the throughput for High-Priority, Regular and Low-Priority BE traffic. Fig. 6 shows the simulation results. A comparison with theoretical results is also provided.

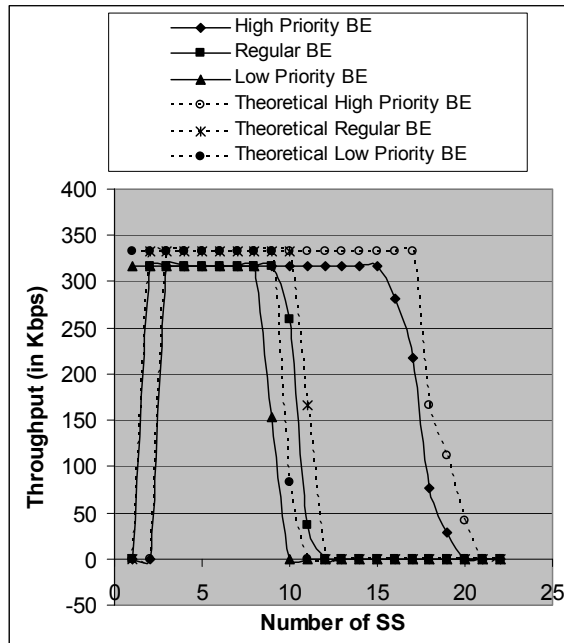


Fig. 6. Throughput for BE traffic for the three different types of user.

From fig. 6 we see that as the number of SS in the network increases, the throughput form Low-priority BE drops. Subsequently the throughput reduces for regular BE and finally the throughput for High-priority BE. Since method-2 prioritized high-priority BE ahead of Regular nRTPS and Low-priority nRTPS, simulations were carried out for the service flow. Results of simulation are shown in Fig.7.

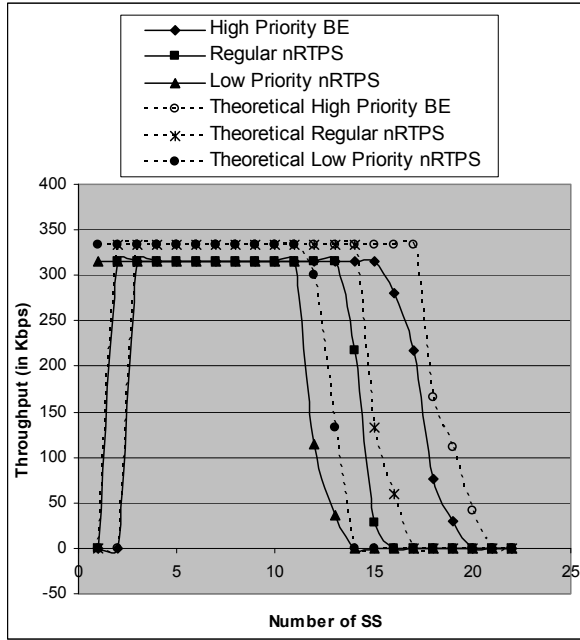


Fig. 7. Throughput for High Priority BE v/s Regular nRTPS v/s Low priority nRTPS.

2.2.3.2 Delay results

Simulations were carried out to find the delay incurred by the service flows. Fig. 8 shows the delay for High-Priority BE, Regular BE and Low-Priority BE.

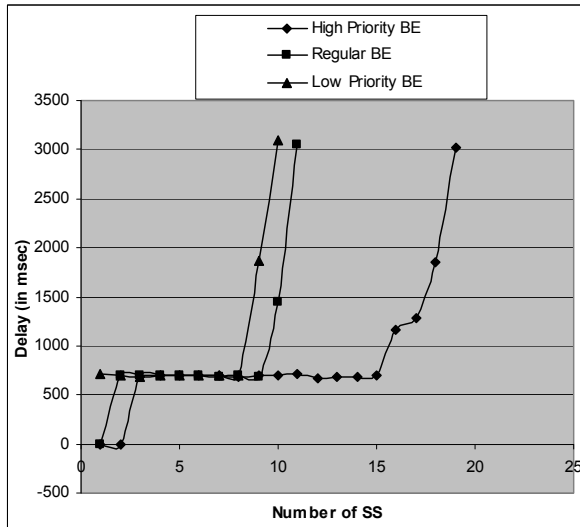


Fig. 8. Delay for High-Priority BE v/s Regular BE v/s Low-Priority BE.

Packet delay was measured for High-Priority BE, Regular nRTPS and Low-Priority nRTPS. Results of simulation are shown in Fig. 9.

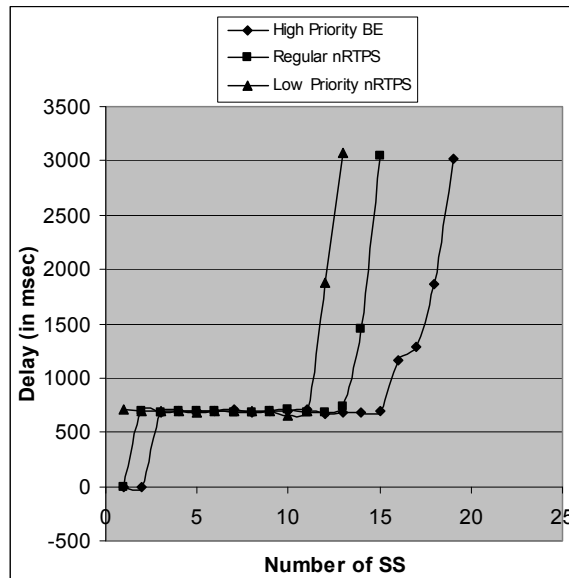


Fig. 9. Delay Results for High Priority BE v/s Regular nRTPS v/s Low Priority nRTPS.

From Fig. 8 and Fig. 9 we see that using eBBAM, packets from high-priority SS are subjected to lesser delay compared to regular and low-priority SS.

2.2.3.3 DBAM v/s eDBAM

Simulations were done to compare the throughput and delay for DBAM and eDBAM. Fig. 10 shows the throughput comparison for DBAM and eDBAM. We consider method-2 for eDBAM.

From Fig. 10 we observe that the throughput for DBAM drops down much before eDBAM. This is because in case of eDBAM, high-priority BE is allotted bandwidth ahead of regular nRTPS and low-priority nRTPS. Figure 11 shows the simulation results for delay. Again eDBAM fairs better than DBAM.

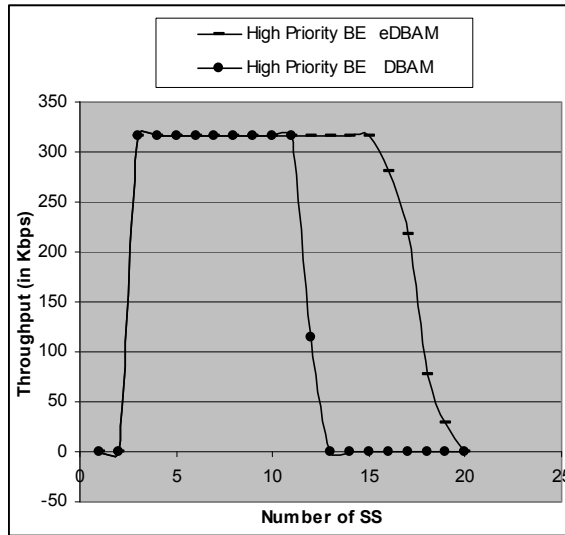


Fig. 10. Throughput comparison for eDBAM and DBAM for high priority BE.

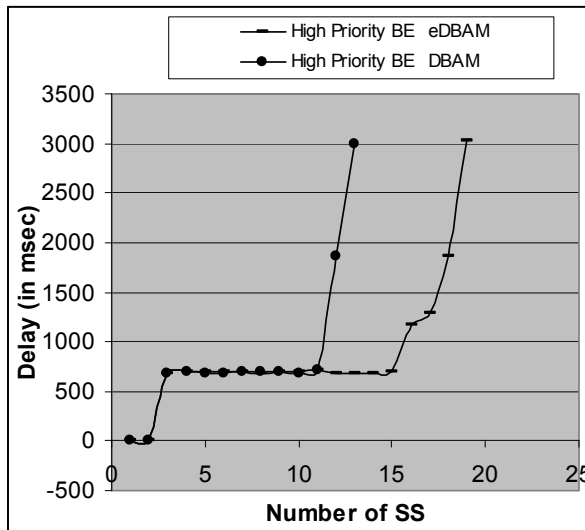


Fig. 11. Delay results for eDBAM and DBAM for High Priority BE.

2.3 Network Aware Differentiated Bandwidth Allocation Mechanism (nDBAM)

Though eDBAM improves the throughput, the algorithm is indifferent to the current network conditions. Especially, if eDBAM method-1 is implemented, it could result in delays for regular and low-priority RTPS. Users might face jitter when they are viewing videos. This might not be desirable. nDBAM takes care of current network conditions before

allocating bandwidth to the different service flows. The steps for nDBAM algorithm as given below.

Step 1. Users shall be allotted bandwidth as per one of the selected Seven methods of eDBAM.

BS keeps monitoring the network condition. BS could poll the SS to know their current queue length and the average queuing delays faced for each service flow. BS and SS can use the ranging mechanism to pass the information between them.

Step 2. If the average queuing delay exceeds the QoS limits for the service class then the BS shall fallback from eDBAM to DBAM bandwidth allocation mechanism

Step 3. BS checks with the SS if the average queuing delay has reduced. If yes then BS sticks to DBAM. If the average queuing delay is still high then BS falls back to First-come-first-serve (FCFS) method of bandwidth allocation.

Step 4. BS keeps monitoring the queuing delay. If the delay reduces and stays within acceptable limits then BS moves back to eDBAM algorithm

2.3.1 Implementation

BS does ranging at periodically with the SS. Ranging process is generally done to adjust the power levels and the clock skews. During the ranging process, BS can also request for the current queue state for the different service flows. As a part of ranging response (RNG-RSP) The SS can send the queue state to BS. The information is generally sent as a TLV (Type-Length-Value) header. A new header will be required to send the queue state information. Table 9 lists an example for the TLV.

Type	Length	Value	Scope
Unused TLV type (ex: 105)	1	Average Queue delay for Service flow	RNG-RSP

Table 9. TLV header used to send Queue state.

BS receives the RNG-RSP from all the SS for each of their service class. BS then checks if the queuing delay is within the QoS limits for the service class. If not then it means that the eDBAM algorithm is introducing delay for regular and low-priority users. So, BS shifts from eDBAM to DBAM.

3. User based packet classification algorithm

We know that WiMAX supports 5 different types of service classes i.e. UGS, RTPS, eRTPS, nRTPS, BE. When a user generates data (ex: video packets) they are classified and placed into one of the 5 queues at SS (ex: Video packets are classified as RTPS packets and placed in the RTPS queue). As the user keeps generating data packets, these are classified and placed in one of the queues.

This method of classification is application specific. i.e. if the user keeps generating video packets they are always classified as RTPS packets and placed in RTPS queue and if the user generates web browsing/email packets they are generally classified as BE packets and places in BE queue. Packet classification is not user specific. i.e. there may be some users

who are ready to pay more if their browsing packets are treated as high priority packets i.e. the browsing packets generated by such users are treated as RTPS packets instead of BE packets and placed in RTPS queue.

There may be some users who may wish to pay less and still enjoy broadband facility. For such users we may want to downgrade even their high priority packets like RTPS packets and treat them as low priority BE packets. A third set of users may fall in-between the high-priority and low-priority users.

There shall be 8 different ways of classifying the packets as given in Table 10 (Lagare & Das 2009).

Priority	Bit Value	Description
0	000	802.16e's existing packet classification mechanism is retained. i.e. real time packets will be placed in RTPS queue. Non real time packets are placed in nRTPS queue and delay tolerant packets are placed in BE queue.
1	001	RTPS, nRTPS and BE packets are classified as real-time packets and placed in RTPS queue.
2	010	nRTPS packets are promoted as RTPS packets and all BE packets are promoted as nRTPS packets
3	011	Only the BE packets are promoted as RTPS packets. Other Packets are placed in their respective priority queues.
4	100	Only the BE packets are promoted as nRTPS packets. Other Packets are placed in their respective priority queues.
5	101	RTPS, nRTPS and BE packets are classified as delay tolerant packets and moved to BE queue.
6	110	RTPS packets will be blocked. This priority level can be set to a certain set of users so that these users can be blocked from transmitting RTPS packets like MPEG videos.
7	111	RTPS and nRTPS packets will be blocked. This priority can be set to very low priority users.

Table 10. Eight different ways of packet classification.

3.1 Implementation

When the MS enters the network, it sends the RNG-REQ to BS. On receiving the range request, BS shall check the priority value associated with the SS. This priority value is passed to the SS in the RNG-RSP. On receiving the priority value the SS shall classify the packets as per table 10.

3.2 Simulation

Simulations were carried out to observe the improvement in throughput by implementing user based packet classification. Priority 3 scenario of table 11 was simulated. The simulation network consists of one priority MS whose packets are prioritized as per Priority

3. Other MS are regular users whose packets are prioritized as per priority 1. Table 11 lists the simulation parameters used.

Parameter	Value
Uplink Bandwidth	2Mbps
Uplink Frame Duration	1msec (2000 bits)
Number of Uplink frames per second	1000/Sec
Maximum Uplink bandwidth per SS per Frame	400 bits/frame
Minimum Reserved Traffic Rate for RTPS	240Kbps
Arrival Pattern for RTPS Traffic	Variable bit rate packets at regular interval of time
Arrival Pattern for BE Traffic	Poisson Arrival
Average arrival Rate for RTPS traffic	160Kbps
Average arrival Rate for BE traffic	72Kbps

Table 11. Simulation Parameters.

Figure 12 shows the simulation results for BE traffic when the priority MS and regular MS generate both RTPS and BE packets. For priority MS, the BE packets are classified as RTPS packets.

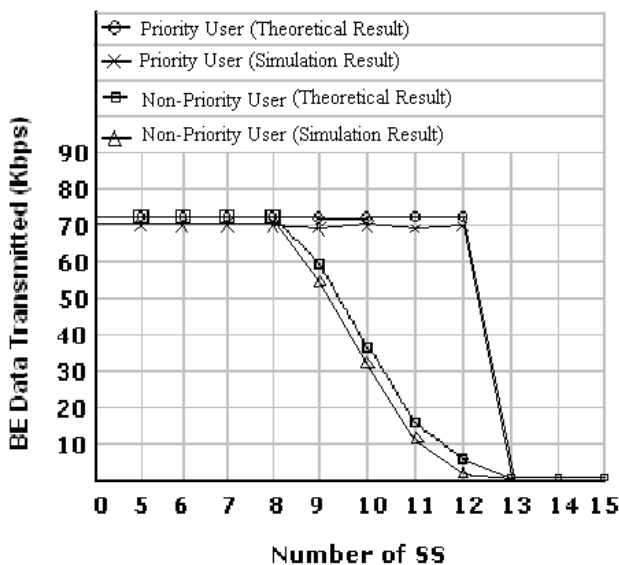


Fig. 12. BE data (in Kbps) transmitted by priority-SS compared to regular-SS.

From Fig. 12 we see that, when the number of MS in the network are less than 8, both Priority MS and non-priority MS are able to transmit all their data. When the number of MS in the network goes beyond 8, there isn't enough bandwidth to support the BE traffic for non-priority users. So the average throughput for non-priority user drops. Since priority MS

request bandwidth for their BE traffic as RTPS traffic, priority MS continue to receive bandwidth. Beyond 12 SS there isn't enough bandwidth to support elevation of BE traffic as RTPS traffic. So throughput for even priority-MS drops down. Fig. 13 shows the simulation results when the network consists of only BE traffic.

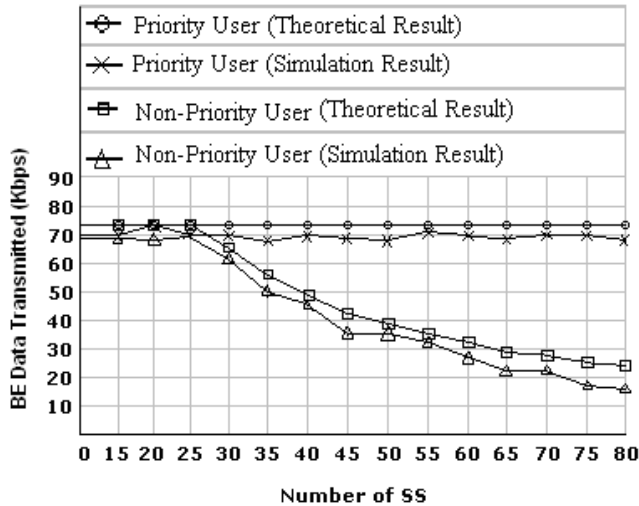


Fig. 13. BE data transmitted by priority-MS and regular-MS when only BE packets are present in the network.

In Fig. 13 we see that priority MS enjoy constant throughput of 70Kbps where as the throughput for non-priority MS keeps decreasing as the number of MS increases. This happens because BE traffic for priority MS is treated as RTPS traffic. So bandwidth is allotted to priority MS. The leftover bandwidth is shared by non-priority MS. So, by implementing priority based packet classification we can provide graded QoS to the users.

4. User based Call Admission Control (CAC) algorithm

Call admission control (CAC) plays a very important role in the IEEE 802.16 based wireless network. WiMAX networks aim to ensure that the QoS requirements for each service class are met. In order to provide QoS, the network should have a robust CAC algorithm.

When an SS/MS wants to establish a connection for a particular service class, it sends a DSA (Dynamic Service Addition) request to BS. This DSA request also contains the QoS parameters for the service class. Upon receiving the DSA request the BS decides to accept or reject the connection. If BS accepts the connection then it has to support the QoS needs of that connection.

When a BS decides to accept a connection, various factors need to be considered. For example the minimum and maximum data rates on the connection, the delay and jitter parameters for the connection etc. There can be other criteria like fairness, revenue per connection that can also play a role while admitting a connection.

Many CAC algorithms have been proposed both for wired and wireless medium. Because of the unique characteristics of wireless medium, many of the CAC algorithms of wired world cannot be applied to the wireless networks. Researches have proposed some CAC algorithms for WiMAX. In (Chen et. al, 2005) a simple bandwidth based CAC algorithm is proposed. A new connection is accepted if the bandwidth requirements for the connection can be satisfied by the BS. This algorithm does not take into consideration the deadline consideration of the connections. Once the bandwidth is allocated to the connection, the available bandwidth is calculated using the below equation:

$$BW_{avail} = BW - \sum_{s \in \{UGS, RTPS, nRTPS\}} \sum_{i=1}^{N^s} C_i^s [rate] \quad 13$$

Where $C_i^s [rate]$ represents the data rate for the i^{th} connection which belongs to s service class. In (Chandra & Sahoo, 2007) a QoS aware CAC is proposed. BS contains CAC queues for each service class. So there shall be 5 CAC queues (one each for UGS, RTPS, eRTPS, nRTPS and BE). When an SS makes a CAC request for a particular connection, the BS shall queue the request in one of queues based on the QoS requirements for the Class. BS then goes through each of the queues and accepts the connections. (Chandra & Sahoo, 2007) also provides criteria for call admission for each of the service class. In (Shu'aibu et. al, 2010) (Shu'aibu et. al, 2011) a partition based CAC algorithm is proposed. The total bandwidth is divided into many partitions like constant bit rate partition (CBR), variable bit rate partition (VBR) and Handover partition (HO) etc. CAC is applied to each of these partitions. CAC algorithms proposed above, are all service class based algorithms. In this section we shall look at user based CAC algorithm. The algorithm is based on (Chandra & Sahoo, 2007).

4.1 User based CAC algorithm

Fig. 14 shows the control flow at the SS when a new connection request is sent.

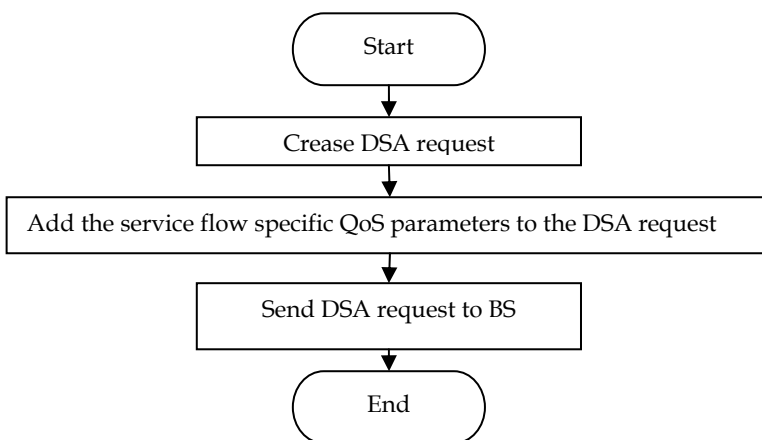


Fig. 14. User based CAC at SS.

Fig. 15 shows the classification of DSA request into different queues based on the priority of user.

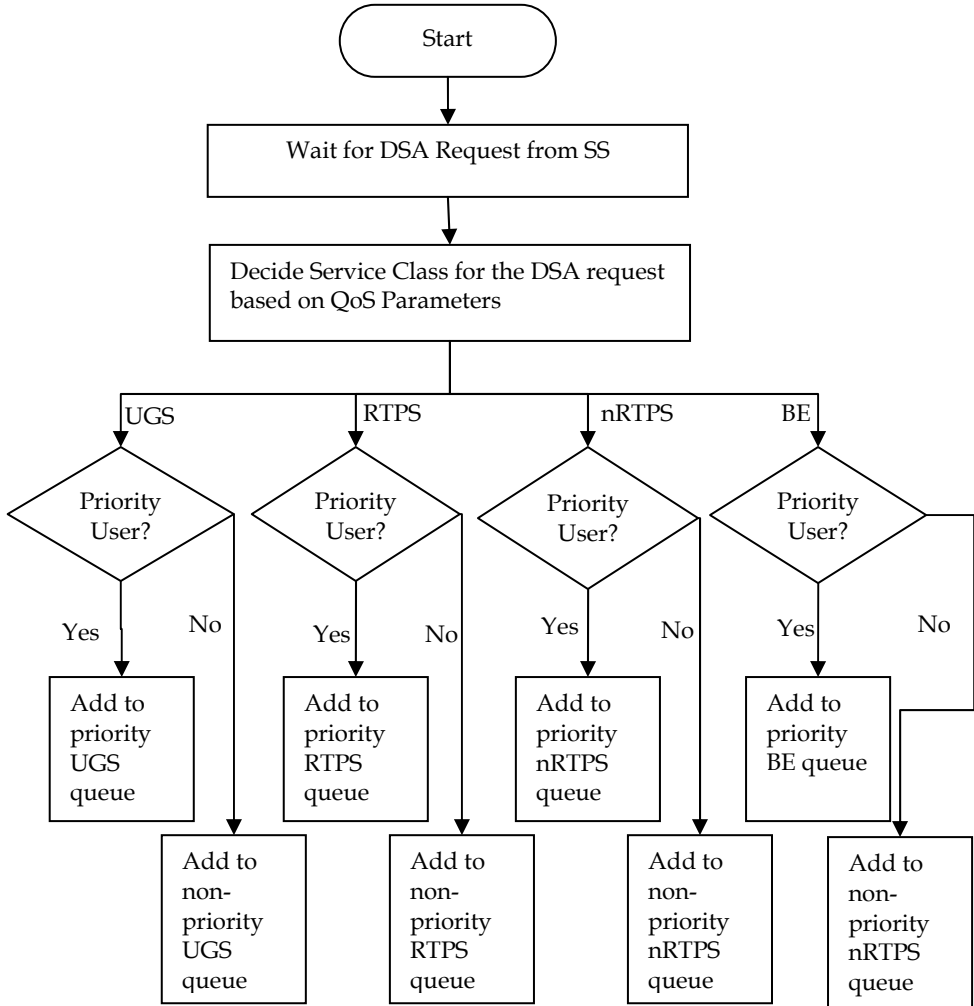


Fig. 15. Classification of DSA request at BS based on priority of User.

Because of lack of space, the control flow of eRTPS service class cannot be shown. However the logic for classifying the DSA request for eRTPS would be similar to UGS.

Once DSA requests are classified into the respective queues, BS goes through the DSA requests in each queue to admit the connection. High Level view of Admission control algorithm is given in Fig 16.

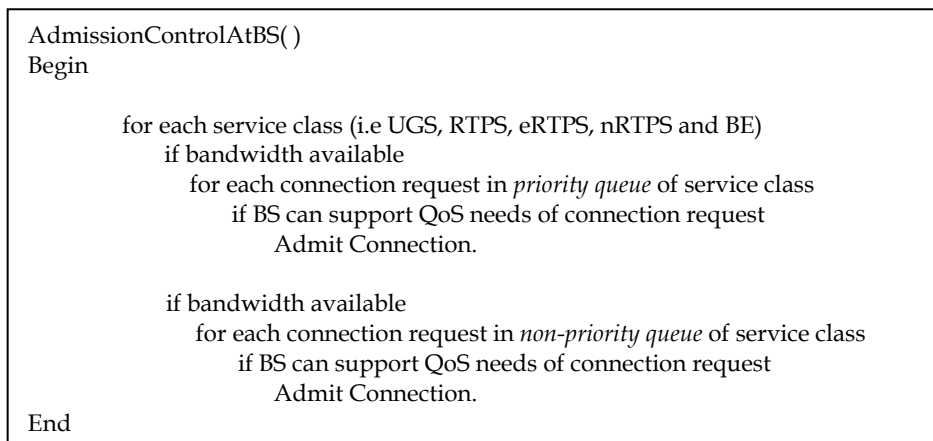


Fig. 16. User Based Admission control.

So, first, connections of priority UGS shall be accepted, followed by Non priority UGS. Priority RTPS and Non Priority RTPS follow next. Once RTPS connections are taken care, priority and Non priority eRTPS connection as admitted in that order. Subsequently priority and Non priority nRTPS connections are admitted. And finally priority and non priority BE connections are admitted.

4.2 Simulation results

Silulations were carried out to evaluate the performance of user based admission control algorithm. Simulation Parameters are given in Table-12.

Parameter	Value
Uplink Capacity	16 Mbps
Arrival of Connection Requests	Poisson arrival pattern
Lifetime of Connections	2 - 6 seconds
Data rate of UGS connections	256 kbps
Data rate of RTPS connections	256 kbps
Data rate of eRTPS connections	256 kbps
Data rate of nRTPS connections	256 kbps
Data rate of BE connections	256 kbps
Simulation Lifetime	200 seconds

Table 12. Simulation parameters.

Simulations were carried out to find the acceptance ratio for the connection requests for priority users and non-priority users. Acceptance ratio is defined as the ratio between the number of connections accepted to the total number of connections requested. Fig. 17 shows the simulation results for acceptance ratio when the network contains only RTPS connections.

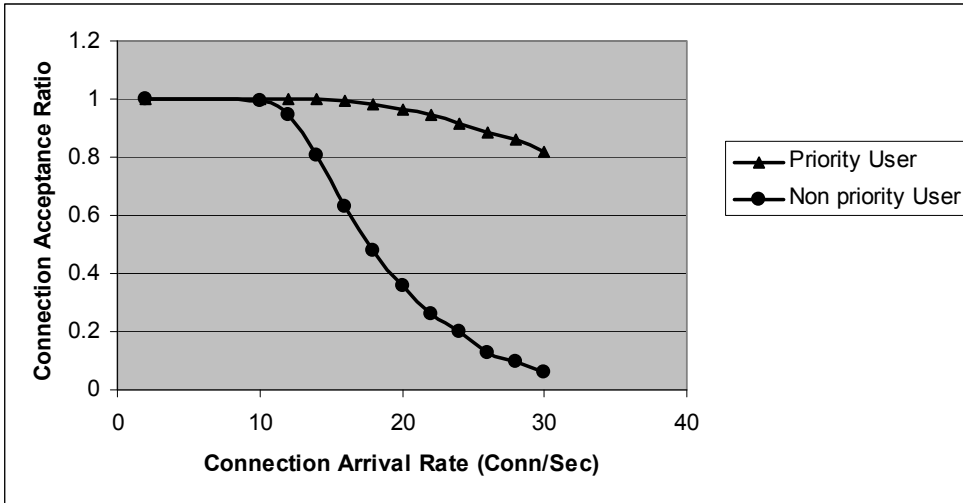


Fig. 17. Connection acceptance ratio for RTPS connections.

From Fig. 17 we can see that till the connection arrival rate is 10 connections/sec, there is enough capacity to accept both priority and non-priority connections. But beyond that the network cannot support the connection. So it starts to reject the connection. Since connection requests from priority users are processed first, the acceptance ratio for priority users would be higher compared to non-priority users.

Fig. 18 shows the simulation results for RTPS connections for different uplink capacities.

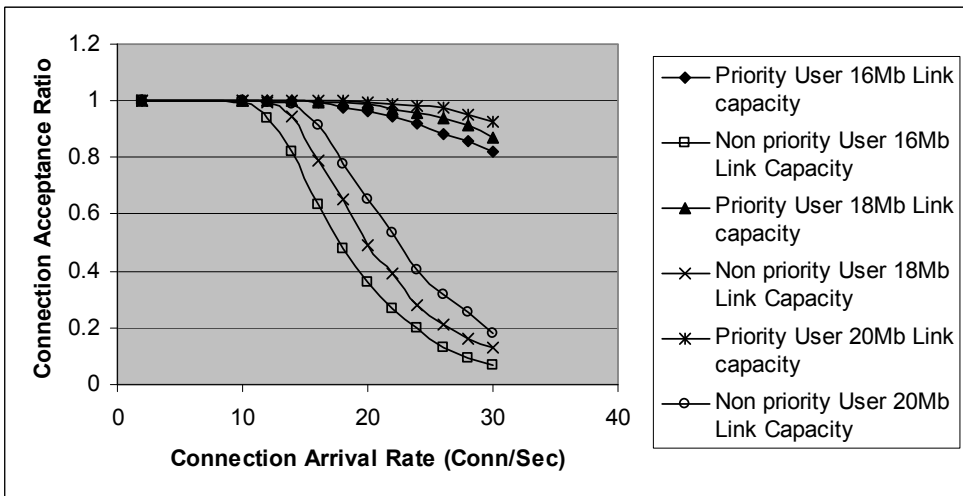


Fig. 18. Connection acceptance ratio for RTPS connection at different uplink capacities.

Simulations were also carried out to check the performance of user based admission control algorithm when BS receives connection requests for all the types of service classes i.e UGS, RTPS, nRTPS, eRTPS and BE. Fig 19. illustrates the simulation results for this scenario.

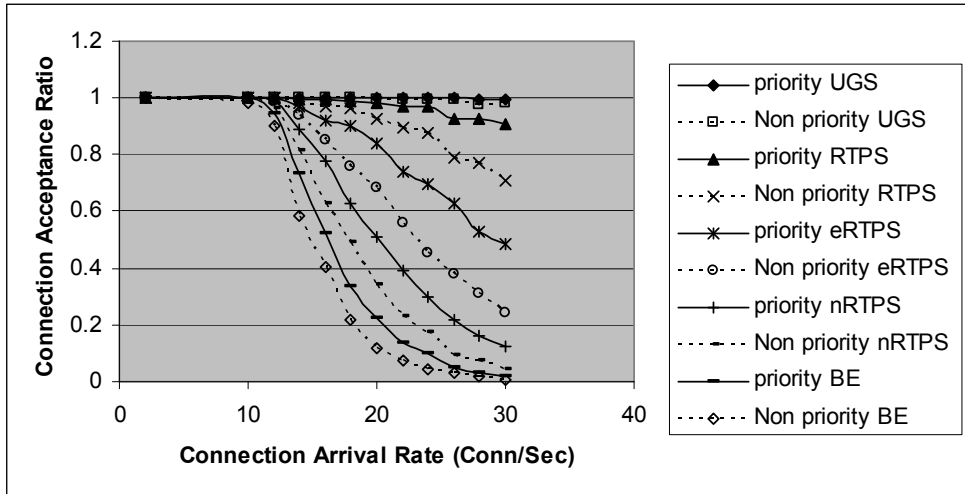


Fig. 19. Connection acceptance ratio for connections from different service classes.

From the simulation results it is clear that implementing user based admission control improves the connection acceptance ratio for priority users, there by improving the broadband experience for this section of users.

4.3 Drawback of user admission control

If, at any point in time, the network receives many connection requests from priority users then there is a chance that the non-priority users might see higher rejections of their connections. This can be tackled at the operator level. Based on the capacity of the network, a network service provider can limit the number of priority users that he can support. So when signing a new user, the operator can decide whether he wishes to provide the user the privilege of being a priority user.

5. Conclusion

In this chapter a comprehensive user based framework is proposed across various modules in WiMAX. Though operator can provide graded services by having different data rates at different price points, it does not give the flexibility that user based framework provides. Using the user based framework, graded services can be managed at the MAC layer and users can be up-graded/down-graded dynamically.

6. References

Chandra S. & Sahoo A. (2007), An efficient call admission control for IEEE802.16 networks, in *Proceedings of the 15th IEEE Workshop on Local and Metropolitan Area Networks*, pp. 188-193, ISBN 1-4244-1100-9, Princeton, NJ, USA, June 2007

- Chen J.; Jiao W. & Wang H. (2005), A Service Flow Management Strategy for IEEE 802.16 Broadband Wireless Access Systems in TDD Mode, *IEEE International conference on Communication*, pp. 3422 - 3426, ISBN 0-7803-8938-7, May 16-20 2005
- Chen Y. H. (2008), Light WiMAX Module, Available from <http://code.google.com/p/lwx/>
- Chiang C.H.;Liao W.; & T. Liu (2007), Adaptive Downlink/Uplink Bandwidth Allocation in IEEE 802.16 (WiMAX) Wireless Networks: A Cross-Layer Approach, *IEEE Global Telecommunications Conference*, pp. 4775-4779, ISBN 978-1-4244-1043-9, Washington DC, US, Nov. 26-30 2007
- IEEE (2004). IEEE 802.16, *Air Interface for Fixed Broadband Wireless Access Systems*, ISBN 0-7381-4070-8
- IEEE (2005). IEEE 802.16e, *Amendment to IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems - Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands*
- Kumar N.; Murthy K. N. B. & Lagare A. M. (2011), DBAM: Novel User Based Bandwidth Allocation Mechanism in WiMAX, *2nd International Conference on Recent Trends in Information, Telecommunication and Computing*, pp. 229-236, ISBN 978-3-642-19541-9, March 2011
- Kumar N.; Murthy K. N. B. & Lagare A. M. (2011), User oriented Network aware bandwidth allocation in wimax, *International Journal on Recent Trends in Engineering and Technology*, vol. 5, no. 1, pp 8-14, ISSN 2158-5555, March 2011
- Lagare A.M.; Das D. (2009). Novel user-based packet classification in 802.16e to provide better performance, *IEEE 3rd International Symposium on Advanced Networks and Telecommunication Systems (ANTS)*, pp. 1 - 3, ISSN: 2153-1676, Dec 2009
- Lee N.; Choi Y.; Lee S. & Kim N. (2010), A new CDMA-based bandwidth request method for IEEE 802.16 OFDMA/TDD systems, *IEEE Communications Letters*, Vol 14, Iss 2, pp 124-126, ISSN 1089-7798, Feb 2010
- Lin Y.N.; Wu C.W.; Lin Y.D. & Lai Y.C. (2008), A Latency and Modulation Aware Bandwidth Allocation Algorithm for WiMAX Base Stations, *IEEE Wireless Communications and Networking Conference*, pp. 1408-1413, ISBN 978-1-4244-1997-5, Las Vegas, Nevada, US, March 31 - April 3 2008
- Liu C.Y. & Chen Y.C. (2008), An Adaptive Bandwidth Request Scheme for QoS Support in WiMAX Polling Services, *Proceedings of The 28th International Conference on Distributed Computing Systems Workshops*, pp. 60-65, ISBN 978-0-7695-3173-1, Beijing , China, June. 17-20 2008
- ns2, Available from <http://www.isi.edu/nsnam/ns/>
- Park E.C. (2009), Efficient Uplink Bandwidth Request with Delay Regulation for Real-Time Service in Mobile WiMAX Networks, *IEEE Transaction on Mobile Computing*, Vol. 8, Iss. 9, pp. 1235-1249, ISSN 1536-1233, Sept. 2009
- Peng Z.; Guangxi Z.; Haibin S. & Hongzhi L. (2007), A Novel Bandwidth Scheduling Strategy for IEEE 802.16 Broadband Wireless Networks, *Proceedings of International Conference on Wireless Communications, Networking and Mobile Computing*, pp. 2000-2003, ISBN 978-1-4244-1311-9, Shanghai, China, Sept. 21-25, 2007
- Rong B.; Qian Y. & Chen H.H. (2007), Adaptive power allocation and call admission control in multiservice WiMAX access networks, *IEEE Wireless Communications*, Vol. 14, Iss. 1, pp. 14-19, ISSN 1536-1284, Feb 2007

- Shu'aibu D. S.; Syed-Yusof S. K. & Fisal N. (2010), Partition-base bandwidth managements for mobile WiMAX IEEE802.16e, *International Review on Computers and Software*, vol. 5, no. 4, pp. 445-452
- Shu'aibu D. S. (2011), Fuzzy Logic Partition-Based Call Admission Control for Mobile WiMAX, *ISRN Communications and Networking*, Article ID 171760, ISSN 2090-4355, April 11, 2011
- Tao S. Z. & Gani A. (2009), Intelligent Uplink Bandwidth Allocation Based on PMP Mode for WiMAX, *Proceedings of the 2009 International Conference on Computer Technology and Development*, pp. 86-90, ISBN 978-0-7695-3892-1, Kuala Lumpur, Malaysia, Nov. 13-15 2009
- Vaughan-Nichols, S. J. (2004), Achieving Wireless Broadband with WiMAX, *IEEE Computer*, Vol. 37 Iss. 6 , pp. 10-13, ISSN: 0018-9162
- Wee K. K. & Lee S. W. (2009), Priority based bandwidth allocation scheme for WIMAX systems, *Proceedings of 2nd IEEE International Conference on Broadband Network & Multimedia Technology*, pp. 15-18, ISBN 978-1-4244-4590-5, Cyberjaya, Malaysia, Oct. 18-20 2009

Introduction to the Retransmission Scheme Under Cooperative Diversity in Wireless Networks

Yao-Liang Chung¹ and Zsehong Tsai²
*Graduate Institute of Communication Engineering,
National Taiwan University Taipei
Taiwan, R.O.C.*

1. Introduction

As implied by the word “Cooperative Diversity (CD),” mobile users in a multi-user environment can share their antennas in a manner that creates a virtual Multiple-Input Multiple-Output (MIMO) system, which can be conceptually viewed as a multichannel transmission environment in the network layer, to achieve individual or common purposes of those users. By employing CD for transmissions, the quality and reliability of users’ data in wireless networks can thus be improved, mainly owing to the reason that the effect of wireless channel fading can be reduced. In this chapter, we aim to introduce existing representative retransmission schemes under various environments and further present a novel packet retransmission scheme for Quality-of-Service (QoS)-constrained applications in a general CD environment.

Transmit diversity of MIMO systems is an important technique which can bring significant gain to wireless systems with multiple transmit antennas. This technique is clearly advantageous to be employed on a cellular base station; however, it may not be practical for other scenarios. To be more specific, due to size, cost, or hardware limitations, small handsets/cellular phones may not be able to support certain types of multiple transmit antennas. For example, the size of an antenna must be several times the wavelength of the carrier frequency. Therefore, the use of multiple antennas is not an attractive way to achieve the transmit/receiving diversity in small handsets/cellular phones. To overcome such a naturally fundamental problem, CD is in nature an effective strategy to allow a single-antenna mobile device to achieve the benefit of MIMO systems with the help of cooperative mobile devices.

CD, which is a form of spatial diversity, is through cooperating users’ (usually called partners) relaying signals to the destination. This technique is achieved without the use of additional antennas of any user. That is to say, the antennas of the sender and partners together form a multiple-transmit antenna situation. Basically, the relay mechanism can be decode-and-forward or amplify-and-forward. Moreover, CD is an emerging and powerful technique that can mitigate fading and improve robustness to interference in wireless environments. Thus, CD becomes a promising candidate for emulating MIMO systems.

Recently, many research groups have turned their attention to the CD-related topics. Individual aspects of these problems have been considered, for example, in various papers [1-4]. In [1], Mahinthan *et al.* proposed a Quadrature Signaling (QS) mechanism in the CD system for transmissions. CD transmissions considering issues related to power allocation algorithms were explored by Mahinthan *et al.* in [2-3]. In [4], Chen *et al.* exploited that the use of space-time block coding in the multi-user CD to improve the performance of the transmission in wireless local area networks. Other abounding literature survey and investigation regarding the issue related to CD including principles and applications can be referred to Ray Liu *et al.* in [5] and Fitzek *et al.* in [6], respectively. Recently, a simple method to evaluate the performance of complex networks under CD using sampling property of a delta function was proposed by Jang in [7].

However, because of fundamental physical characteristics of wireless channels, data packets often cannot be delivered to the destination successfully. As a result, the design focusing on the efficient retransmission scheme under such a CD environment still plays a highly crucial role. Due to the evolution of the communication technology, most packet retransmission schemes under CD in literatures were based on the rich results from those retransmission schemes on point-to-point transmissions. Thus, we will first provide an overview of retransmission schemes on point-to-point transmissions, and then, investigate the issue on the retransmission scheme under the CD environment.

While there have been many papers exploring various retransmission schemes in the CD environment, there were no elaborations on the issue considering the time constraint for delay-sensitive services. Consequently, in such a CD environment their throughput formulas did not reflect the effective throughput (goodput) that must satisfy the typical delay constraints of streaming-type or real-time multimedia flows. Motivated by the above point, we therefore pay our attention to design a novel fast packet retransmission scheme to be employed in a general CD environment for delay-sensitive flows as a case study.

The rest of this chapter is organized as follows. A survey of various retransmission schemes is included in Section 2. Next, Section 3 proposes a novel fast packet retransmission scheme in a general CD environment for delay-sensitive applications as a case study. Section 4 makes a summary of this chapter and suggests the future work of interest. Finally, the list of references is provided in the end.

2. A survey of various retransmission schemes

The traditional retransmission scheme designed to combat the loss of transmission data for single-radio single-channel environments was first introduced by Lin *et al.* in [8]. Thanks to the advances of multi-band radio technologies, for many broadband wireless systems and short range communication networks, there may be many communication channels available to use. Consequently, it is natural to arrange the link layer packets to be transmitted over multiple channels to boost bandwidth. Issues regarding how to design and analyze the multi-channel transmission schemes have recently become an important research direction. A vast amount of research groups have thus started to pay attention to related topics. Individual aspects of these problems have been separately considered in many related papers [9-18]. Most literatures in this area can be conceptually categorized into 2 research directions; one is the single-radio multi-channel transmission discussed in [9-11],

and the other direction is the multi-radio multi-channel transmission discussed in [12-18]. Meanwhile, for the single-radio single-channel cases, the performance results of the Automatic-Repeat-reQuest (ARQ) retransmission schemes based on Markov analysis were available in [19-22], and optimized in both the power and the packet drop probability aspects respectively were recently studied in [23]. These studies [9-23] together provide a basis for the analysis and comparison of the multi-channel transmission. Furthermore, various kinds of ARQ and Hybrid ARQ schemes designed to be employed for CD environments were explored in [24-27]. The design approaches of these related research works are elaborated as follows.

In [9], a protocol was proposed to enable hosts to utilize multiple channels by switching channels dynamically, and their simulation showed that the effective throughput was improved, especially when the network was highly congested. Centralized and distributed algorithms to perform efficient channel assignments in component-based approach were proposed and implemented in [10]. A joint multi-channel and multi-path control protocol was proposed in [11], where it combined multi-channel link layer with multi-path routing, and simulations showed that the scheme can improve the throughput significantly.

The uses of switchable interfaces and multi-channel routing were proposed in [12], and simulation results showed that the throughput in a wireless ad hoc network can be improved. In [13], a hybrid channel assignment scheme was modeled into an integer linear programming formulation, and an approximation method for simplification was also studied. In the ad hoc network related areas, [14] maximized the network throughput subject to fairness requirements using the proposed wireless network coding schemes for a variety of multi-radio multi-channel environments with different routing strategies.

Several works focusing on the ARQ schemes in multi-radio multi-channel transmissions have been discussed in [15-18]. Formulas of the link layer throughput for the Stop-and-Wait (SW), the Go-Back-N (GBN) or the Selective-Repeat (SR) schemes were derived under different assumptions of channel characteristics. The multi-channel SR ARQ scheme in [15] assumed equal transmission rate and allowed the transmitter dynamically to assign the retransmission link packet to the channel using the link packet error probability as the selection criterion, and retransmissions can continue till its success. In [16], both throughput and delay performances of the multi-channel SW, GBN, and SR ARQ schemes were investigated and validated via simulations, while all channels were assumed statistically independent and identical. Throughput analysis of the multi-channel SW, GBN, and SR ARQ schemes were generalized in [17], where the generalization took the form of packet-to-channel assignment rules, and radio channels can be with different transmission rates and different link packet error probabilities. Two fast ARQ/HARQ packet retransmission schemes have been proposed to transport delay-sensitive flows in a multi-radio multi-channel environment in [18], where they can incorporate various retransmission policies, which are adjusted by the channel signal-to noise ratio (SNR) and the APDU size.

Closed form equations of the service data unit delay under the SR ARQ scheme were successfully derived and validated via simulation in [19]. An exact Markov model proposed to evaluate the delay statistics of the link packet for the SR ARQ scheme was available in [20]. The queuing models using dynamic link adaptation for the GBN and the SR ARQ schemes were formulated in [21], where the exact queue length and the delay statistics were

obtained. For HARQ schemes, a Markov model was presented to analyze the SR truncated type II HARQ scheme employing Reed Solomon linear erasure block codes in [22], where the link packet throughput, error probability, and delay performance were analyzed. In [23], a suboptimal root-finding solution was developed to solve the exhaustive search for the optimization problem formulated based on the incremental-redundancy HARQ scheme.

The delay performance of several truncated ARQ and HARQ schemes in a CD environment under the assumption of Poisson arriving packets were evaluated in detail by Boujemâa in [24]. An analytical model to quantify end-to-end performances for a CD ARQ scheme in a cluster-based multi-hop wireless network was proposed by Le *et al.* in [25]. Markov models developed to evaluate the CD system were also investigated by Mahinthan *et al.* in [26] and by Issariyakul *et al.* in [27], respectively.

While papers [24-27] have widely explored various ARQ/HARQ schemes in the CD environments, the issues regarding time constraints for delay-sensitive flows were not addressed and elaborated on. Therefore, their throughput formulas did not reflect the effective throughput that must satisfy the typical delay constraints of streaming-type or real-time multimedia flows in such an environment.

Due to the aforementioned reasons, we herein propose a novel fast packet retransmission scheme, where a new approach of retransmission strategy is designed and appropriately combines the encoding/decoding mechanism presented in [18], in such a CD environment for delay-sensitive flows as a case study. In the proposed scheme, there are 2 retransmission policies that can be employed adaptively according to both the channel quality and the Application layer Protocol Data Unit (APDU) size. The retransmission is designed to be allowed only one time. Here, APDU flows in the sender are further assumed to always have a link packet ready for transmission. As a result, it is not much meaningful to analyze the packet delay involving the queueing analysis. In this paper, we only focus on the complete throughput analysis to gain the main insight of optimizing the number of channels for retransmission between the 2 proposed retransmission policies under such the CD environment. All of the derived formulas are then verified via simulations. The effective throughput of our proposed scheme is shown better than that of other CD retransmission schemes (such as [26]) and non-CD retransmission schemes.

3. Case study: On the effective throughput gain of cooperative diversity with a fast retransmission scheme for delay-sensitive flows [33-34]

3.1 System description

3.1.1 Cooperative diversity system

A general CD system model composing of a sender, a partner, and a receiver is considered, as shown in Fig. 1, where two cooperative users (i.e., sender and partner) transmit their information to the same destination (i.e., receiver). It is assumed that each user' device in this system only has one radio transceiver. Additionally, Orthogonal Frequency Division Multiplexing (OFDM) is employed as the underlying transmission technique.

In the present system, channels among sender, partner and receiver are modeled as non-identical but independent Nakagami- m slow-fading channels corrupted by additive white Gaussian noise. The fading channels and the noise are assumed to be independent of each other.

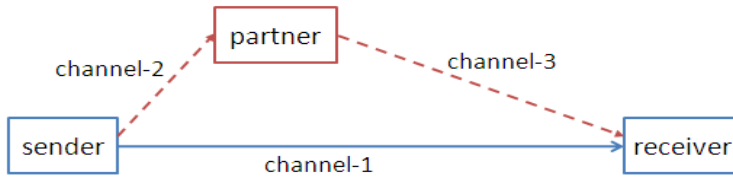


Fig. 1. System model for two cooperative users (sender and partner) transmission.

Generally speaking, for the transmission of flows of the sender, application layer flows are composed of APDUs. We assume that an APDU consists of s link packets. Each link packet will be encoded (described in detail in Section 3.1.3) in sequence for transmission. Here, APDU flows in the sender are assumed to always have a link packet ready for transmission.

For the convenience of the following analysis, channels that between sender and destination, between sender and partner, and between partner and destination are denoted as channel- j , $j=1,2,3$, respectively.

Last, but not least, we herein choose to employ only 1 partner for study since the significant improvement of the overall system performance with CD is usually owing to the contribution of the best partner [3].

3.1.2 Principles of fast retransmission strategy

The design philosophy of the retransmission strategy is to improve the application layer throughput while the effective control of the transmission delay is also assured. We assume that the underlying coding scheme is HARQ and that if a packet is retransmitted, then only its complementary packet is sent.

The packet retransmission strategy can be described via the following 4 principles:

- A link packet will be duplicated a copy in the sender buffer before its first transmission. When the sender begins to transmit this link packet, it will be broadcasted to the receiver and the partner.
- When an original link packet is transmitted to the receiver, an acknowledgment (ACK) or Negative ACK (NACK) packet (assumed error-free) will be sent to both the sender and the partner.
- There are 2 retransmission policies designed for the retransmission, indexed as $policy_k$: for $k=0, 1$. If a NACK is received, a complementary link packet will be retransmitted only one time via the partner (indexed as $policy_0$) or via both the sender and the partner (indexed as $policy_1$).
- The best retransmission policy is selected based upon both the APDU length and the expected long term link packet error probability among sender, partner and receiver, using the average application throughput as the performance objective. Under different APDU sizes and different link packet error probabilities, the corresponding best retransmission policy can be different.

Since an OFDM system is assumed, we assume each channel only uses a subset of OFDM subchannels. For the above retransmission strategy, note that both channel-1 and channel-2

are with orthogonal subchannel set 1, while channel-3 is with orthogonal subchannel set 2. The intersection of subchannel set 1 and set 2 is arranged to be an empty set; therefore, for the receiver, the signals from channel-1 and channel-3 will not interfere with each other.

Typical retransmission operations under *policy_0* and *policy_1* are illustrated in Fig. 2 and Fig. 3, respectively. Also, we assume that the delay threshold of the considered delay-sensitive APDU is set equal to the maximum of maximum delays for 2 policies. Notice that in this model the terminology *delay* only indicates the air-transmission delay component for the APDU under the proposed retransmission principle.

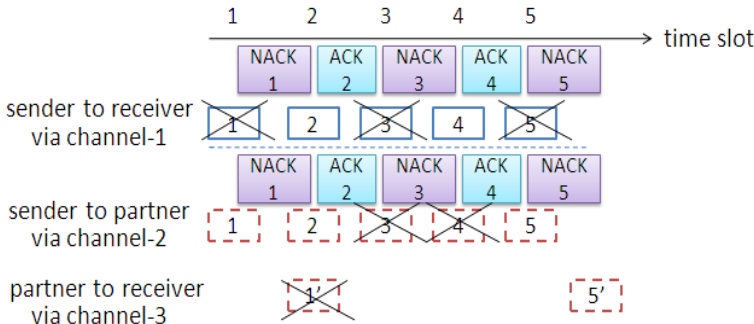


Fig. 2. A typical example of a fast HARQ with *policy_0*. When any original link packet is found failed at the receiver, only the partner will retransmit a complementary link packet if a link packet is successfully received by the partner. Link packet *i'* means the complementary packet of link packet *i*.

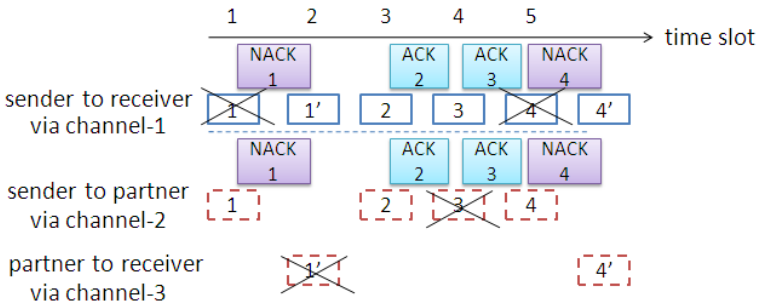


Fig. 3. A typical example of fast HARQ with *policy_1*. When any original link packet is found failed in the receiver, both the sender and the partner will retransmit a complementary link packet. Link packet *i'* is the complementary packet of link packet *i*.

3.1.3 Cooperative diversity with fast HARQ scheme

Two codes, a block code C_0 and a convolutional code C_1 , are together used as the coding mechanism employed in each user (see [29], [36] for examples). For more detail description, C_1 is a rate-1/2 convolutional code with constraint length c consisting of a c -stage shift

register and two generator polynomials $G_1(x)$ and $G_2(x)$. This code C_1 is used as an inner code for error detection and correction. Next, the outer code C_0 is a high rate $(n-(c-1), n-(c-1)-r)$ block code used for error detection only, where n is the length of each link packet that is transmitted in this scheme, and r is the length of parity-check bits for error detection.

When an APDU arriving at the sender, there will be s new information sequences $I_i(x)$, $1 \leq i \leq s$, each length $(n-(c-1)-r)$, generated in sequence. They are encoded into $J_i(x)$ with C_0 and then encoded into $V_i(x) = J_i(x)G_1(x)$ with C_1 in sequence. Each of them will be broadcasted in sequence only one time to the partner and the receiver.

Let $\tilde{V}_i(x)$ and $\hat{V}_i(x)$ be the noisy versions of $V_i(x)$ arriving at the receiver and the partner, respectively. For the receiver, the syndrome of $\tilde{V}_i(x)$ is checked in two steps. In step_1, $\tilde{V}_i(x)$ is regarded as a noisy version of a codeword in the $(n, n-(c-1))$ shortened cyclic code generated by $G_1(x)$. In step_2, an estimate $\tilde{J}_i(x)$ of $J_i(x)$ is then checked in the high rate $(n-(c-1), n-(c-1)-r)$ block code. If the syndromes are all zero in any step, an estimate $\tilde{I}_i(x)$ of $I_i(x)$ is obtained and delivered to the *receiving buffer*, which is a buffer used for waiting other link packets back for an APDU. Subsequently, an ACK packet will be sent to the sender and the partner. However, if the aforementioned syndrome check in any step is not zero, $\tilde{V}_i(x)$ is stored in the receiver buffer, and a NACK packet will be transmitted to the partner and the sender for possible retransmission. For the partner, if $\hat{V}_i(x)$ is error-free, then either *policy_0* or *policy_1* is adopted; otherwise, it is directly dropped.

Under *policy_0*, if $\hat{V}_i(x)$ is error-free, $\tilde{V}_i(x)$ will be decoded to $V_i(x)$ and then re-encode via $G_2(x)$ to $V'_i(x)$. Thereafter, $V'_i(x)$ is transmitted to the receiver. Let the noisy version of $V'_i(x)$ be denoted as $\tilde{V}'_i(x)$. $\tilde{V}'_i(x)$ will be checked in the same way as in the first transmission. If $\tilde{V}'_i(x)$ is found failed, $\tilde{V}'_i(x)$ shall be combined with $\tilde{V}_i(x)$, to form a combined codeword, which is decoded by the Viterbi decoding. The result is checked in the high rate $(n-(c-1), n-(c-1)-r)$ block code. If the syndrome is zero, it is claimed as a correct result. Its information sequence is then estimated and delivered to the *receiving buffer*; otherwise, it is discarded and the retransmission for this link packet is stopped.

Under *policy_1*, a complementary link packet $V'_i(x)$ via $G_2(x)$ of $V_i(x)$ will be transmitted from the sender to the receiver, and let the noisy version be denoted as $\tilde{V}'_i(x)$. Meanwhile, if $\hat{V}_i(x)$ is error-free, $\tilde{V}_i(x)$ will be decoded to $V_i(x)$ and then re-encode via $G_2(x)$ to $V'_i(x)$. Let the noisy version of $V'_i(x)$ be denoted as $\tilde{V}'_i(x)$. Following that, $\tilde{V}_i(x)$ and $\tilde{V}'_i(x)$ will be checked via the two-step decoding procedure, respectively. If the syndrome of any one is zero, it is claimed as a correct result, its information sequence is then estimated and delivered to the *receiving buffer*; otherwise, $\tilde{V}_i(x)$ shall be combined with $\tilde{V}'_i(x)$ and $\tilde{V}_i(x)$, respectively, to form two combined codewords, which are decoded by the Viterbi decoding. If any result is successful, its information sequence is then estimated and delivered to the *receiving buffer*; otherwise, a new codeword, $\hat{V}_{i,mrc}(x)$, will be further generated based on the Maximal Ratio Combining (MRC) (see [31] for more details) technique via $\tilde{V}'_i(x)$ and $\tilde{V}_i(x)$. The syndrome of the new codeword is checked in the same concept of the two-step decoding procedure. If the result is successful, the estimated information sequence will be delivered to the *receiving buffer*; otherwise, it is discarded and the retransmission for this link packet is stopped.

3.2 Throughput analysis

Performances of the application layer throughput for the present scheme will be first analyzed in detail in this Section. It is assumed that the time axis is partitioned into equal size slot. In each time slot, it is separated into two parts. That is to say, the main part of a time slot is used for link packets transmissions and the rest of the time slot is reserved for ACK/NACK packets transmissions. Here, the SNR is assumed staying constant in a time slot. In addition, the M -ary, $M = 2^b$ where b is even, the Quadrature Amplitude Modulation (QAM) scheme is assumed in the OFDM subchannels of the proposed model.

3.2.1 Link packet error probability

For M -ary QAM in Nakagami- m slow-fading channels, the average BERs for channel- j , $j=1,2,3$, denoted as $\bar{\varepsilon}_j$, can be derived by

$$\bar{\varepsilon}_j = \int_0^\infty p_j(\gamma_j) \varepsilon_{ins,j}(\gamma_j) d\gamma_j, \quad j=1,2,3, \tag{1}$$

where, in channel- j , $\varepsilon_{ins,j}(\gamma_j)$ is the instantaneous BER conditional on γ_j for M -ary QAM, γ_j is the instantaneous SNR per bit, $\gamma_j > 0$, $p_j(\gamma_j) = m^m \gamma_j^{m-1} e^{-m\gamma_j/\bar{\gamma}_j} / \bar{\gamma}_j^m \Gamma(m)$ is the probability density function (pdf) of γ_j in Nakagami- m fading given in [28], $m \geq 1/2$, $\Gamma(\cdot)$ is the gamma function, and $\bar{\gamma}_j$ is the average SNR per bit. The instantaneous BER $\varepsilon_{ins,j}(\gamma_j)$ was previously derived in [30], [32] as

$$\varepsilon_{ins,j}(\gamma_j) = \frac{1}{\sqrt{M} \log_2 \sqrt{M}} \sum_{z=1}^{\log_2 \sqrt{M}} \sum_{t=0}^{f(z,M)} \left\{ \operatorname{erfc}((2t+1)\sqrt{g\gamma_j}) f(t,z,M) \right\}, \quad j=1,2,3, \tag{2}$$

where $f(z,M) = (1 - 2^{-z})\sqrt{M} - 1$, $g = 3\log_2 M / (2M - 2)$, $f(t,z,M) = (-1)^{\lfloor t2^{z-1}/\sqrt{M} \rfloor} \left(2^{z-1} - \lfloor t2^{z-1}/\sqrt{M} + 1/2 \rfloor \right)$, and $\operatorname{erfc}(\cdot)$ is the error function.

The average link packet error probability in a single transmission in channel- j , $j=1,2,3$, denoted as $\bar{P}_{j,e}$, can be given by

$$\bar{P}_{j,e} = \int_0^\infty p_j(\gamma_j) (1 - (1 - \varepsilon_{ins,j}(\gamma_j))^n) d\gamma_j, \quad j=1,2,3. \tag{3}$$

Furthermore, the average link packet error probability after the Viterbi decoding conditional on the event that both $\hat{V}'_i(x)$ and $\hat{V}_i(x)$ are corrupted, denoted as $\bar{P}_{f,0}$, can be approximately by (see eq. (28) in [29])

$$\bar{P}_{f,0} \cong 1 - (1 - p_b)^{n-(c-1)}, \tag{4}$$

where p_b is the corresponding bit error probability obtained via the Viterbi decoding. As shown in [29], p_b is bounded by

$$p_b \leq \frac{1}{2} \frac{\partial T(X, Y)}{\partial Y} \Big|_{X=2\sqrt{\varepsilon'(1-\varepsilon')}, Y=1}, \tag{5}$$

where ε' , the upper bound of the conditional BERs given that the two-step (mentioned in Section 3.1.3) decoding syndromes in channel-1 and channel-3 are non-zero, is given by

$$\varepsilon' = \max \left\{ \frac{\varepsilon_1}{1-(1-\varepsilon_1)^n}, \frac{(1-\varepsilon_2)\varepsilon_3}{1-(1-(1-\varepsilon_2)\varepsilon_3)^n} \right\}, \tag{6}$$

and $T(X, Y)$ is the generating function of the convolutional code. In addition, $\bar{P}_{f,1}$, the average link packet error probability after the Viterbi decoding conditional on the event that both $\tilde{V}_i(x)$ and $\tilde{V}_i(x)$ are corrupted, can be given by (4)-(6) together with ε' replaced by $\varepsilon = \varepsilon_1 / (1 - (1 - \varepsilon_1)^n)$, which is the conditional BER given that the two-step decoding syndrome in channel-1 is non-zero.

Last, but not least, the average link packet error probability after the MRC decoding, under *policy_1*, denoted as \bar{P}_{mrc} , can be given by

$$\bar{P}_{mrc} = \int_0^\infty \tilde{p}(\gamma_b) (1 - (\varepsilon_{mrc}(\gamma_b))^n) d\gamma_b, \tag{7}$$

where γ_b is the instantaneous SNR per bit at the output of the MRC decoder, $\tilde{p}(\gamma_b)$ represents the pdf of γ_b , $\gamma_b > 0$, and $\varepsilon_{mrc}(\gamma_b)$ is the instantaneous BER conditional on γ_b for M -ary QAM after the MRC decoding. According to [28], [30], $\tilde{p}(\gamma_b) = m^{2m} \gamma_b^{2m-1} e^{-m\gamma_b/\bar{\gamma}_c} / \bar{\gamma}_c^{2m} \Gamma(2m)$, where $\bar{\gamma}_c$ means the equivalent average SNR for each channel. The instantaneous BER $\varepsilon_{mrc}(\gamma_b)$ can be given by (2) with γ_j replaced by γ_b .

3.2.2 Throughput

For the fast HARQ scheme with *policy_0*, the application layer throughput in APDU/slot, denoted as T_0 , can be derived as

$$T_0 = \frac{1}{s} (1 - \bar{P}_{1,e} \bar{P}_{2,e} - \bar{P}_{1,e} (1 - \bar{P}_{2,e}) \bar{P}_{3,e} \bar{P}_{f,0})^s, \tag{8}$$

where $1/s$ represents the average number of the APDUs transported per slot, and the second term indicated the success probability of an APDU transmission.

Next, for the fast HARQ scheme with *policy_1*, the application layer throughput in APDU/slot, denoted as T_1 , can be derived as

$$T_1 = \frac{\alpha}{s} (1 - \bar{P}_{1,e} \bar{P}_{2,e} \bar{P}_{f,1} - \bar{P}_{1,e}^2 (1 - \bar{P}_{2,e}) \bar{P}_{3,e} \bar{P}_{f,0} \bar{P}_{f,1} \bar{P}_{mrc})^s, \tag{9}$$

where α/s represents the average number of the APDUs transported per slot, and similar to the concept in (8), the second term means the success probability of an APDU transmission. In (9), \bar{P}_{mrc} is the average link packet error probability after the MRC decoding conditional on $\hat{V}_i(x)$ and $\tilde{V}_i(x)$ all found failed. Since \bar{P}_{mrc} is the unconditional probability of a link packet error after the MRC decoding and the result will be correct after the MRC decoding as long as there is at least a link packet that is correct, \bar{P}_{mrc} can be derived as $\bar{P}_{mrc} = \bar{P}_{mrc} / (\bar{P}_{1,e}^2(1 - \bar{P}_{2,e})\bar{P}_{3,e})$ by the definition of conditional probability [35]. Moreover, α in (9) can be obtained via the equality

$$\bar{P}_{1,e}\alpha = 1 - \alpha, \quad (10)$$

since the average number of retransmission link packets generated per slot should equal the average number of retransmission completed, after normalization.

Last, but not least, for delay-sensitive flows, the maximum air-transmission delay of an APDU allowed is usually subject to a specific QoS requirement. In this case, based on the similar derivation and argument in [18], one can appropriately tune the key parameter, namely, s , in the system to achieve the highest effective throughput under a given delay constraint.

3.3 Analytical and simulation results

In this section, the considered CD environment with a sender, a partner, and a receiver remains the same as shown in Fig. 1. We assume that an APDU is composed of 5 link packets. A 16 QAM modulation scheme is adopted. The coding mechanism is referred to Section 3.1.3. Also, we set $r = 6$ bytes, $c = 9$, and $n = 257$ bytes. The ACK/NACK packet size for ARQ related schemes is set equal to 25 bytes and for HARQ related schemes is set equal to 26 bytes. The link speed is set equal to 10Mbps. Besides, excluding the error-correcting codes, the ratio of the additional header overhead associated with the lower layer protocols from the application one is set equal to 0.04.

First, we will evaluate and compare the performance results among all schemes to see main potential insights of our proposed scheme by considering the ideal case that the channel between the sender and the partner is error-free. Next, we further investigate the impact on the system performance when there is an error probability on the channel between the sender and the partner.

3.3.1 With an error-free channel-2

For the fast retransmission scheme, based on (8)-(9), analytical results of application throughputs under $\bar{\varepsilon}_1$, with $\bar{P}_{3,e} = 0.9$ and $\bar{P}_{3,e} = 0.1$, in the Nakagami-3 slow-fading environment, are depicted in Fig. 4 and Fig. 5, respectively. In Fig. 4, it can be found that if $\bar{\varepsilon}_1 \leq 0.4 \times 10^{-2}$, the optimal throughput can be achieved with only 1 channel for retransmission (via the partner); if $\bar{\varepsilon}_1 \geq 0.4 \times 10^{-2}$, it can be achieved by parallel retransmissions via 2 channels (via both sender and partner). However, in Fig. 5, it is seen

that the throughput of *policy_0* is always better than that of *policy_1*. Because the average link packet error probability of channel-3 is small, the retransmission of duplicated link packet on channel-1 via *policy_1* will waste bandwidth.

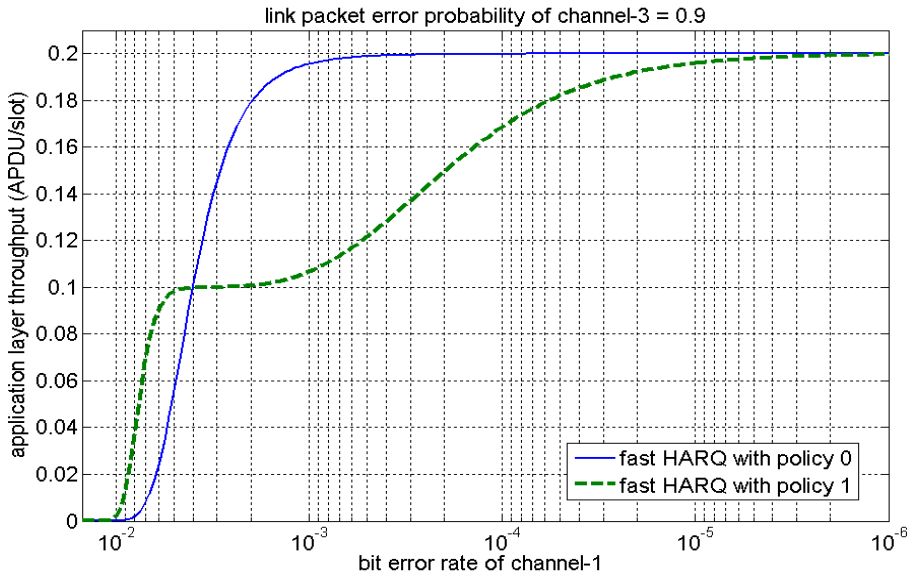


Fig. 4. Application throughputs under typical $\bar{\varepsilon}_1$ with $\bar{P}_{3,e} = 0.9$ under a fast HARQ scheme.

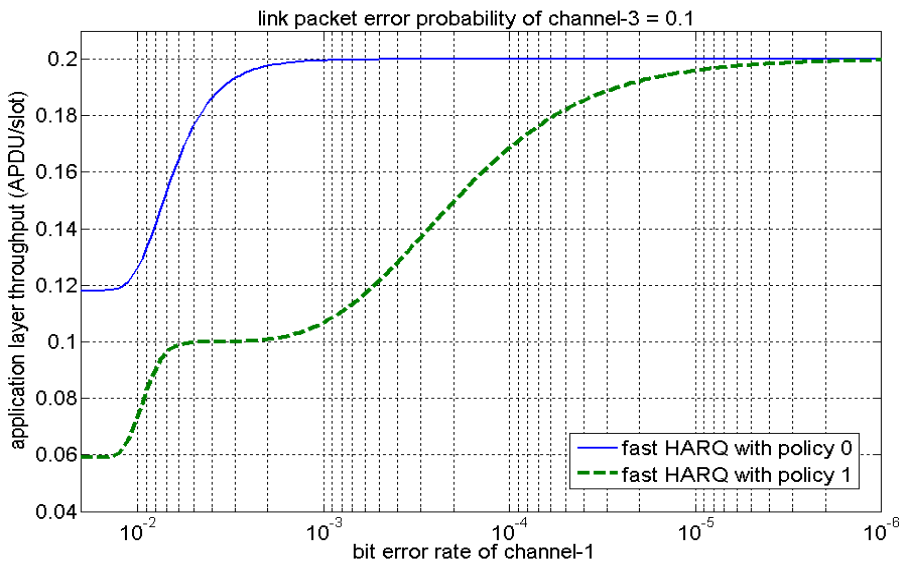


Fig. 5. Application throughputs under typical $\bar{\varepsilon}_1$ with $\bar{P}_{3,e} = 0.1$ under a fast HARQ scheme.

In what follows, we compare the throughput of our proposed scheme with that of the previous work [26] and non-CD HARQ scheme under $\bar{\epsilon}_1$, with $\bar{\epsilon}_3 = 10^{-3}$ in Nakagami-3 slow-fading channels, as shown in Fig. 6. Notice that in Fig. 6, the *CD with optimized fast HARQ* scheme represents the case that the retransmission policy is adaptively adjusted to be optimal on the basis of the channel quality in the CD environment. For a fair comparison among all schemes, all throughput results are in bit/second, and other 2 schemes are modified to allow only 1 retransmission and time slots for those discarded retransmissions are then used for new transmissions. Notice due to this modification, their throughput formulas are modified versions of (9) with the unused parameters removed. In details, one should set $\bar{P}_{f,1} = 1$ and replace the parameter n in (3) by $n-(c-1)$ for the scheme in [26], and set $\bar{P}_{2,e} = 1$ for the non-CD HARQ scheme. With the help of Fig. 6, it can be found that better performance is achieved by the optimized fast HARQ scheme except when $\bar{\epsilon}_1$ is extremely small due to the additional overhead of the HARQ.

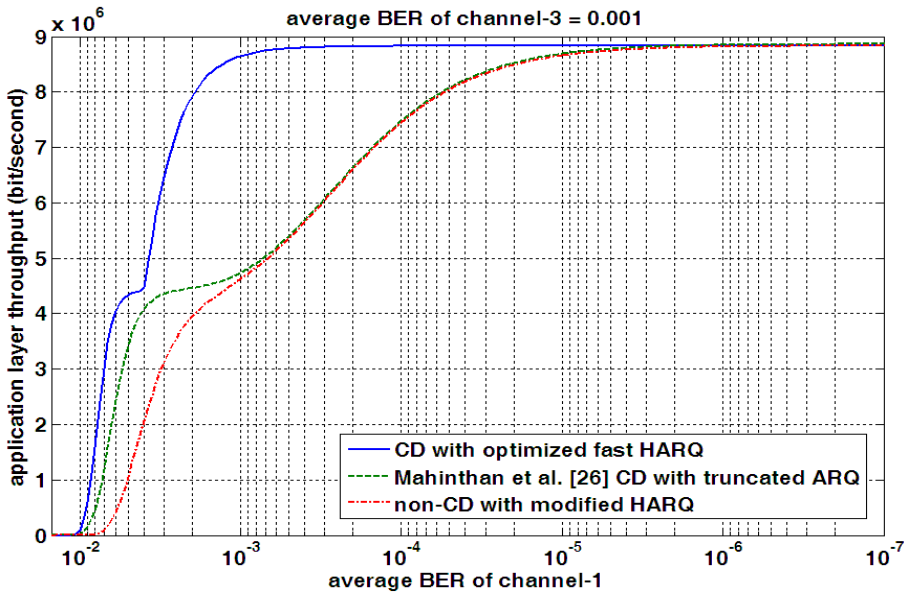


Fig. 6. Application layer throughput (in bit/sec.) comparison among various schemes under Nakagami-3 slow-fading channels when channel-2 is error-free.

Notice that, generally speaking, BER=0.001 is fairly high (in our parameter setting, which is about equal to the packet error rate =0.9 when without employing any error correcting mechanism), that is to say, the channel condition is extremely bad. Here, the reasons for setting channel-3's BER=0.001 are explained as follows. Although the sender would like to select the neighboring partner having good channel condition for helping transmission, in the worst case when the partner is far away from the receiver and both of them are at the edge of a cell such that the channel-3's condition degrades. In this case, with the validation of analytical and simulation results, the performance result of our scheme is much better than that of other schemes. It means our scheme is very powerful. Thus, it can be easily

reasoned that when channel-3's BER decreases, our scheme still remains the best although the performance results for these schemes will all be improved.

Furthermore, taking $\bar{\epsilon}_1 = 2 \times 10^{-3}$ and $\bar{\epsilon}_3 = 10^{-3}$ as an example, effective throughput performances of these schemes under different Nakagami- m , $m=1/2, 1, 3$, slow-fading channels in the CD environment are compared, as listed in Table 1. From Table 1, both analytical results based on (8)-(9) and simulation results show that the optimized fast HARQ scheme always achieves better throughput performance than other schemes since the optimized fast HARQ scheme can adaptively adjust the retransmission policy according to the channel quality. Again in Table 1, it is found that the analytical results are slightly lower than the simulation ones for both the optimized fast HARQ scheme and the non-CD HARQ scheme since the Viterbi decoding mechanism via (5) is employed for them. Note that the upper bound in (5) is tight and can be regarded as an excellent approximation when the BER is lower than 10^{-2} [36].

Environment \ Scheme		Effective Throughput $\times 10^6$	CD with optimized fast HARQ	[26] CD with truncated ARQ	Non-CD with modified HARQ
Nakagami-1/2	analytical result		7.765	4.308	3.704
	simulation result		7.853	4.356	3.767
Nakagami-1	analytical result		7.889	4.413	3.822
	simulation result		7.968	4.462	3.887
Nakagami-3	analytical result		7.997	4.512	3.998
	simulation result		8.077	4.567	4.068

Table 1. Comparisons of the application layer throughputs (in bit/second) at $\bar{\epsilon}_1 = 2 \times 10^{-3}$ and $\bar{\epsilon}_3 = 10^{-3}$ among various schemes under different Nakagami- m , $m=1/2, 1, 3$, slow-fading channels.

3.3.2 With a non-error-free channel-2

Due to the fundamental physical characteristics of wireless channels, there often exists an error probability for each transmission channel in the real-world environment. However, in order to take the advantage of CD, the sender usually selects the neighboring partner having good channel condition between them. Thus, we herein set $\bar{\epsilon}_2 = 10^{-4}$ for demonstrating performance results. The throughput comparisons of various schemes under $\bar{\epsilon}_1$, with $\bar{\epsilon}_2 = 10^{-4}$ and $\bar{\epsilon}_3 = 10^{-3}$ in Nakagami-3 slow-fading channels, are shown in Fig. 7.

It is found in Fig. 7 that the performance of the optimized fast HARQ scheme obviously degrades when the BER of channel-1 is smaller than 3×10^{-3} when compared with that in Fig. 6. Because there exists an error probability on channel-2 and *policy_0* only uses the cooperative path (i.e., channel-2 together with channel-3) for retransmissions, the power of *policy_0* decreases. However, the throughput result of the optimized fast HARQ scheme in Fig. 7 is also shown better than that of the other 2 schemes. In addition, it can be observed

that when $\bar{\varepsilon}_1 \geq 3 \times 10^{-3}$, the performance results of the first 2 good schemes are almost the same as those in Fig. 6 due to the fact that MRC is much powerful.

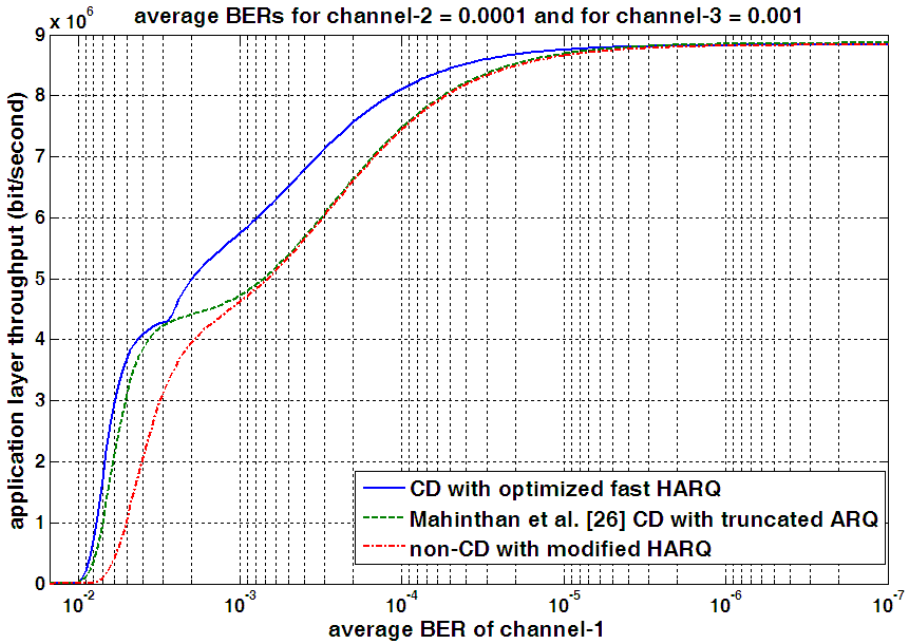


Fig. 7. Application layer throughput (in bit/sec.) comparison among various schemes under Nakagami-3 slow-fading channels when channel-2 is not error-free.

Last, for completeness, we take $\bar{\varepsilon}_1 = 2 \times 10^{-3}$, $\bar{\varepsilon}_2 = 10^{-4}$, and $\bar{\varepsilon}_3 = 10^{-3}$, as an example, to illustrate the effective throughput results for various schemes under different Nakagami- m , $m=1/2, 1, 3$, slow-fading channels, and summarize the results in Table 2. From Table 2, it can be found that both analytical results based on (8)-(9) and simulation results of the optimized fast HARQ scheme also always have better throughput results than those of other schemes as in Table 1. The results of the non-CD HARQ scheme for both Table 1 and Table 2 are the same since its performance only depends on channel-1's BER. We also notice that throughput improvement of our scheme is significant even with $\bar{\varepsilon}_2 = 10^{-4}$ in the sender-to-partner channel.

In summary, based on Figs. 6 and 7 and Tables 1 and 2, we can thus conclude that the fast HARQ scheme is an excellent approach for transporting delay-constrained streaming-type or real-time multimedia flows in CD environments even when there is an error probability on the cooperative path. It is for the reasons that the retransmission strategy can be adaptively adjusted according to the channel condition and that the decoding procedure involving MRC and the Viterbi decoding are appropriately designed.

Environment	Scheme	CD with optimized fast HARQ	[26] CD with truncated ARQ	Non-CD with modified HARQ
	Effective Throughput $\times 10^6$			
Nakagami-1/2	analytical result	4.805	4.258	3.704
	simulation result	4.848	4.258	3.767
Nakagami-1	analytical result	4.902	4.351	3.822
	simulation result	4.951	4.351	3.887
Nakagami-3	analytical result	5.017	4.456	3.998
	simulation result	5.067	4.456	4.068

Table 2. Comparisons of the application layer throughputs (in bit/second) at $\bar{\epsilon}_1 = 2 \times 10^{-3}$, $\bar{\epsilon}_2 = 10^{-4}$, $\bar{\epsilon}_3 = 10^{-3}$ among various schemes under different Nakagami- m , $m=1/2, 1, 3$, slow-fading channels.

4. Conclusions

A fast HARQ packet retransmission scheme has been successfully proposed to transport delay-sensitive flows in a general CD environment. The presented scheme incorporates 2 retransmission policies, and these 2 policies can be selected adaptively by the channel SNRs and the APDU sizes. In ideal conditions, the best retransmission policy can always be selected to achieve optimized performance.

In this case study, our cooperative fast retransmission scheme has been shown to be an excellent approach for improving the effective throughput in transporting delay-sensitive flows in CD environments. Numerical results verified via simulations, show that when optimized, the proposed scheme can achieve effective throughput much better than that of other ARQ schemes (such as [26]) and non-CD HARQ schemes, especially when the sender-to-partner channel condition is good. The performance improvement is still significant even when there is an error probability (e.g. $\text{BER} \leq 10^{-4}$) in the sender-to-partner channel. Moreover, in the aspect of the battery saving, the presented scheme should save much more power than that of other schemes due to the one-time retransmission design. It is thus concluded that the proposed fast HARQ retransmission scheme is an excellent ARQ candidate for the multimedia or real-time transport in CD environments, when the time-constraint is imposed.

5. Summaries and future works

The issues of improving fast retransmission schemes under CD environments can be essential to many delay-sensitive applications. This chapter has widely covered the conceptual description of many representative retransmission schemes under various environments and presented a novel fast packet retransmission scheme intended for effectively transporting delay-sensitive flows in a general CD environment. The presented retransmission scheme and other related works should have provided a sufficient collection of schemes and analysis methodologies for designing further wireless communication systems with similar requirements.

Furthermore, due to the fact that in most practical scenarios the terminals are battery-powered, the design of the energy-efficiency transmission satisfying the respectively specific QoS requirements of these users in the network is very crucial to prolonging the battery life of these terminals. Consequently, the issue concerning the energy consumption has been increasingly paid much attention. We suggest incorporating such a concern with the present work to design an efficiently power-saving fast packet retransmission scheme in a general CD environment for delay-sensitive flows in the future. Additionally, the well design of an efficient retransmission scheme to be employed in such a CD environment simultaneously considering the issue of effective throughput, QoS, fairness, complexity, and power saving is still an open issue for research.

6. Acknowledgements

The authors wish to express their sincere appreciation for financial support from the National Science Council of the Republic of China under Contracts NSC 98-2219-E-002-002 and NSC 99-2219-E-002-002.

7. References

- [1] V. Mahinthan, J. W. Mark, and X. Shen, "A cooperative diversity scheme based on quadrature signaling," *IEEE Trans. Wireless Commun.*, vol. 6, no. 1, pp. 41-45, January 2007.
- [2] V. Mahinthan *et al.*, "Maximizing cooperative diversity energy gain for wireless networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2530-2539, July 2007.
- [3] V. Mahinthan *et al.*, "Partner selection based on optimal power allocation in cooperative diversity systems," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 511-520, January 2008.
- [4] J. Chen and K. Djouani, "A multi-user cooperative diversity for wireless local area networks," *Int. J. Communications, Network and System Sciences*, vol. 3, pp. 207-283, 2008.
- [5] K. J. Ray Liu, A. K. Sadek, W. Su, and A. Kwasinski, *Cooperative communication and networking*, Cambridge University Press, 2009.
- [6] F. H. P. Fitzek and M. D. Katz, *Cooperation in wireless networks: principles and applications*, Springer, Netherlands, 2006.
- [7] W. M. Jang, "Quantifying performance of cooperative diversity using the sampling property of a delta function," *IEEE Trans. Wireless Commun.*, vol. 10, no. 7, pp. 2034-2039, July 2011.
- [8] S. Lin, *et al.*, "Automatic-repeat-request error control schemes," *IEEE Commun. Mag.*, pp. 5-16, December 1984.
- [9] J. So and N. H. Vaidya, "Multi-channel MAC for ad hoc networks: handling multi-channel hidden terminals using a single transceiver," in *Proc. ACM MobiHoc*, pp. 222-233, Roppongi Hills, Tokyo, Japan, May 2004.
- [10] R. Vedantham *et al.*, "Component based channel assignment in single-radio multi-channel ad hoc networks," in *Proc. ACM MobiCom*, pp. 378-389, Los Angeles, CA, USA, May 2006.
- [11] W. H. Tam and Y. C. Tseng, "Joint multi-channel link layer and multi-path routing design for wireless mesh networks," in *Proc. IEEE INFOCOM*, pp. 2081-2089, May 2007.

- [12] P. Kyasanur and N. H. Vaidya, "Routing and link-layer protocols for multi-channel multi-interface ad hoc wireless networks," *ACM Mobile Computing and Comm. Rev.*, pp. 31-43, vol. 10, no.1, January 2006.
- [13] A. K. Jeng and R. H. Jan, "Optimization on hybrid channel assignment for multi-channel multi-radio wireless mesh networks," in *Proc. IEEE GLOBECOM*, November 2007.
- [14] H. Su and X. Zhang, "Modeling throughput gain of network coding in multi-channel multi-radio wireless ad hoc networks," *IEEE Journal on Selected Areas in Commun.*, vol. 27, no. 5, June 2009.
- [15] N. Shacham and B. C. Shin, "A selective-repeat-ARQ protocol for parallel channels and its resequencing analysis," *IEEE Trans. Commun.*, vol. COM-40, pp. 773-782, April 1992.
- [16] J. F. Chang and T. H. Yang, "Multichannel ARQ protocols," *IEEE Trans. Commun.*, vol. COM-41, pp. 592-598, April 1993.
- [17] Z. Ding and M. Rice, "ARQ error control for parallel multichannel communications," *IEEE Trans. Wireless Commun.*, vol. COM-5, no. 11, pp. 3039-3044, November 2006.
- [18] Y.-L. Chung and Z. Tsai, "Performance analysis of two multichannel fast retransmission schemes for delay-sensitive flows," *IEEE Trans. Veh. Technol.*, vol. 59, no. 7, pp. 3468-3479, September 2010.
- [19] W. Luo, K. Balachandran, S. Nanda, and K. K. Chang, "Delay analysis of selective-repeat ARQ with applications to link adaptation in wireless packet data systems," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1017-1029, May 2005.
- [20] L. Badia, M. Rossi, and M. Zorzi, "SR ARQ packet delay statistics on Markov channels in the presence of variable arrival rate," *IEEE Trans. Wireless Commun.*, vol. 5, no. 7, pp. 1639-1644, July 2006.
- [21] L. B. Le, E. Hossain, and M. Zorzi, "Queueing analysis for GBN and SR ARQ protocols under dynamic radio link adaptation with non-zero feedback delay," *IEEE Trans. Wireless Commun.*, vol. 6, no. 9, pp. 3418-3428, September 2007.
- [22] L. Badia, M. Levorato, and M. Zorzi, "Markov analysis of selective repeat type II hybrid ARQ using block codes," *IEEE Trans. Commun.*, vol. 56, no. 9, pp. 1434-1441, September 2008.
- [23] T. V. K. Chaitanya and E. G. Larsson, "Outage-optimal power allocation for hybrid ARQ with incremental redundancy," *IEEE Trans. Wireless Commun.*, vol. 10, no. 7, pp. 2069-2074, July 2011.
- [24] H. Boujemâa, "Delay analysis of cooperative truncated HARQ with opportunistic relaying," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 4795-4803, November 2009.
- [25] L. Le and E. Hossain, "An analytical model for ARQ cooperative diversity in multi-hop wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1786-1791, May 2008.
- [26] V. Mahinthan *et al.*, "Cross-layer performance study of cooperative diversity system with ARQ," *IEEE Trans. Veh. Technol.*, vol. 58, no. 2, pp. 705-719, February 2009.
- [27] T. Issariyakul, and V. Krishnamurthy, "Amplify-and-forward cooperative diversity wireless networks: models, analysis, and monotonicity properties," *IEEE/ACM Trans. Networking*, vol. 17, no. 1, pp. 225-238, February 2009.
- [28] M. Nakagami, "The m-distribution - a general formula of intensity distribution of rapid fading," in *Statistical Methods of Radio Wave Propagation*, pp. 3-36, W. C. Hoffman Ed. Elmsford, Pergamon Press, New York, 1960.

- [29] L. R. Lugand, *et al.*, "Parity retransmission hybrid ARQ using rate- $\frac{1}{2}$ convolutional codes on a nonstationary channel," *IEEE Trans. Commun.*, Vol. 37, no. 7, pp. 755-765, July 1989.
- [30] M. S. Patterh, *et al.*, "BER performance of MQAM with L-branch MRC diversity reception over correlated Nakagami-m fading channels," *International Journal of Wireless Communications and Mobile Computing*, vol. 3, pp. 397-406, May 2003.
- [31] D. G. Brennan, "Linear diversity combining techniques," *Proc. IRE*, vol. 47, pp. 1075-1102, June 1959.
- [32] D. Yoon, K. Cho, and J. Lee, "Bit error probability of M-ary quadrature amplitude modulation," in *Proc. the IEEE VTC*, Boston, MA, September 2000.
- [33] Y.-L. Chung and Z. Tsai, "On the effective throughput gain of cooperative diversity with a fast retransmission scheme for delay-sensitive flows," *IEICE Trans. Commun.*, vol. E94-B, no.12, pp.-, December 2011.
- [34] Y.-L. Chung and Z. Tsai, "Cooperative diversity with fast HARQ for delay-sensitive flows," in *Proc. IEEE 71st Veh. Technol. Conf. (IEEE VTC)*, Taipei, Taiwan, May 2010.
- [35] A. Leon-Garcia, *Probability, statistics, and random processes for electrical engineering*, 3rd ed., Prentice Hall, 2008.
- [36] A. J. Viterbi and J. K. Omura, *Principles of digital communication and coding*, McGraw-Hall, 1979.

Intelligent Transport Systems: Co-Operative Systems (Vehicular Communications)

Panagiotis Lytrivis and Angelos Amditis
*Institute of Communication and
Computer Systems (ICCS)
Greece*

1. Introduction

The term *Intelligent Transport Systems* (ITS) is used to illustrate the application of information and communication technologies in the transport domain. The intention of ITS is to enhance road safety and traffic efficiency, minimize environmental impact and in general maximize the benefits for the road users (Zhou et al., 2010; Popescu-Zeletin et al., 2010; Hartenstein & Laberteaux, 2010).

In turn, *Co-operative Systems* are the most promising technology within the ITS framework. The word “co-operative” indicates that vehicles are collaborating with each other and with the infrastructure, exploiting wireless communications, in order to increase their awareness about the road environment. There are two types of communication in co-operative systems, namely vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I).

The scope of this chapter is to highlight the significant role of vehicular communications in future ITS. Standalone sensors and sensor systems can support the drivers in certain cases (e.g. maintain a safe speed and safe distance from the vehicle ahead, avoid a possible rear-end collision etc.) but are not sufficient enough. Vehicles exchanging real-time messages and sharing information about the perception of the road environment could significantly extend the benefits of the abovementioned standalone systems and also satisfy the requirements of a large number of applications (see Figure 1).

Over the past years significant efforts have been performed for the bandwidth allocation and the standardization of vehicular communications worldwide. The Federal Communication Commission (FCC) decided the allocation of a frequency spectrum for vehicular applications. In Europe under the European Commission Decision 676/2002/EC the radio spectrum dedicated to ITS is in the 5.8 GHz frequency band. ETSI and CEN have formed working groups and technical committees dedicated to the ITS domain.

Although the benefits from the use of co-operative systems in transport are numerous there are also some difficulties. Some of the concerns are the following: wide uptake of such systems, market penetration, standards finalization and consensus among different standardization organizations, all the inherent problems of wireless technologies (multi-path propagation, security issues etc.).

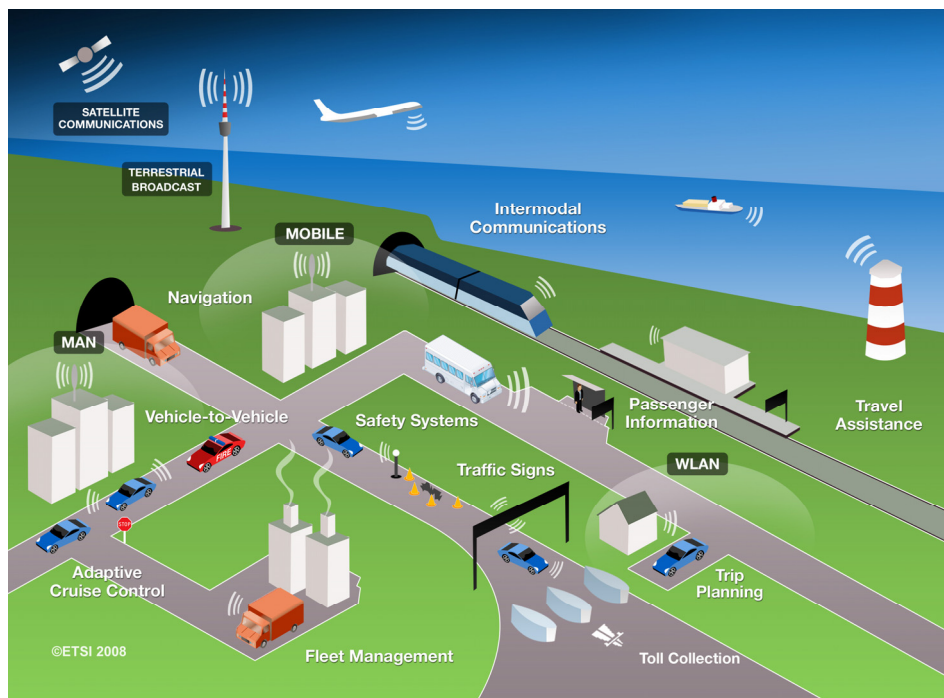


Fig. 1. Indicative ITS applications (ETSI, 2011).

The remainder of this chapter is organized as follows. In the next section, the architecture of co-operative systems is described. In the following wireless technologies used within the co-operative systems framework are outlined. The applications of vehicular networks and their corresponding categories are highlighted. Emphasis is given on hot research topics concerning co-operative systems such as data fusion, routing, security and privacy. Eventually, conclusions are drawn.

2. Architecture

In co-operative systems the specification of a unified communication architecture plays a central role for further deployment. As a result of the deployment of co-operative systems the road users will benefit from improved safety, reduced traffic congestion, environmental friendly driving and much more. The key to achieving these benefits lies in the specification of a common and standardized communication architecture among the various components of such systems. This architecture comprises four main components which can be composed arbitrarily to form a co-operative intelligent transport system. To form such a system there is no need to have all these four components available but a subset of them is sufficient. The components can communicate with each other either directly within the same communication network or indirectly across several communication networks. These four components are depicted in Figure 2 and are briefly described in the following. For a more detailed description one can refer to (Bechler et al., 2010; ETSI, 2010).

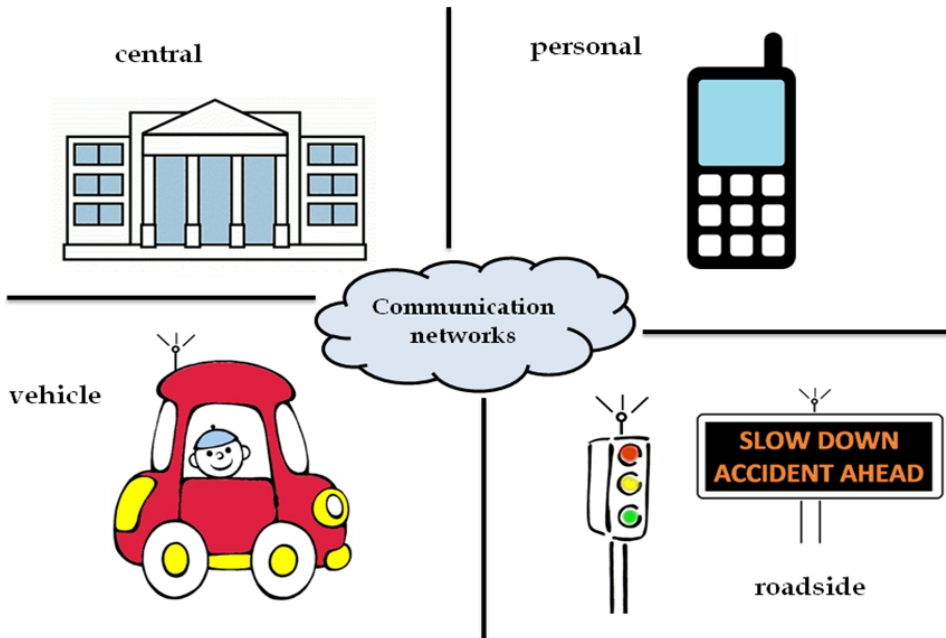


Fig. 2. Communication Architecture Components.

2.1 Vehicle component

The vehicle is equipped with communication capabilities (i.e. a router, embedded PCs) to establish communication with other vehicles and the roadside infrastructure. These modules have access to the CAN network of the vehicle as well as to other vehicle data which they collect, process and communicate to other vehicles, the roadside units or the central system. The exact HW solution is not strictly defined and it can be a unique HW unit or several units which form a LAN inside the vehicle.

2.2 Roadside component

The roadside component includes variable message signs, traffic lights and other units which are equipped with communication capabilities. This way, the roadside component can communicate with other vehicles by sending them information or acting as a relay station supporting multi-hop communication. Moreover, this component can communicate with other roadside units and the central system and therefore forward information received from vehicles. The roadside component can be also connected to the Internet.

2.3 Personal component

The personal component is actually a nomadic device, that is a personal navigator or a smartphone, which can host a variety of ITS applications. These devices can also support co-operative ITS applications based on communication with other road users or the road infrastructure.

2.4 Central component

The central component is a public authority or a road operator who manages the co-operative applications or services. An example of such component is a traffic management center which uses roadside units to inform the drivers about traffic status or accidents in a specific road network and suggests alternative routes. The central component can receive information from vehicles or roadside units and in turn send information to them.

2.5 Reference protocol stack

Each one of the above components contains an ITS station which in turn comprises a number of ITS specific functions and a set of devices implementing these functions. From a communication's point of view an ITS station is based on the reference protocol stack depicted in Figure 3. This protocol stack follows the ISO/OSI reference model and consists of four horizontal layers and two vertical ones that flank the horizontal stack. Access, networking and transport, facilities and applications layers are the horizontal ones, whereas management and security are the vertical ones.

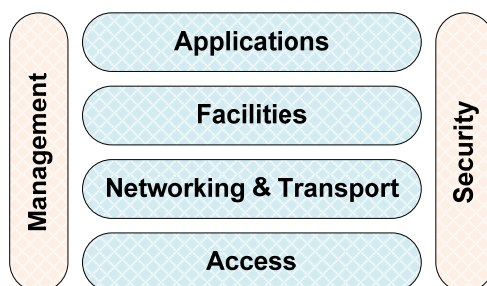


Fig. 3. Reference protocol stack of an ITS station.

3. Wireless technologies

The wireless technologies used for the continuous communication among different vehicles and between vehicles and the road infrastructure are the cornerstone of co-operative systems. These technologies concern the *networking & transport* and *access* layers of the reference protocol stack of an ITS station (Figure 3) and can be divided into two main categories: general and vehicular specific communication technologies.

Therefore, for the connection among vehicles and the road infrastructure a mixture of general and vehicular specific technologies is needed. Some of these technologies are already in use, while some others are still under development.

3.1 General communication technologies

This category comprises well known wireless communication technologies such as cellular networks, WiMAX, WiFi, infrared, bluetooth, DVB/DAB etc. which are not specifically developed for vehicular networks but play a significant role for future deployment of co-operative systems. Currently, in ITS the main focus is on cellular networks and WiMAX and for this reason only those two will be analyzed below.

3.1.1 Cellular networks

Cellular networks are evolving rapidly to support the increasing demands of mobile networking. Although these networks are designed for voice data exchange they can be applied also to vehicular networks especially for information and entertainment applications. Nowadays, cellular networks migrate from GPRS, to UMTS, to LTE standards, increasing bandwidth and reducing delay times making these networks appropriate also for other kind of applications such as efficiency and trip planning.

Cellular networks have several characteristics suitable for co-operative systems like large scale usage and long range communication. However, some drawbacks of cellular networks which are relevant to vehicular connectivity are summarized below:

- Increased latency (i.e. voice data higher priority than text data, data sent via base stations)
- No broadcasting capabilities, only support of point-to-point communication
- Operation fees (e.g. internet access, roaming)

However, despite all the above disadvantages, cellular networks can be used for ITS applications which require moderate delay, long range communication, and low data rate. With the migration from 3G towards 4G (such as LTE) the focus remains on technologies that can serve a circular area with Internet connectivity, with no special provision for following the road infrastructure and optimising for connected car services.

3.1.2 WiMAX

WiMAX (Worldwide Interoperability for Microwave Access) is based on IEEE 802.16 standard and aims at providing wireless data over long distances in a variety of ways, from point-to-point links to full mobile cellular type access. WiMAX provides support for mobility and it will fill the gap between 3G and WLAN standards. It offers high data rates (<40Mb/sec), portable connectivity at low speeds (<60km/h) and wide area coverage (<10km) required to deliver high speed internet access to mobile clients. WiMAX provides a wireless alternative to cable and xDSL for last mile broadband access and can be used for V2I or I2I long range communication. At this point it should be mentioned that WiMAX supports several service levels including guaranteed Quality of Service (QoS) for delay sensitive applications and an intermediate QoS level for delay tolerant applications that require a minimum guaranteed data rate.

3.2 Vehicular specific communication technologies

This category includes communication technologies which are dedicated to vehicular applications and actually were the result of additional communication requirements posed by ITS applications. Dedicated communication standards are in development for co-operative systems. At the access layer, a convergence towards the IEEE 802.11p standard can be observed, while standardisation on the network and transport layer is still in progress. Several prototype implementations exist and are used in demonstrations and pilots. IP communication (focus is on IPv6) can be used on top of 802.11p, but due to the highly dynamic character of the network (i.e. movement of vehicles, relatively short communication distances) dedicated standards have been developed and are being

standardised by ISO and ETSI. The most important of them, namely DSRC, WAVE and CALM, are illustrated below.

3.2.1 Dedicated Short Range Communications

Dedicated Short Range Communications (DSRC) is a short to medium range communications service that supports both public safety and private operations in V2V and V2I communication environments (DSRC, 2003). DSRC is meant to be a complement to cellular communications by providing very high data transfer rates in circumstances where minimizing latency in the communication link and isolating relatively small communication zones are important.

DSRC is designed for vehicular wireless communications and operates on radio frequencies in the 5.725 to 5.875 GHz (Industrial, Scientific and Medical - ISM) band in Europe and in the 5.850 to 5.925 GHz band in the United States. DSRC systems consist of Road Side Units (RSUs) and On Board Units (OBUs) with transceivers and transponders. The DSRC standards specify the operational frequencies and system bandwidths, but also allow for optional frequencies which are covered (within Europe) by national regulations.

The range of communication using DSRC is up to 1000m with data rates of 6–27 Mb/s, where vehicles may be moving at speeds up to 140 km/h. As mentioned previously, DSRC is divided into two types of communication, namely V2V and V2I. V2V communication is used when vehicles need to exchange data among themselves in order for co-operative applications to work properly, whereas V2I communication is used when roadside units are part of the co-operative application. In co-operative systems, some applications are required to send messages periodically (e.g. every 100ms), whereas other applications send messages when an event occurs.

At this point it should be highlighted that DSRC systems are used in the majority of European Union countries, but these systems are currently not totally compatible. Therefore, standardization is essential in order to ensure pan-European interoperability, particularly for applications such as electronic fee collection, for which the European Union imposes a need for interoperability of systems.

Standardization will also assist with the provision and promotion of additional services using DSRC, and help ensure compatibility and interoperability within a multi-vendor environment.

3.2.2 Wireless Access in Vehicular Environments

The design of an efficient communication protocol in the automotive sector that deals with privacy, security, multi-channel operation and management of resources is a difficult task, which is under intensive scientific investigation. This task is assigned to a special IEEE working group and the ongoing suite of protocols is the IEEE 1609, mostly known as Wireless Access in Vehicular Environments or simply WAVE (WAVE, 2007).

The WAVE standards define an architecture and a complementary, standardized set of services and interfaces that collectively enable secure V2V and V2I wireless communications. Together these standards provide the foundation for a broad range of

applications in the transportation environment, including vehicle safety, automated tolling, enhanced navigation, traffic management and many others.

The architecture, interfaces and messages defined in WAVE support the operation of secure wireless communications among vehicles and between vehicles and the road infrastructure. Applications can use these standards in conjunction with equipment operating at 5.9 GHz to provide, for example, services for drivers, road operators, facilities operators and maintenance staff.

The IEEE 1609 Family of Standards for WAVE consists of four trial use standards which have full use drafts under development and two unpublished standards under development:

- *IEEE P1609.0 - Draft Standard for Wireless Access in Vehicular Environments (WAVE) - Architecture*
This standard describes the WAVE architecture and services necessary for multi-channel DSRC/WAVE devices to communicate in a mobile vehicular environment.
- *IEEE 1609.1-2006 - Trial Use Standard for Wireless Access in Vehicular Environments (WAVE) - Resource Manager*
This standard specifies the services and interfaces of the WAVE Resource Manager application. It describes the data and management services offered within the WAVE architecture. It defines command message formats and the appropriate responses to those messages, data storage formats that must be used by applications to communicate between architecture components, and status and request message formats.
- *IEEE 1609.2-2006 - Trial Use Standard for Wireless Access in Vehicular Environments (WAVE) - Security Services for Applications and Management Messages*
This standard defines secure message formats and processing. This standard also defines the circumstances for using secure message exchanges and how those messages should be processed based upon the purpose of the exchange.
- *IEEE 1609.3-2007 - Trial Use Standard for Wireless Access in Vehicular Environments (WAVE) - Networking Services*
This standard defines network and transport layer services, including addressing and routing, in support of secure WAVE data exchange. It also defines WAVE Short Messages, providing an efficient WAVE-specific alternative to IPv6 (Internet Protocol version 6) that can be directly supported by applications. Further, this standard defines the Management Information Base (MIB) for the WAVE protocol stack.
- *IEEE 1609.4-2006 - Trial Use Standard for Wireless Access in Vehicular Environments (WAVE) - Multi-Channel Operations*
This standard provides enhancements to the IEEE 802.11 Media Access Control (MAC) to support WAVE operations.
- *IEEE P1609.11 Over-the-Air Data Exchange Protocol for Intelligent Transportation Systems (ITS)*
This standard will define the services and secure message formats necessary to support secure electronic payments.

Additionally, the IEEE 1609 standards rely on IEEE P802.11p. This proposed standard specifies the extensions to IEEE 802.11 that are necessary to provide wireless communications in a vehicular environment.

3.2.3 Communications Access for Land Mobiles

The Communications Access for Land Mobiles (CALM) framework is an ISO TC204 initiative that specifies a common architecture, network protocols and communication interface definitions for wired and wireless communications using various access technologies including cellular 2nd generation, cellular 3rd generation, satellite, infra-red, 5 GHz micro-wave, 60 GHz millimetre-wave, and mobile wireless broadband (CALM, 2007). These and other access technologies that can be incorporated are designed to provide broadcast, unicast and multicast communications between mobile stations, between mobile and fixed stations and between fixed stations in the ITS sector.

The CALM concept is therefore developed to provide a layered solution that enables continuous or quasi continuous communications between vehicles and the infrastructure, or between vehicles, using such (multiple) wireless telecommunications media that are available in any particular location, and have the ability to migrate to a different available media where required. Media selection is at the discretion of user determined parameters.

The motivations behind this standardization effort are the following:

- different countries use different ITS media,
- different ITS applications have different requirements, therefore it is impossible to use a single carrier to support all types of applications.

The following communication types are supported by CALM:

- *Vehicle-to-Infrastructure*: Multipoint communication parameters are automatically negotiated and subsequent communication may be initiated by either roadside or vehicle.
- *Infrastructure-to-Infrastructure*: The communication system may also be used to link fixed points where traditional cabling is undesirable.
- *Vehicle-to-Vehicle*: A low latency peer-to-peer network with the capability to carry safety related data such as collision avoidance and other vehicle-vehicle services such as ad-hoc networks linking multiple vehicles.

At a high level, on the one side there are multiple services possibly operating simultaneously all requesting communications services, whereas on the other side there is a possibility of multiple communications media opportunities in the vehicle to handle the transaction. In the middle CALM is located managing quasi continuous communications using the available media, to satisfy the needs of one or multiple applications. It is important to understand that the vehicle may be maintaining multiple simultaneous sessions.

Finally, it is important to highlight that the specifications and standards of CALM are not a physical piece of equipment. While CALM may indeed operate through a "box" designed to achieve its tasks, it is actually a set of protocols, procedures and management actions. The implementation is actually a commercial decision.

4. Applications

Together with the evolution of vehicular networks numerous novel ITS applications have emerged. Typical examples of co-operative applications include remote diagnostics,

collision avoidance, online navigation, map update, congestion avoidance for the driver, internet in the vehicle for passengers (e.g. gaming, downloading videos, reading the news etc.). These applications can be divided into three major categories, namely safety, efficiency and infotainment. In the following some indicative applications in each category are selected and will be described briefly.

4.1 Safety

One of the main goals of transport authorities is the minimization of traffic accidents and the increase of road safety. The exploitation of wireless technologies will be a significant asset towards this direction as it has been obvious from the results of the SAFESPOT project (SAFESPOT, 2006-2010). Examples of co-operative safety related applications will be given below.

4.1.1 Frontal collision warning

Frontal collisions represent a major proportion of accidents worldwide. Typical causes of such accidents are the distraction of the driver, sudden braking of a vehicle ahead, the presence of a stationary obstacle in front of the vehicle (e.g. right after a turn) etc. Conventional collision warning systems are based on sensors installed in the vehicle. These sensors could be long range radars for adaptive cruise control, camera sensors for objects detection, cameras covering the blind spot area, laser scanners for both detecting and classifying objects. This way a vehicle can be informed about events and targets which are within range of the detection sensors. Figure 4 shows a frontal collision warning application where the vehicle in front brakes while other vehicles are following.

The reliability and accuracy of a conventional collision warning system is based on the number and the type of sensors used, as well as the type of the environment (i.e. urban, inter-urban, highway) around the vehicle. The occlusion of sensors from obstacles, the limited range of sensors and other physical constraints, reduce system's range and degrade its performance. Apart from these factors, a collision warning system to function properly needs a multitude of sensors to cover the entire area around the vehicle, which makes such a system extremely expensive.

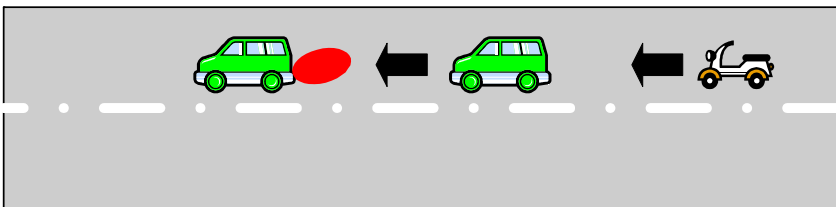


Fig. 4. Frontal collision warning application.

The collision warning system can be much more effective if other neighboring vehicles communicate with the subject vehicle, extending thus the perception of the driver in relation to the limited perception based only on sensors installed in the vehicle. Actually, this is the principle of co-operative collision warning systems. While driving, equipped vehicles

anonymously share relevant information, including their position, speed and direction. This way each vehicle monitors the intentions of other drivers and the location and behavior of all vehicles in the neighborhood. When a vehicle detects a critical situation, the system warns the driver with a visual, audible and/or haptic manner. Thus, the driver has enough time to intervene and avoid a collision.

In time critical situations, immediate intervention to avoid a collision is feasible with the use of communication. This would not be possible in case only onboard sensors were used because of the delay in detecting and classifying objects and analyzing the ongoing situation. The co-operative approach also has great influence on the classification of objects. If vehicles are equipped with wireless communication they can directly exchange information about their type (e.g. truck, car, motorbike).

4.1.2 Intersection safety

A significant number of accidents in urban areas occur at intersections. The reasons for this are the significant burden of the driver from the complex situations that can occur at intersections due to many vehicles that are flooding them from different directions, the variety of road users (cars, trucks, pedestrians, cyclists etc.) as well as buildings and walls that limit the visibility of the driver.

In order for an accident or a dangerous situation to occur at an intersection it is supposed that there should be a violation of traffic rules such as traffic light or "STOP" sign violation. But intersections are complex road environments and accidents are likely to occur even if the rules are obeyed (e.g. abrupt braking while the traffic light turns from green to red). A simplified example of an intersection safety application is shown in Figure 5.

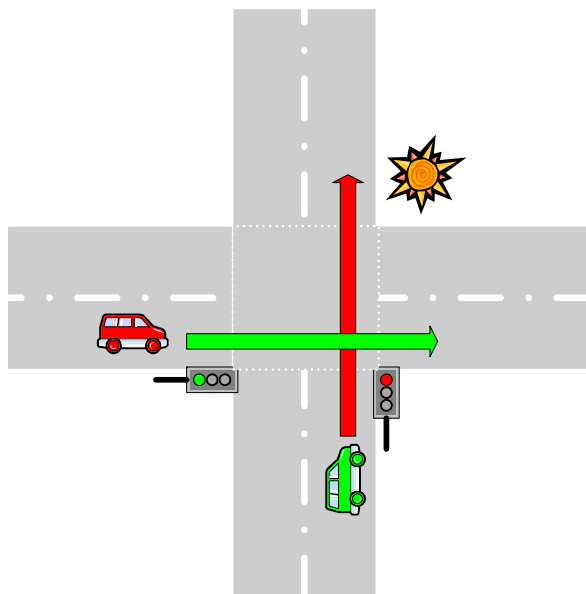


Fig. 5. Intersection safety application.

The benefits from the exploitation of wireless communication for safety reasons at intersections are significant. First of all, the traffic lights and other traffic signs can emit their status, together with information concerning the instant that this status is going to change (e.g. traffic light that turns from red to green), to the interested drivers informing them about the real ongoing situation at the intersection. In addition, vehicles can emit their position and dynamic state (speed, acceleration, steering angle) and thus to inform other drivers at the intersection about their presence. Otherwise it would be impossible for other drivers to be aware of their presence because of the occlusion of the surrounding buildings, walls and other obstacles. Finally, some sensors, such as laser scanners, could be installed at critical points at an intersection to detect pedestrians and cyclists (vulnerable road users) and inform the drivers for their presence through wireless communication.

4.1.3 Slippery road detection

This application informs the driver about the status of a road segment where there is a possible risk. The risk is primarily related to a slippery road surface that may be due to adverse weather conditions (e.g. ice, rain) or to any extraordinary event (e.g. an oil leak of the vehicle ahead). The detection of the slippery road segment can be carried out by a vehicle either directly by specific sensors or indirectly by activation of the ABS or ESP system. Even though this information, about the dangerousness of the road, is transmitted to the driver directly, it may be too late to take action because it is highly likely that the vehicle has slipped already since the ESP or ABS has been activated. A slippery road detection application is depicted in Figure 6.

As it is obvious, this application has almost no interest without the use of wireless communication. By taking advantage of wireless communication the vehicular network can share information related to hazardous road segments such as slippery roads. For example, a vehicle that its ESP system is activated associates this data with its current location and notifies other nearby vehicles and possibly a RSU if it is in communication range. The neighboring vehicles receiving such information shall promptly inform the drivers about the potential risk, while retransmitting this information to other nearby vehicles (multi-hop communication). As an alternative, a RSU can be equipped with some special sensors which can detect the hazardous road conditions and inform the drivers approaching this area.

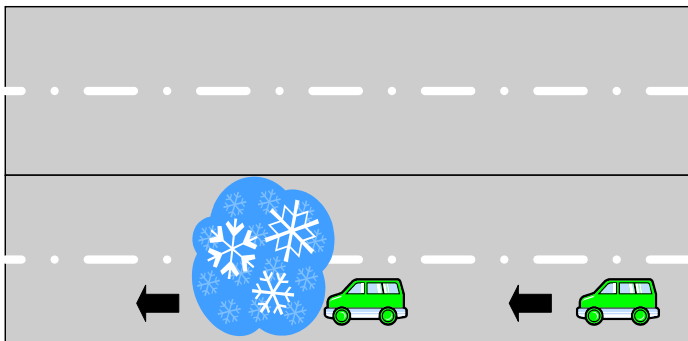


Fig. 6. Slippery road detection application.

As a conclusion, there is an apparent advantage in using wireless communication to broadcast information about the road conditions to the approaching drivers who will then have sufficient time to react. With this co-operative approach the RSU has the ability to transmit road condition information to the traffic management center which then can be analyzed and checked for accuracy and quality and transmitted to other vehicles that drive on this road segment.

4.2 Efficiency

The climate change that has been observed in recent years has also affected the priorities in the transport agenda. Nowadays minimization of CO₂ emissions related to transport and environmental friendly driving comprises a top priority. The results of the CVIS project (CVIS, 2006-2010) have shown the added benefit from the use of wireless communication to enhance efficiency in transport. Examples of co-operative efficiency related applications will be highlighted in the following.

4.2.1 Enhanced route planning

In this application the infrastructure continuously collects information related to traffic density and makes forecasts for road segments with potential traffic jams. Then when an equipped vehicle drives next to a RSU the traffic density information on the neighboring area as well as driving instructions are transmitted to it. This information is processed by the vehicle and then the driver is informed about possible delays and alternative routes to avoid traffic jams. This way a significant number of drivers could be guided around congested areas so the entire transport system to become more efficient. A significant side effect of this application will be the reduction of pollution deriving from vehicles got stack in congested highways.

Figure 7 shows an enhanced route planning scenario in which the RSU informs the driver of the red vehicle about heavy traffic ahead and suggests a faster alternative route to the driver's final destination. It is obvious from the description of this application that for its implementation a RSU with wireless communication is essential to provide the necessary information. A vehicle equipped only with some perception sensors, without communication capabilities, could make a rough estimation about the traffic density in the road segment it is currently driving, but without any information on alternative routes.

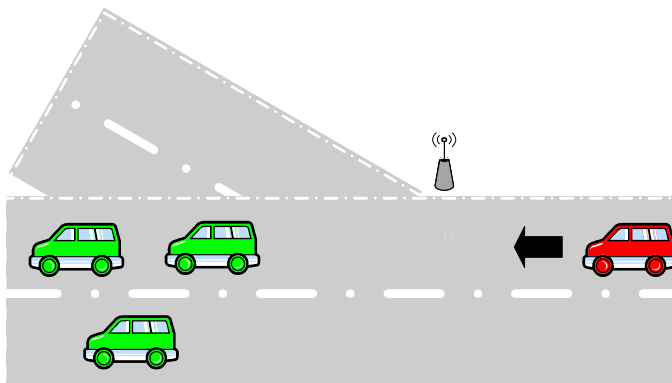


Fig. 7. Enhanced route planning application.

4.2.2 Optimal speed advice

The aim of this application is to provide the needed information to the driver in an effort to make driving smoother and reduce significantly start and stop situations. As the vehicle approaches a signalized intersection it receives information about the exact position of the traffic light and the duration of its current status (e.g. the remaining seconds for the traffic light to become red).

Based on this information, that is using the distance from the intersection and the time until the traffic light turns to green, the approaching vehicle calculates the speed that it should follow to avoid stopping at the intersection. Then this "optimal" speed is provided to the driver and if this suggestion is followed it is very likely that the traffic light will turn into green as soon as the vehicle reaches the intersection and there is no need for it to stop. Fewer stops result in increased traffic flow and reduced fuel consumption for the equipped vehicles.

Figure 8 shows an example of this application in which the driver follows the optimal speed suggestion and thus will not have to stop at the intersection. For such kind of application to become a reality it is essential that the traffic light should be equipped with wireless communication capabilities.

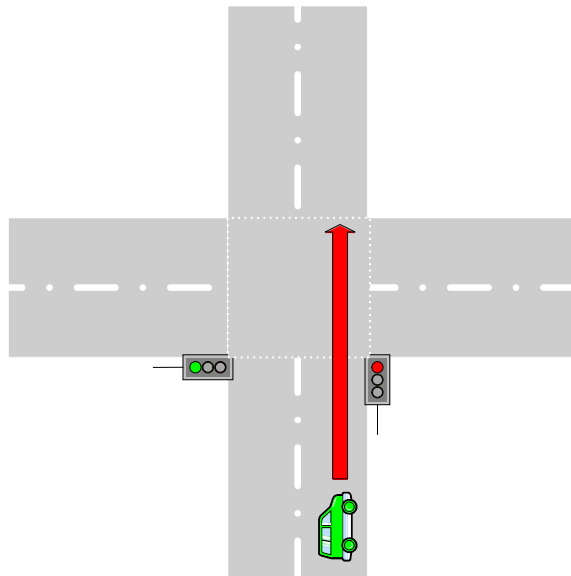


Fig. 8. Optimal speed advice application.

4.2.3 Traffic merge assistant

The traffic merge assistant application allows the merging of cars in a joint traffic flow without the need to interrupt their smooth flow. When a vehicle enters a highway from an entrance communicates this intention to the neighboring vehicles. The vehicle entering the

highway requires specific maneuvers by surrounding vehicles to adapt in a safe and continuous way to the traffic. If there are no objections from other drivers, then either the traffic will be adjusted automatically or advice will be given to the drivers on how to act. In this way the vehicle entering the highway can adjust smoothly into the flow of traffic without causing major disruptions to it. This application can be enhanced by using a RSU which can determine the movements of each participant.

This application, as shown in Figure 9, requires a significant number of vehicles to be equipped with wireless capabilities. Moreover, it is one of the most difficult and complex applications which is based on the harmonious and trustworthy co-operation among highly dynamic vehicles.

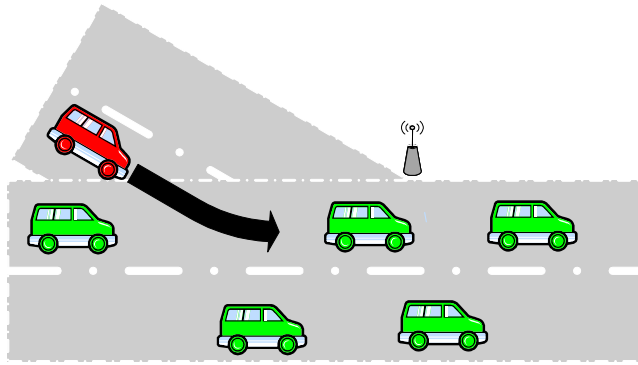


Fig. 9. Traffic merge assistant application.

4.3 Infotainment

The entertainment and information provision to the driver and the passengers might not be life critical but is still very important in today's society, where internet and exchange of information dominate. Examples of co-operative applications related to information and entertainment (infotainment) will be outlined in this section.

4.3.1 Points of interest

The Points of Interest (PoI) application allows local businesses, touristic attractions, restaurants, gas stations and so on to advertise their availability to nearby vehicles. In this case, a RSU transmits information about a PoI, such as its location, hours of operation and pricing. Then this information is filtered by the vehicles dynamically, depending on the case, and the relevant information is presented to the driver. For example, if the fuel level is low, the vehicle could show to the driver the locations and prices of gas stations in the surrounding area. The benefit of this application is that advertising gets more effective as the driver moves within the geographical area where the PoI is located and it is more likely to visit it rather than if he listened about it to a radio station or found it on the web. Moreover, another benefit is that consumers receive up to date information directly from a business in the neighborhood.

4.3.2 Internet access

This application allows drivers and passengers to access the Internet. This in turn means the use of all types of services based on the IP protocol inside the vehicle. Therefore, a multi-hop route from a RSU to the relevant vehicle is installed and maintained to act as a gateway to the Internet. This path of multiple hops takes place transparently to the upper layers of the protocol stack and enables almost any service based on IP protocol to be used inside vehicles. Finally, this application allows access to the driver and passengers in any type of information available on the Internet (e.g. downloading of updated digital maps).

4.3.3 Remote diagnostics

This application allows an authorized service station to assess the condition of a vehicle without needing a physical connection with it. When a vehicle enters the parking building of the station, remote diagnostics system may ask the vehicle about relevant information to support the diagnosis of the problem reported by the client. Moreover, as the vehicle is approaching its service history and the necessary customer information can be retrieved from a database and be ready for use by the technician. If software updates are necessary, the system can install the updates without a physical connection. This application can reduce the amount of time required for a customer during a visit to an authorized service station. This fact will also reduce both the repair cost and the waiting time for the customers.

5. Research topics

Co-operative systems have received particular attention, with respect to the research activities in the field of ITS, the last decade. The recent advances in information and communication technologies have enabled the deployment of co-operative systems as an exciting platform for developing new and useful vehicular applications. The research activities focus mainly on data fusion, routing as well as on privacy and security issues which will be analyzed in the following.

5.1 Data fusion

Data fusion plays an important role in co-operative systems. A stand alone sensor or several sensors installed in a vehicle cannot overcome certain physical limitations as, for example, the limited range and field of view. Therefore combining information coming from both onboard sensors and wireless messages, encompassing information from other vehicles, broadens the awareness of the driver and increases the reliability of the whole system in case of sensor failure. However, fusing information from highly mobile vehicles, forming a wireless network, is a challenging task (Ahlers & Stimming, 2008; Lytrivis et al., 2008).

The Joint Directors of Laboratories (JDL) functional model, which is the most prevalent in data fusion community, is depicted in Figure 10. According to this model the data processing is divided to the following levels: signal, object, situation and application. All these levels communicate and exchange data through a storage and system manager (Liggins et al., 2008).

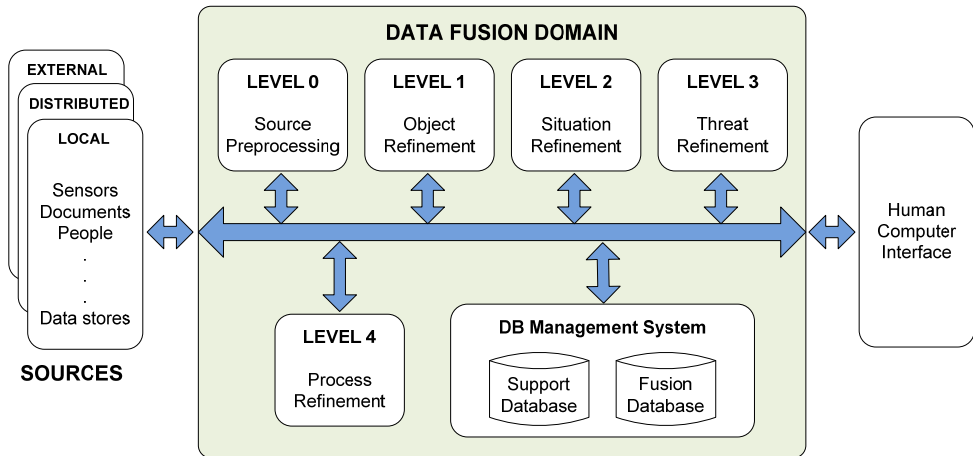


Fig. 10. Joint Directors of Laboratories (JDL) model.

In the data fusion process the main focus is on object and situation refinement levels, which refer to the state estimation of objects and the relations among them, correspondingly. The discrimination between these levels is also made by using the terms low and high level fusion instead of object and situation refinement. The different levels of the JDL model are summarized below:

- *Level 0:* Preprocessing of sensor measurements (pixel/signal-level processing).
- *Level 1:* Estimation and prediction of entity states on the basis of inferences from observations.
- *Level 2:* Estimation and prediction of entity states on the basis of inferred relations among entities.
- *Level 3:* Estimation and prediction of effects on situations of planned or estimated/predicted actions by the participants.
- *Level 4:* Adaptive data acquisition and processing related to resource management and process refinement.

In the past decade the advances in autonomous sensor technologies and the major objective of the European Union to reduce to a half road accidents and fatalities by 2010, led to the development of advanced driver assistance systems (ADAS). The fusion of data coming from different advanced in-vehicle sensors was initially in the centre of this attempt. However, this approach suffers from serious limitations. Specifically:

- the perception environment of the vehicle cannot go beyond the sensing range,
- the sensor systems cannot perform well in all environments (the urban roads comprise a major challenge),
- in several cases the system is not able to perceive the situation in time in order to warn the driver and suggest a corrective action,
- the cost of the sensor systems is too high and so their installation is feasible only at luxurious vehicles.

Recently the focus of research activities on co-operative systems is driven by the attempts to overcome all the above limitations. The limited bandwidth, security issues, privacy, reliability and propagation are some of the emerging disadvantages of the wireless connectivity in vehicles. All these issues poses additional challenges to the data fusion process. The association and synchronization of data from on-board sensors together with the wireless network data is the main challenge. Moreover, the manipulation of delayed information and the reliability of the information transferred via the network are other important issues.

5.2 Routing

Routing is the process of finding a path from a source node to a destination node. In this section the word "node" will be used interchangeably with the word "vehicle" because a vehicle is actually a node of a vehicular network. Since each node has limited transmission range, messages often need to be forwarded by other nodes in the network to reach their final destination (i.e. multi-hop communication).

Despite the fact that there are already some routing protocols available, which are mainly derived from the Mobile Ad-hoc Network (MANET) domain, it is an intensive scientific research area due to the highly dynamic nature of vehicular networks.

The routing protocols designed specifically for co-operative systems (Lee et al., 2010; Li & Wang, 2007) can be divided into two broad categories: *topology-based* routing and *location-based* routing. The former use information about the existing links of the network to forward the relevant messages. In the latter forwarding decisions are based on the location of the nodes. Moreover, position based routing protocols can be further divided into *proactive* and *reactive*.

Proactive algorithms are using classical routing strategies such as distance-vector routing or link-state routing. Proactive algorithms maintain routing information about the available paths in the network even if these paths are not currently used. The main disadvantage of this approach is that the maintenance of unused paths occupies a significant part of the available bandwidth if the network topology changes frequently.

In response to the problem of maintaining the paths of proactive protocols, reactive routing protocols were created. Reactive protocols maintain only routes that are in use, thereby reducing the load on the network when only a small subset of available paths are used.

In location-based routing, forwarding decisions are based on the location of the node that forwards the message according to the location of the source and destination nodes. In contrast to pure ad hoc approaches which are based on topology-based routing, here it is not necessary to setup or maintain a path since packets are forwarded directly. Location-based routing protocols consist of location services and geographical forwarding.

Geographical forwarding takes advantage of a topological assumption which works well for wireless ad hoc networks: nodes that are physically close are likely to be close in the network topology too. Each node is aware of its location using technologies such as GPS and periodically broadcasts its presence, location and speed to its neighbors. Thus, each node maintains a table with the identities and locations of its current neighbors. When one node

needs to forward a packet it includes the identifier of the destination-node and its geographical location into the header of the packet. Each node along the forwarding path consults its list of neighbors and forwards the packet to the neighbor closest to the destination in terms of physical location, until it reaches its final destination.

Although the geographical forwarding works well for networks where nodes are uniformly distributed, perhaps cannot find a route to a packet's destination when the packet has to travel around a topology "hole" - that is, when an intermediate forwarding node has no neighbors who are closer than itself to the destination of the packet.

An overview of some *topology-based* routing algorithms is given below:

- **Ad Hoc On Demand Distance Vector (AODV)** is a routing algorithm where the nodes of the network upon receiving a broadcast query they record the address of the querying node to their routing table. The process of recording the previous hop is called backward learning. When a packet reaches its destination a reply packet is sent back to the source through the full path retrieved from the process of backward learning. At every node of the path, the previous hop should be recorded, creating this way the forward path from the source. The query together with the response create a complete bidirectional path. After setting the path, it is maintained as long as the source uses it. A failure on a link will be reported recursively to the source and in turn this will trigger another query-response process for finding the new route. More details about AODV one can find in (Perkins & Royer, 1999).
- **Dynamic Source Routing (DSR)** is an algorithm that uses source routing, that is the source indicates to a data packet the sequence of intermediate nodes on the routing path. In DSR, the query packet copies in its header the identities of the intermediate nodes it has already visited. Afterwards, the destination uses the query packet to retrieve the entire path to respond to the source. As a result, the source can establish a path to the destination. If the destination node is allowed to send multiple routes responses, the source node may receive and store these multiple routes. An alternative route can be used in case a link of the current path is broken. In a low mobility network DSR has the advantage over AODV in case the alternative route can be tested before the DSR initiates another query to discover the route. There are two major differences between AODV and DSR. The first is that in AODV data packets carry the destination address, while in DSR data packets carry all the routing information. This means that DSR has probably more routing burden than AODV. Moreover, as the diameter of the network increases, the burden on the data packet will continue growing. The second difference is that in AODV route response packets carry the destination address and the sequence number, while in DSR they carry the address of each node along the route. The interested reader in DSR can refer to (Johnson & Maltz, 1996).

A brief description of some *location-based* routing algorithms is given below:

- **Connectivity-Aware Routing (CAR)** is a routing algorithm which derives from the work performed by the Preferred Group Broadcast (PGB) to reduce the broadcasted packets during the discovery of the AODV route taking also into account the mobility of the nodes. CAR uses the route discovery of AODV to find routes with reduced broadcasting from PGB. However, the nodes forming the route record neither their previous node from the backward learning nor their previous node which forwards

the response route packet from the destination. Only anchor points, which are nodes near an intersection or a curve of the road, are recorded in the route discovery packet. A node defines itself as an anchor point if its velocity vector is not parallel to the velocity vector of the previous node in the packet. The destination may receive multiple route discovery packets. If this happens it chooses the path that provides the best connectivity and the shortest delays. More details about CAR can be found in (Naumov & Gross, 2007).

- **Geographic Source Routing (GSR)** is based on the availability of a map. It calculates the shortest Dijkstra path of the cascading graph where vertices are intersection nodes and edges are the roads connecting these vertices. The sequence of intersections is setting up the route to the destination. Then the packets are greedily forwarded between intersections. GSR does not take into account the connectivity between two intersections, so the route might not be fully connected. In case such a situation occurs a recovery with greedy forwarding takes place. The most significant difference between the GSR and CAR is that CAR does not use a map and uses proactive discovery of anchor points that indicate a turn at an intersection. More details about GSR can be found in (Lochert et al., 2003).

5.3 Security and privacy

Security in V2V and V2I communications is a prerequisite for future development of co-operative systems and actual deployment in the real world. Co-operative systems have to ensure that data transmission derives from a trusted source and has not been counterfeited. For example, in a red light violation warning application, the in-vehicle system receives data from the equipment which is installed in the traffic light and then decides to issue or not a warning to the driver. An incorrect transmission from a malfunctioning or compromised unit might jeopardize vehicle's safety as well as others' safety in the vicinity. Similarly, the future development of safety applications is jeopardized without securing that transmissions are coming from a trusted source.

Privacy and anonymity are primary issues that also have to be addressed. In co-operative applications vehicles are broadcasting messages about their current location, speed and heading. It is desirable for the users to maintain their privacy since they fear that such a system could be used to build tracking mechanisms which would allow harassment, automatic issue of tickets for speeding or otherwise act in an undesirable way for them.

Unfortunately, on the other hand anonymity may be abused. Some examples are sending fake information or spamming. If the system ensures accountability¹ then the users know that there will be consequences for others if their data is abused. The challenge here is ensuring anonymity and at the same time accountability, as they seem to be conflicting.

There are many ongoing research activities on security and privacy in co-operative systems. Some ideas that have been proposed for solving such issues include public key certificates or digital signatures. For more information the interested reader can refer to (Fischer et al., 2007; Raya & Hubaux, 2007).

¹Accountability is the ability to attribute actions to the entity that caused those actions.

6. Conclusion

An overview of co-operative systems and their importance in future transportation systems has been presented in this chapter. The chapter started with a short introduction about the historical background and the purpose of co-operative systems. Emphasis was given to the communication architecture, including its components, which is mainly the outcome of the efforts carried out so far in Europe. Also the importance of the definition of a common architecture for further deployment of co-operative systems was stressed.

In the following, the focus was on the wireless technologies used within the co-operative systems framework which are divided into two categories: general and vehicular specific communication technologies. These technologies are the cornerstone of co-operative systems and their objective is the continuous communication among different road users (vehicles, motorbikes, trucks, roadside units, infrastructure etc.). To achieve this continuous and seamless communication a mixture of general and vehicular specific technologies is needed. Some of these technologies are already in use, while some others are still under development.

Additionally, some co-operative applications were described which are categorized into three main groups: safety, efficiency and infotainment. The applications addressing safety and efficiency are of great importance today because the minimization of accidents and their consequences as well as the reduction of CO₂ emissions are the primary targets worldwide. Finally, emphasis was given on hot research topics concerning co-operative systems such as data fusion, routing, security and privacy. A general description highlighting each of these topics, the research challenges as well as some solutions were indicated.

Although many problems are not yet solved, the general feeling is that vehicles could benefit from evolving wireless communications in the near future, making "talking vehicles" a reality. Co-operative systems will not only provide lifesaving and environmental friendly applications, but they will become a powerful communication tool for their users.

7. References

- Ahlers, F. & Stimming, C. (2008). "Cooperative Laserscanner Pre-Data-Fusion", in Proc. *IEEE Intelligent Vehicles Symposium (IV 2008)*, Eindhoven, 2008, pp. 1187-1190
- Bechler, M. et al. (2010). *European ITS Communication Architecture, Overall Framework, Proof of Concept Implementation*, version 3.0, COMeSafety, February 2010
- CALM (2007). International Organization for Standardization, Intelligent Transport System-Continuous Air Interface Long and Medium - Medium Service Access Point, Draft International Standard ISO/DIS 21218, 2007
- CVIS (2006-2010). "Cooperative Vehicle-Infrastructure Systems", Integrated Project co-funded by the European Commission, Available from <http://www.cvisproject.org>
- DSRC (2003). Standard Specification for Telecommunications and Information Exchange Between Roadside and Vehicle Systems-5GHz Band Dedicated Short Range Communications (DSRC) Medium Access Control (MAC) and Physical Layer (PHY) Specifications, September 2003

- ETSI (2010). *ETSI EN 302 665 - Intelligent Transport Systems (ITS): Communications Architecture*, v1.1.1, European Standard (Telecommunications series), September 2010
- ETSI (2011). European Telecommunications Standards Institute (ETSI), Intelligent Transport Systems - ITS, Available from <http://www.etsi.org/WebSite/Technologies/IntelligentTransportSystems.aspx>
- Fischer, L.; Stumpf, F. & Eckert, C. (2007). "Trust, Security and Privacy in VANETs - A Multilayered Security Architecture for C2C-Communication", In *VDI/VW-Gemeinschaftstagung: Automotive Security*, Wolfsburg, Germany, November 2007
- Hartenstein, H. & Laberteaux, K. (2010). *VANET Vehicular Applications and Inter-Networking Technologies* (Intelligent Transport Systems), Wiley, 1st edition, March 2010
- Johnson, D. & Maltz, D. (1996). "Dynamic Source Routing in Ad Hoc Wireless Networks," *Mobile Computing*, T. Imielinski and H. Korth, Eds., Ch. 5, Kluwer, 1996, pp. 153- 81
- Lee, K.; Lee, U. & Gerla, M. (2010). *Survey of Routing Protocols in Vehicular Ad Hoc Networks*, Advances in Vehicular Ad-Hoc Networks, M.Watfa Book Editor, IGI Global
- Li, F. & Wang, Y. (2007). "Routing in vehicular ad hoc networks: A survey", *IEEE Vehicular Technology Magazine*, June 2007, Vol. 2, Issue 2, pp. 12-22
- Liggins, M.; Hall, D. & Llinas J. (2008). *Handbook of Multisensor Data Fusion: Theory and Practice*, Second Edition, CRC Press
- Lochert, C.; Hartenstein, H.; Tian, J.; Fussler, H.; Hermann, D. & Mauve, M. (2003). "A routing strategy for vehicular ad hoc networks in city environments," *IEEE Intelligent Vehicles Symposium 2003*, 9-11 June 2003, pp. 156-161
- Lytrivis, P.; Thomaidis, G. & Amditis, A. (2008). "Cooperative Path Prediction in Vehicular Environments," in *Proc. 11th Int. IEEE Conf. on Intelligent Transportation Systems (ITSC 2008)*, Beijing, 2008, pp. 803-808
- Naumov, V. & Gross, T. (2007). "Connectivity-Aware Routing (CAR) in Vehicular Ad-hoc Networks," *26th IEEE International Conference on Computer Communications (INFOCOM 2007)*, May 2007, pp.1919-1927
- Perkins, C. & Royer, E. (1999). "Ad-Hoc On-Demand Distance Vector Routing," *Proc. IEEE WMCSA '99*, New Orleans, LA, Feb. 1999, pp. 90-100
- Popescu-Zeletin, R.; Radusch, I. & Rigani, M. A. (2010). *Vehicular-2-X Communication: State-of-the-Art and Research in Mobile Vehicular Ad hoc Networks*. Springer, 1st Edition, May 2010
- Raya, M. & Hubaux, J.P. (2007). "Securing vehicular ad hoc networks", *Journal of Computer Security* 15 (2007), IOS Press, pp. 39-68
- SAFESPOT (2006-2010). "Cooperative vehicles and road infrastructure for road safety", Integrated Project co-funded by the European Commission, Available from <http://www.safespot-eu.org>
- WAVE (2007). IEEE Standards Association, IEEE P1609.1 - Standard for Wireless Access in Vehicular Environments (WAVE) - Resource Manager, IEEE P1609.2 - Standard for Wireless Access in Vehicular Environments (WAVE) - Security Services for Applications and Management Messages, IEEE P1609.3 - Standard for Wireless Access in Vehicular Environments (WAVE) - Networking Services, IEEE P1609.4 - Standard for Wireless Access in Vehicular Environments (WAVE) - Multi-Channel

Operations, adopted for trial-use in 2007, IEEE Operations Center, 445 Hoes Lane, Piscataway, NJ, 2007

Zhou, M.-T.; Zhang, Y. & Yang, L. (2010). *Wireless Technologies in Intelligent Transportation Systems* (Transportation Issues, Policies and R & D), Nova Science Pub Inc, June 2010

Wireless Technologies in the Railway: Train-to-Earth Wireless Communications

Itziar Salaberria, Roberto Carballedo and Asier Perallos
*Deusto Institute of Technology (DeustoTech), University of Deusto
Spain*

1. Introduction

Since the origins of the railway in the XIX century most of the innovation and deployment efforts have been focused on aspects related to traffic management, driving support and monitoring of the train state (Shafiullah et al., 2007). The aim has been to ensure the safety of people and trains and to meet schedules, in other words, to ensure the railway service under secure conditions. To achieve this it has been necessary to establish a communication channel between the mobile elements (trains, infrastructure repair machinery, towing or emergency vehicle, and so on) and the earth fixed elements (command posts and stations, signals, tracks, etc.) (Berrios, 2007).

Nowadays, safety is a priority too, but new requirements have arisen, mainly concerning the quality improvement of the transport service provided to the passengers (Aguado et al., 2005). Moreover, the current European railway regulation by establishing that railway services be managed by railway operators independent of railway infrastructure managers makes it necessary for infrastructure fixed elements to share information with mobile elements or trains (handled by railway operators). This new policy results in additional requirements on the exchange of information between different companies. How to fulfil these requirements is a new technological challenge in terms of railway communications (Shafiullah et al., 2007) that is explored in this chapter.

The use of wireless technologies and Internet is growing in the railway industry which allows the deployment of new services that need to exchange information between the trains and terrestrial control centres (Shafiullah et al., 2007). In this sense, there are suitable solutions for other environments which allow to manage the bandwidth in terms of data rates. Therefore, these solutions are not designed for railway needs, and do not cover all the requirements that the railway industry has (California Software Labs, 2008; Marrero et al., 2008).

This chapter describes a specific wireless communications architecture developed taking into account railway communications needs and the restrictions that have to be considered in terms of broadband network features. It is based on standard communication technologies and protocols to establish a bidirectional communication channel between trains and railway control centres.

The second section of this chapter includes a brief description of the state of art in railway communications. The third one describes a specific train-to-earth wireless communication architecture. The fourth section describes the main challenges concerning with the management of the quality of service in train-to-earth communications. The fifth identifies some services that are arising as result of using this connectivity architecture and the way in which they interoperate. The sixth section shows the future lines of work oriented to improve the proposed communication channel. Finally, the seventh section of the chapter establishes the main conclusions of this work.

2. State of the art in railway communications

Railway communications emerged almost exclusively from the communication between fixed elements to carry out traffic management and circulation regulation. The technologies that communicate fixed elements with mobile elements (trains) are relatively recent, and they have contributed to improve and simplify the work required for rail service exploitation. Therefore, focusing on the network topology, two categories can be identified within the field of railway communications: a first one involving only fixed elements, and a second one involving both, fixed and mobile elements (called "train-to-earth" communications) (Salaberria et al., 2009). For the former, the most efficient solutions are based on wired systems. The latter has undergone great change in recent years, requiring wireless and mobile communications (Laplante & Woolsey, 2003).

Traditionally, the communication between fixed elements and trains has been established using analogical communication systems, such as the traditional telephone or PMR (Private Mobile Radio) based on radio systems (ETSI, 2008). These analogical systems are still used for voice communications and issues related with signalling. However, their important limitations in terms of bandwidth are causing the migration to digital systems, which offer a higher bandwidth.

Among the technologies of communication "train-to-earth", one of the most important advances of the last decade has been the GSM-R (Global System for Mobile Communications - Railway) (International Union of Railways, 2011). This system is based on the GSM telephony, but has been adapted to the field of railways. GSM-R is designed to exchange information between trains and control centres, and has as key advantages its low cost, and worldwide support.

Another technology that provides a wide circulation in the rail sector is the radio system TETRA (Terrestrial Trunked Radio) (ETSI, 2011). TETRA is a standard for digital mobile voice communications and data communication for closed user groups. The system includes a series of mobile terminals, similar to walkie-talkies, which allow establishing direct communication between control centres, train drivers and maintenance personnel, in addition to being able to establish communications with earthlines and mobile phones. Being a private mobile telephone system, its deployment in the rail sector is very simple, because it is based on the placement of a series of antennas at stations or control centres along the route.

In addition, the special-purpose technologies mentioned so far include the growing use of wireless communication technologies based on conventional mobile telephony (GSM, GPRS

and UMTS) and broadband solutions such as WiFi (IEEE 802.11, 2007) or WiMax (IEEE 802.16.2, 2004). The wireless local area networks WiFi enable the exchange of information, at much higher speeds and bandwidths than with other technologies. The cost of deployment of such networks is very low, but these are limited in terms of coverage or distance they cover. To address this limitation, the WiMax technology has emerged extending the reach of WiFi, and is a very suitable technology to establish radio links, given its potential and high-capacity at a very competitive cost when compared with other alternatives (Aguado et al, 2008).

All technologies discussed so far aim to establish a wireless communication channel between fixed elements and mobile elements of the railway field, but what happens with the services offered by means of this communication channel?, how can they have access to the channel?, how can they share it?. To address these questions, a categorization of railway services is necessary. Traditional applications or services of the railway can be classified into two major groups: (1) services related with signalling and traffic control; and (2) services oriented to train state monitoring.

The first group of services is based on the exchange of information between infrastructure elements (tracks, signals, level crossings, and so on) and control centres, all of them fixed elements. Additionally, it uses voice communication between train drivers and operators in the control centres. Therefore, for this type of service, traditional communication systems based on analogical technology remain significant.

The second group of services requires the exchange of information in the form of "data" between the trains and the control centres. In this case, the new services use any of the wireless technologies mentioned so far, but on an exclusive basis, which means that each application deployed on the train must be equipped with its own wireless communications hardware. This leads to have an excessive number of communications devices, often underused. In addition, there are still many applications that require a physical connection "through a wire" between the train devices and a computer for information retrieval and updating tasks.

On the other hand, a new set of services around the end user (passenger, or companies who need to transport some goods) is emerging. These services are oriented to providing a transport service of higher quality that not only is safe, but provides additional benefits such as: detailed information about the location of trains and schedules, contextual advertising services, video on demand, and so on. All these services are characterized by their need of a wireless communication channel with high bandwidth and extensive coverage (Garstenauer & Pocuca, 2011). As a result, the following needs are identified: (1) to standardize the way to exchange train state information between the trains and the control centres; and (2) to define a wireless communications architecture suitable for the new 'end user'-oriented services (Aguado et al., 2005).

In this chapter, a specific communications architecture based on standard technologies and protocols; that is designed to manage train-to-earth connectivity at application layer, will be presented in order to fulfil such needs.

3. Managing train-to-earth wireless communications

This section describes a general purpose wireless communications architecture to address the needs for high bandwidth and wide coverage. This solution is based on the

management of a wireless communications channel at the application layer. The architecture proposed is currently being deployed in some railway companies from Spain (Euskotren and ETS from Basque Country, and Renfe from Spain), will be presented (Gutiérrez et al., 2010). It is an innovative general purpose wireless communication channel which allows the train to communicate with the railway control centres in such a way that the applications or services are unaware of communication issues such as: establishment and closure of the communication, management of the state of connectivity, prioritization of information and so on.

This new wireless communication architecture has to respond to the demand for communication and transmission of information from any application, so it will have to take into account the nature of the information to be sent. The information exchanged between two applications (one on earth and the other on a train) may have different urgency degrees depending on their purpose or treatment with respect to the exchanged information. In fact, there is information that needs to be transmitted at the time that it is generated, for example in case of positioning information or alarms in some critical train operation elements. On the other hand, there may be less urgent information whose transmission can be postponed, such as train CCTV images, or audio files used by the background music. In addition, the urgent or priority information is usually smaller than the non-priority information.

3.1 Towards a train-to-earth wireless communications architecture

In this section the core components of the mentioned wireless communications architecture are described. This architecture allows a full-duplex transmission of information between applications and devices deployed in the trains, and applications that are in the railway control centres. The description of the architecture will be made at two levels: conceptual and physical level. The first level defines the basic concepts of the architecture, and the second one illustrates the technologies used to implement the architecture in two real scenarios.

3.1.1 Conceptual level

From a conceptual point of view, two issues are especially important: the elements that manage architecture's behaviour and the ways in which the different applications (terrestrial and on-board) transmit the information.

Our architecture hosts both terrestrial and train-side applications, so in order to manage its behaviour two main entities are defined: Terrestrial Communication Manager (TCM) and On-board Communication Manager (OCM). The former manages terrestrial aspects of the architecture and the latter train-side issues. Although the managers have a different physical location, both of them have nearly the same responsibilities:

- Delivery and reception of the information,
- Dynamic train addressing,
- Medium access control,
- Security and Encryption, and
- Communication error management.

Due to the information transmission needs and for a correct and optimized used of the communication architecture, two types of communications are distinguished: "slight" and

“heavy” communications. These two types take into account characteristics of both information and communication technologies, such as: the volume and the priority of the information, the existence of coverage, and the cost of the communication.

- **Slight communications:** This type of communication is for the transmission of small volumes of information (few kB.) and with high priority. In general, information that has low latency (milliseconds or a pair of seconds) and needs to be transmitted exactly when it is generated or acquired (for instance, the GNSS location of a train, or a driving order to the train driver). For example, in the first case, if the information about positions is not sent immediately after its generation (real-time transmission), it loses all relevance.
- **Heavy communications:** This type of communication is tied to the transmission of large volumes of information (in the order of MB) and with low priority. The importance of this information is not affected by the passage of time, so it doesn't need to be transmitted at the exact time it is generated (no real-time transmission).

3.1.2 Physical level

In this section the technological aspects of the wireless architecture are described. They refer to the protocols and the communication technologies used for the development of the train-to-earth architecture (showed in Fig. 1).

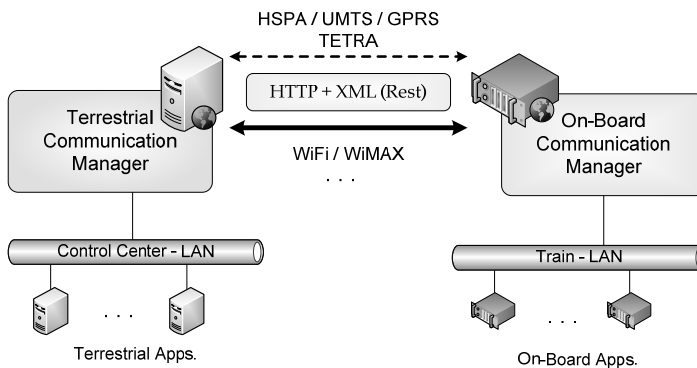


Fig. 1. Train-to-earth wireless communications architecture.

It is important to point out that the protocols and technologies for the development of the new architecture have been selected with regard to: standardization, robustness, security, scalability and compatibility with existing and potential applications and systems. The major aim has been the ease of integration of any application or system into the new communication architecture. Concerning with this objective, Web Services constitute the transport technology for the communication between final applications and the “local” Communication Managers. All the information is interchanged in XML format, in order to allow future extensions.

On the other hand, the communication between the Terrestrial (TCM) and the On-board Communication Managers (OCM) is based on REST (Representational State Transfer) technology. This communication technology uses the HTTP (HyperText Transfer Protocol) protocol and XML formatted messages. This solution is similar to traditional XML Web Services but with the benefit of a low overload and computational resources consumption.

Although the information interchanged between the TCM and the OCM is not encrypted, using the HTTP protocol allows the easy migration to HTTPS (HyperText Transfer Protocol Secure) that offers encryption and secure identification. It can be seen that every communication has to go through two core elements that can result in the loose of channel availability in case of failure. This problem is tackled by means of the use of web services because this solution deploys support web services in a way similar to traditional web architectures. It can be said that the selected technologies and architectures are well known and broadly used in different application areas or contexts, but they are novel in the railway 'train-to-earth' communication field.

In order to establish a wireless communications channel between the trains and the railway control centres, mobile and radio technologies have been selected (Yaipairoj et al., 2005). In this case, slight and heavy communications use different technologies due to different transmission characteristics.

Due to the necessity of delivery of information in real-time, mobile technologies such as GPRS/UMTS/HSPA (Gatti, 2002) are used for the *slight communications*. These technologies do not offer a great bandwidth nor a 100% coverage and they have a cost associated to the information transmission. Despite this, these technologies are a good choice for the delivery of high-priority and small sized information. The selection of the specific technology (GPRS/UMTS/HSPA) depends on whether the service is provided or not, (by a telecommunications service provider), and the coverage in a specific area.). To increase coverage availability, the hardware installed in each train has two phone cards belonging to different telephone providers. This allows switching from one to the other depending on coverage availability. Therefore, the idea is to have a predetermined operator, and only switch to the second when the former is unable to send.

On the other hand, for the *heavy communications*, WiFi radio technology has been chosen. This technology allows the transmission of large volumes of information, does not have any costs associated to the transmission and its deployment cost is not very expensive. In this case, a private net of access points is needed. This net does not need to cover the complete train route because the heavy communications are thought for the transmission of big amounts of information at the end of train service (for example the video recorded by the security cameras).

Although each separate technology can't achieve 100% coverage of the train route, the combination of both comes very close to complete coverage (Pinto et al., 2004). As the application layer protocols are standard, other radio technologies such as TETRA or WiMAX (Aguado et al, 2008) can easily substitute the ones selected now. These technologies can achieve a 100% coverage and neither one has a transmission cost. However, there are certain limitations such as the cost of deploying a private TETRA network, and the cost and the stage of maturity of the WiMAX technology

3.2 How to manage wifi based broadband communications

As it was explained previously, there are some railway applications that need a high bandwidth to interchange large amount of information without time restrictions (real time communication is not needed). This kind of communications will enable 'train-to-earth' information exchange for train side systems update/maintenance and multimedia information download/upload such as videos or pictures.

With the purpose of providing an innovative broadband communications architecture suitable for the railway, a number of WiFi networks have to be settled in places where the trains are stopped long enough to ensure the discharge of a certain amount of information. This is: stations in the header that starts or ends a tour and garages. In this way, WiFi coverage is not complete, but broadband communications are designed to update large amount of information, which, usually do not need to take place in real time.

Therefore, at this point it has to be taken into account aspects such as bandwidth, coverage or communications priorities. The existing broadband management systems, which are used in other (non mobile) environments, do not satisfy all the needs of the railway applications (California Software Labs, 2008; Marrero et al., 2008). Furthermore, some additional problems have to be solved on this environment. In one hand, it is necessary to find a mechanism to locate the trains because they don't have a known IP address all the time. A dynamic IP assignment is used for every WiFi network so a train obtains a different IP address every time it is connected to a network, and a certain IP address could be assigned to different trains in different moments. On the other hand, there are several applications that want to transmit information to/from the trains at the same time. This implies the existence of a bandwidth monopolization problem.

To tackle these challenges, it is necessary a smart intermediate element which manages when the applications (both terrestrial and train-side) can communicate with each other. That is to say: a Broadband 'train-to-earth' Communications Manager, whose design and functional architecture is described below.

3.2.1 Design of the broadband communications manager

The Broadband Communications Manager (BCM) (Carballedo et al., 2010a) is a system that arbitrates and distributes shifts to communicate terrestrial applications and train-side systems (see Fig. 2); in this way, the terrestrial applications request a turn when they want to establish a communication with a train. This distribution shift is managed on the basis of the state of the train connection to a WiFi network (known at all times) and a system of priorities, which are allocated according to the terrestrial application that wants to communicate with a specific train.

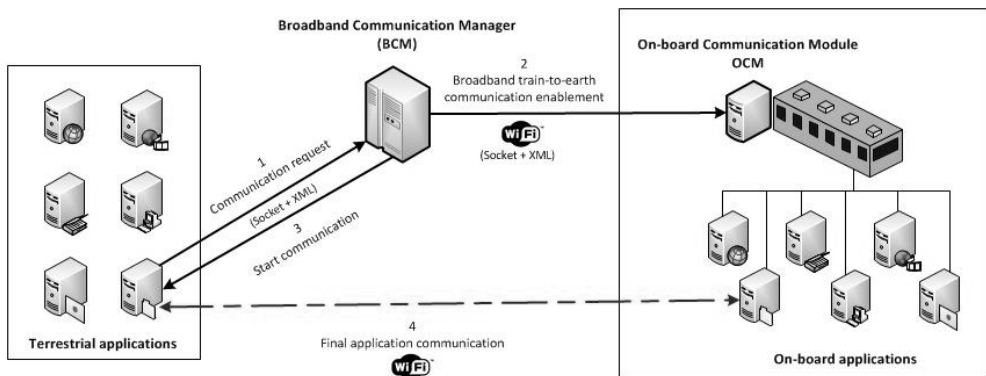


Fig. 2. The Broadband Communications Manager (BCM) arbitrating the communications between terrestrial applications and train-side systems.

1. **Communication establishment protocol.** When the BCM decides to give a shift to communicate a terrestrial application and a train-side system, it sends an authorization to both (application and train system). To do this, the manager establishes a communication with each entity through TCP Sockets. Within these TCP Sockets a series of XML messages, that define the communication protocol, are used.

To explain in a simple way the operation of the BCM, here is the description of a typical scenario:

- Firstly, a terrestrial application designed to communicate with a train system is connected to the manager through a TCP Socket.
- The terrestrial application will make communication request, and will give it a certain priority. By the time the manager receives the request, it orders the request in the queue of the destination train's requests. This queue is always sorted by different criteria.
- When a train arrives at a station, it connects to the WiFi network and it gets an IP address. This address is supplied to the BCM. If the train has pending communication requests, the terrestrial application is notified so that it can start the communication.
- At this moment there is a direct communication between the terrestrial application and the train-side system, through the WiFi network. The responsibility for starting the communication relies on the terrestrial application because it knows the IP address of the train.
- When the communication ends, the terrestrial application informs the BCM, which is ready to serve the next communication request.

It is important to emphasize that the BCM does not set any limitation or condition in the communication between the terrestrial application and the train-side system. The manager's work focuses only in defining the time at which this communication must be carried out, and warns of this fact to the entities involved in the information interchange. It does not define any structure or format of the information being exchanged; it only establishes a mechanism to know the IP address of the destination train (because it is dynamic), and manages the transmission shifts to prevent the monopolization of the communications channel.

2. **Multithreading management.** To carry out its work, the BCM must establish connections with multiple applications and train-side systems at the same time. To manage all these communications efficiently a multithreaded design has been chosen for the management of the connections. Every communication that the BCM performs with any external element (terrestrial applications and train-side systems) is carried out independently and concurrently, using a dedicated thread in each case.

Both the Train-Side Systems Handler and the Terrestrial Applications Handler (see below, Fig. 4) are separate threads that are responsible for receiving connections from external agents. Upon receiving the connection message specified in the protocol, they generate a separate communication thread with the element which has sent the message.

3. **XML based protocol for data transmission.** All the communications are done through an architecture based on TCP sockets (one for the terrestrial applications and another one for the train-side systems) and XML messages exchange. A message will be defined for each requests/responses exchanged between the three elements that form the

architecture: terrestrial applications, train-side systems and Broadband Communication Manager. Fig. 3 shows a XML message of the communication protocol.

```
<?xml version="1.0" encoding="UTF-8"?>
<request>
  <application name="CCTV" ip="130.88.10.56" />
  <train name="UT204" />
  <port number="3556" priority="1" />
</request>
```

Fig. 3. An XML message with a request from terrestrial application to the BCM asking a communication with an on-board system of the train UT204.

In order to communicate terrestrial applications, train-side systems and BCM a XML messages base protocol has been defined. The choice of the TCP Socket schema and XML messages was taken due to the flexibility to add new functionality, and the simplicity of implementation (independent of platform and programming language).

Moreover, all the data handled by the BCM is stored in a relational database. These data contain information about the communication requests, trains, train-side systems, terrestrial applications, and the available communication ports between them. The BCM's design contains a data layer that abstracts the data source of the business layer, so that the changes of this data by another for a different data source does not create any problems in the proper functioning of the BCM.

4. **Port to IP address translation schema.** To finish, we will make a brief description of the management of the applications installed on the trains (train-side systems), which are the target of the communication from terrestrial applications. These train-side systems are implemented on a computer that will have a private IP address (within the on-board Local Area Network) and is not accessible from outside the train. Therefore, it has been defined an addressing scheme to allow access from the IP address of the terrestrial application to the IP address of the train-side system. This is achieved through PAT filtering, associating each private IP address to a port number. Thus, whenever a train acquires an IP address from a WiFi network, the port number becomes the way to access the train-side systems. PAT filtering schema also ensures the security of communications and the information transmitted.

In each train there is a communications module which is responsible for performing this filtering of port numbers to IP addresses. This module is also responsible for communicating with the BCM, and manages the opening and closing of the ports that are associated to each train-side system.

3.2.2 Functional architecture of the broadband communications manager

Functional architecture of the BCM is based on message exchange between the manager itself and two types of external entities such as terrestrial applications and train-side systems. The BCM is divided into 5 modules (Fig. 4) that handle processing and deployment of all the functionality.

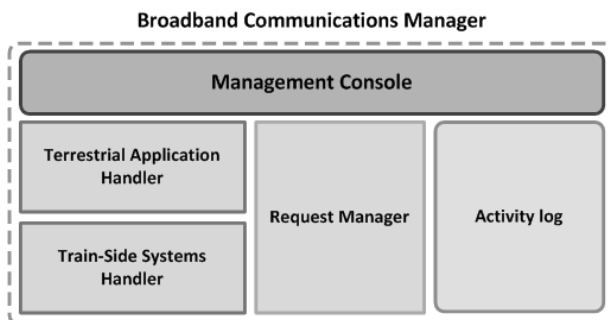


Fig. 4. Functional architecture of BCM, composed of five modules: Terrestrial Application Handler, Train-Side Systems Handler, Request Manager, Activity Log and Management Console.

To have a global vision of the performance of the BCM, it is necessary to focus on three modules which carry out the most important functionality:

1. **Terrestrial Applications Handler.** It will be responsible for managing all the messages exchanged between each terrestrial application and the BCM. Its basic functionality is to receive the XML messages coming from terrestrial applications and generate an appropriate response. This communication is bidirectional, and it is the responsibility of the terrestrial application to start and finish it.

To streamline the management of communications between terrestrial applications and the BCM, the connections are managed independently (through a dedicated thread). The main functionality offered by this module would be the next one:

- Establish and close the connection between terrestrial applications and the BCM.
 - Receive communication requests.
 - Send messages to a terrestrial application in order to start communication requests.
 - Receive communication completed messages from terrestrial applications.
2. **Train-Side Systems Handler.** It will be responsible for managing all the messages exchanged between the communication module of each train and the BCM.

This module is similar to the Terrestrial Applications Handler. It receives XML messages from the communication module of each train and generates the responses. In this case, the primary goal of the module is to indicate when a train is connected to a WiFi network and its IP address. This data is very important for terrestrial applications to communicate with train-side systems.

There is a very important task that Train Communication Module manages, it is: the disconnection or closure of the connection between the train and the BCM. When a train reaches a station with WiFi connectivity, it connects to the WiFi network and establishes a communication with the BCM. After that, two scenarios can occur: in the first one, the train has no pending communication request from terrestrial applications. In this case, the manager sends a connection ending message to the train and the connection is closed. In the second scenario, a train is disconnected from the WiFi network because of its movement or a

communication failure. The manager is constantly checking if the connection with the train is lost so this situation is detected as soon as it happens. There is a problem when the connection fails in the middle of a communication between a terrestrial application and a train-side system because the communication request has not finished correctly. To solve this problem, the next time the train connects to the BCM it sends back the start message of the broken communication to the terrestrial application in order to regain restart the communication. This pattern is repeated until the communication request is completed correctly, or is discarded because it exceeded the threshold of retries.

3. **Request Manager.** It will be responsible for managing communication requests between terrestrial applications and train-side systems, and to control when and under what circumstances the requests need to be attended.

As discussed above, the BCM splits communication shifts to terrestrial applications based on requests that they have performed. These requests are grouped by train, so the manager handles requests addressed to each train independently. The communication request for each train is sorted by the following criteria: (1) priority, which represents the 'urgency' by which a request must be addressed; (2) retries, it is taken into account the number of attempts to start a communication, to avoid the monopolization of the communication channel; and (3) parallelism, the manager can handle communications from multiple applications simultaneously with several trains.

The priorities associated with the communication requests are managed centrally and the BCM assigns these priorities to each terrestrial application. In addition, the manager also controls the train-side systems that can communicate with each single terrestrial application, identifying the ports that can be accessed by each of those terrestrial applications.

To complete the communication shifts service and management algorithm, it has prepared a final criteria, variable in this case (Noh-sam & Gil-Haeng, 2005). This approach takes into account two factors that are related directly with the communications that have been carried out previously. (1) The first factor is based on the calculation of average duration that takes the communications of a particular application. (2) The second factor takes into account the average duration of trains stopping in a particular station. Thus, the manager calculates a numeric value that represents the fitness of serving a request, knowing that the lower average duration of both factors will be most appropriate, since the risk of communication to be split because the train leaves the station will be less. Once calculated this criteria, it is used to discern which communication request is served, if the criteria explained a few paragraphs above is not sufficient

4. **Management Console (MC) and Activity Log (AL).** As for the remaining two modules, the MC contains a small management utility for monitoring the status of existing communication requests, and cancelling or changing the priority of unfinished requests. Through this interface it is possible to configure parameter and information settings of the BCM such as terrestrial applications, train-side systems, communication ports and priorities. The MC utility is based on standard design patterns like Model View Controller (MVC) so that in the future this presentation layer can be replaced by a more suitable one. The other supporting module is the AL. It stores each of the activities undertaken by the BCM.

To validate the improvement in the management of broadband communications produced by the BCM, there were a series of laboratory tests, which have subsequently been carried out in a real scenario. At first, the Broadband Communications Manager was tested in devising single communications between train and CCTV application. But to prove the performance improvement of the available bandwidth use, it has been necessary to include other terrestrial applications such as a document updating tool, and two other fictitious applications that simulate communications with the train.

The performance tests have taken into account two key parameters: 'train-to-earth' data transfer average time; and average waiting time between each communication. Table 1, shows the results obtained in the management of communications between four terrestrial applications (with different volume of data) and a train at the same station without the Broadband Communications Manager, while the second table shows the same scenario with the Broadband Communications Manager.

Data Volume (MB)	Data Transfer Time (seconds)	Waiting time (seconds)
< 1	1.10	0
1-10	11.30	0
11-50	58.84	0
51-100	184.62	0

Table 1. Results without the Broadband Communications Manager.

In the first table we can see that the absence of a communications manager allows communications to be made in parallel sharing the bandwidth. This greatly slows down the transfer rate, increasing the transfer time as the volume of information grows.

Data Volume (MB)	Data Transfer Time (seconds)	Waiting time (seconds)
< 1	0.76	0
1-10	7.69	0.76
11-50	38.46	8.45
51-100	115.38	49.91

Table 2. Results with the Broadband Communications Manager.

The second table shows how communications are conducted from smaller to larger amounts of data transferred thanks to the algorithm developed for the communication request service. The average time of transfer is lower than in Table 1, and the fact that communications are conducted one-by-one implies that there is a timeout that does not exist if they were carried out all at same time. At the conclusion of the tests it was determined that communications are carried out about 30% faster with the Broadband Communications Manager than without it, although there are wait times.

4. Considering the quality of service

In previous sections there have been described a train-to-earth communication architecture that enables two kinds of communications schemes: (1) slight communications and (2) heavy communications. Slight communications aims to respond priority and real-time application communication needs that no requires broadband communications bandwidth capabilities,

whereas heavy communications were designed to lower priority large information volumes transmission management with no real-time requirements.

Based on this previous work, future work aims to go a step further by combining both schemes mentioned before to enable a real-time broadband communication platform which responds to train-to-earth applications communication needs. Thus, the objective is to enable several physical network communication links between train and ground system, choosing the network link considered as the best at every moment according with the bandwidth availability. Not having final applications to get involved in the network management. So, the system should respond to several requirements:

- **High availability:** each train should be enabled with one or more physical network communication links (3G, WiFi, etc.). Providing continuous train-to-earth connectivity in order to respond to the final applications communications demand in real-time.
- **The best bandwidth:** the purpose of this platform is to enable real-time train-to-earth broadband communications, using the best possible bandwidth. Thus, the system will always select the physical link considered as the best in order to respond to final application communication requirements.
- **Quality of Service (QoS):** this solution aims to make a service quality management too. Therefore it is necessary to know the bandwidth availability offered by the network link which is active at every moment, as well as the bandwidth offered by the rest of communications links (although they are not being used). At this point it is essential to establish a set of connection procedures which permit to reserve a certain bandwidth for a particular communication.

Hence this broadband train-to-earth communication platform has three principal functions:

1. Multiple physical device management, considering the dynamic selection of the best one (best bandwidth) and their abstraction into a single virtual device.
2. QoS implementation enabling the reservation and release of channels (virtual links) with a given bandwidth.
3. Message routing.

So, to carry out the train-to-earth communications management and arbitration, presented solution manages a set of criteria for prioritizing final applications communication requirements which will focus primarily on the criticality of the information transmitted and the required bandwidth, as well as their chronological arrival order.

4.1 Capabilities of the Real-time broadband communications platform

The main capabilities of this communication platform are related to (1) communication prioritization, (2) selection of the physical communication network link and (3) train-to-earth information exchange management.

4.1.1 Communication prioritization

The objective is to prioritize train-to-earth communications based on several criteria so that the transmission of critical information have more priority over other information that need less "immediacy" when being transmitted.

Therefore, this platform proposes a set of communication requests prioritization criteria in order to respond final applications communication demand (both on ground and onboard train). So, these criteria are applied to establish the order of the communication requests in train-to-earth communication prioritization queues.

4.1.2 Selection of the physical communication network link

The communication platform is based on different physical communication link existence so that the combination of these independent links offers a continuous train-to-earth connectivity in order to respond to the final applications communications demand in real-time. Thus, depending on the status of each of these media and their characteristics and restrictions, the platform must be designed to utilize the link that offers better performance in order to provide a high availability.

The system is designed to select at every moment the physical link that is most favorable for communications. Therefore, taking into account the availability of enabled different physical links, the system selects always as active link one that offers the best bandwidth (based on the features and coverage of the physical link).

At this point it should be emphasized that the basis is that the system always defines a single train-to-earth network link as active for communications (most favorable). So, all communications will always be generated by the channel set as active (WiFi, GSM / GPRS, Tetra, etc.) regardless of the availability of other physical channels simultaneously.

4.1.3 Train-to-earth Information exchange management

The main feature of the system is to manage the transmission of real-time broadband train-to-earth information. Therefore, the platform has to offer:

- Real-time bidirectional communication between train and terrestrial applications allowing generating new digital services with quality of service (QoS) guarantees. Thus, different kind of information exchange between final applications (multimedia, text, bytes, etc.) has to be supported.
- Train-to-earth communication management without requiring the participation of the final applications. However, transmission retries are delegated to the application logic. When an application information transmission is cut by the platform (because there is another higher priority request), and then it is re-established, it is responsibility of this application to decide if it continues transmitting from the point where it had left, or if the transmission is restarted from the beginning.
- Changing the requests bandwidth allocation in cases where the data traffic on the physical environment allows platform to assign applications' communications a greater bandwidth than initially requested from them.

4.2 Design of the Real-time broadband communications platform

This platform defines two main entities (Fig. 5): Terrestrial Communications Manager (TCM) and On-board Communications Manager (OCM). The former manages terrestrial aspects of the architecture and the latter train-side issues.

TCM interact with the OCM installed on each train in order to (1) select the best network link available at each time and then (2) manage applications communication requests serving those that are considered most priority first.

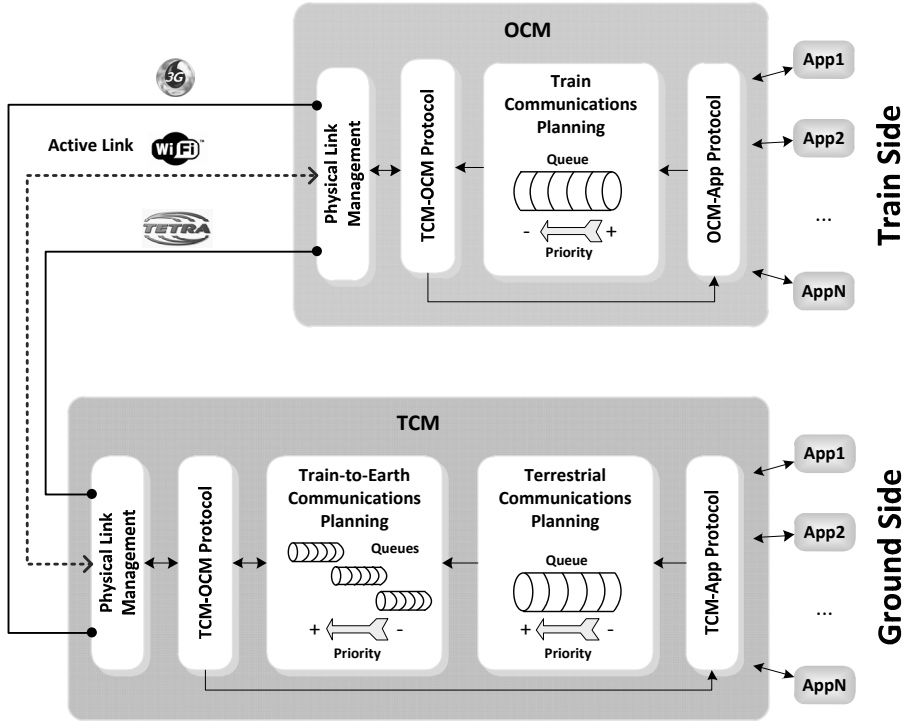


Fig. 5. Train and ground side components which conform the architecture of the real-time broadband communications platform.

4.2.1 Network active link selection

To establish train-to-earth communications, OCM and TCM can communicate through different communications network physical links. These two entities communicate each other to select the active link considered most favorable for communications. Therefore, OCM and TCM are continuously monitoring all enabled network link status, and switch from one to other in two cases: (1) when active link connectivity is lost and (2) when OCM and TCM select another link to be the new active link. In these two cases, the active communication link change is transparent for final applications that do not detect connection interruptions if these link changes occur while they are transmitting.

4.2.2 Priority application communication requests management

The broadband communication platform enables train and terrestrial railway applications to communicate each other. So when an application attempts to start a new communication

makes a communication request to the platform. Then the system make a decision about what priority requests can be served concurrently by the system taking into account active link bandwidth limitations and requests QoS requirements.

On the train side, the OCM must be able to prioritize communication requests made by the on board applications. So, the OCM queues train applications' requests in base of established prioritization criteria. Then taking into account communication active link bandwidth properties, the OCM notifies to TCM about the on board most priority requests that could be served concurrently by the system respecting these requests QoS requirements. TCM will ultimately decide and notify the OCM which communications can be addressed at every moment, considering the rest of the terrestrial applications' request.

Therefore, on the ground side the TCM will manage terrestrial applications' requests as OCM do in the train. Besides, TCM will have a queue for each train on the system containing that train's requests (notified by its OCM) and terrestrial requests in order to make decisions about what applications' requests can communicate at every moment.

5. Developing railway services over train-to-earth communications

Now we will explore the benefits of having a train-to-earth wireless communication technology like the one presented before. These benefits will be justified by mean of the new valued added railway services which will be able to be developed using this communication architecture. In this section we will show the functionality of two specific services as well as the way in which they interoperate with the train-to-earth wireless communication channel. The first one is a Backup Traffic Management Service (BTMS) which uses the slight communication infrastructure and the second one is a Remote Application Management Service (RAMS) which uses the heavy communication model and integrates with the broadband communication manager.

5.1 Backup Traffic Management Service (BTMS)

Security in railway industry is a critical issue. Intelligent Transportation Systems are becoming a very valuable way to fulfill these critical security requirements. In fact, today, rail traffic management is performed automatically using Centralized Traffic Control systems (CTC) (Ambegoda et al., 2008). These systems are based on sensors and different elements fixed on the tracks. They allow real-time traffic management: (a) location of trains, (b) states of the signals, (c) status of level crossings and (d) orientation of the needles. Most of the infrastructure management entities have a CTC that handles centralized all these issues. The applications and systems that handle these tasks are very robust and have a performance index near 100%. Problems occur when these systems fail. In those situations, traffic management has to be performed manually and through voice communications between traffic operators and railway drivers (Sciutto et al., 2007).

In this section web described a support system to assist traffic operators in emergency situations in which CTC systems fail. The main objective of this system is to reduce human error caused by the situations in which priority systems do not work properly.

5.1.1 Functional requirements

CTC traditional systems are centralized and rely on wired communications. When CTC system or communications fail, no one knows the location of trains, thus increasing the chances of an accident. In these situations, the railway companies put into operation its security procedures that transfer the responsibility of traffic management to traffic operators, who are people that monitor traffic in the terrestrial control centres. These people should manage the traffic manually communicating through analogical radio systems to the drivers of the trains. As people get nervous in emergency situations and that leads to mistakes, the new service aims to reduce these errors by creating a new tool to help traffic operators in emergency situations. This new tool must be based on different technologies to those used by traditional CTC systems so that failure in the former does not cause failure in the latter.

Taking into account these motivations and requirements, a Backup Traffic Management Service (henceforth BTMS) has been developed (Carballedo et al., 2010b). This service will assist traffic operators when the primary system fails. The main functions of this new system are:

- **Traffic situation representation for the track stretches where the main system do not provide information.** The new service represents the affected line stretches situation (train locations, track section occupation states, etc.) from information received from train-side systems through real-time wireless 'train-to-earth' communications (see Fig. 6).
- **Traffic management environment.** The objective is to provide a traffic assistance application in order to assist operators in tasks related to traffic control when the main system fails partial or totally.
- **Statistical analysis.** About aspects related to the system performance and reliability.
- **Control message sending from control centre to trains.** This functionality will allow traffic operators to send messages to the train drivers in order to manage and control the traffic.

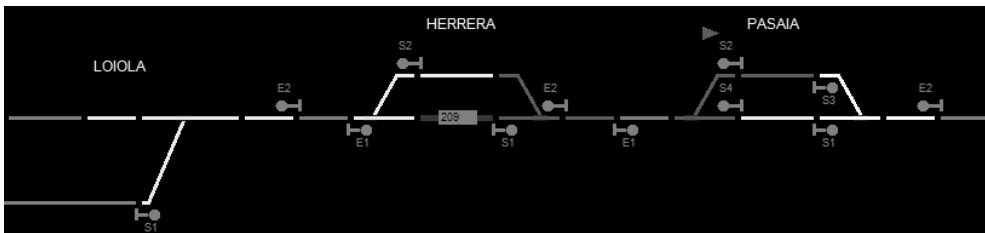


Fig. 6. Traffic situation representation.

The BTMS provides a traffic assistance application that works independently of the main CTC system. Thus, the new service is based on an application that informs about the position of the trains on track and permits to make tasks related to traffic management and control in an easier way. Moreover, this system permits a new way of communication between the traffic operators and trains drivers: exchanging control messages.

5.1.2 Architectural and design issues

In this section, we describe the most important technical considerations about the BTMS. The main issues are those related to (1) trains positioning, (2) wireless ‘train-to-earth’ communications and (3) added value services. These three issues are described below.

1. **Train positioning system.** The BTMS permits a new way of train positioning which works independently of the main system operation. In order to achieve this target, a new hardware/software module is boarded on each train. This module combines the positioning data provided by some hardware devices (accelerometers, gyroscope, odometer, etc.) with the coordinates given by a GPS module. To generate the most accurate positioning information, this system parts from a railway lines different tabulation ways. In this case, the tabulation is related to lines lengths (in kilometres) and the traffic signals positions. Based on this information, and the data extracted from the hardware and software modules boarded on trains (including GPS), this system translates this information to kilometric points (Shang-Guan et al., 2009). Then this positioning information is sent to the control centre in real-time. Besides, the BTMS communicates with an external positioning information system which permits the reception of train positioning information generated by the main CTC system.
2. **Real-time train-to-earth wireless communications.** The BTMS permits real-time train traffic management, so it is necessary to enable a real-time wireless communication channel between the BTMS installed on the control centre and the trains. For this reason, this service needs to use the previously explained train-to-earth wireless communications architecture, which enables slight communications and is based on mobile technologies (Aguado et al., 2005).
3. **Additional Services.** Using the mentioned train-to-earth communication architecture and the information provided by the on board positioning system positioning, two services have been developed related with the functionality of the BTMS.

The first one is a *Statistical Analysis Service*, which using the information stored by the positioning system on a data base, the BTMS can make statistical analysis related to the system’s reliability level, GPS and GPRS coverage, and other system functionality aspects. Thus, one of the main goals of this service is to compare the received information, determining if the positioning provided by the train-side systems is according to the information generated by the primary CTC system. And the second is a *Control Message Exchanged Service*, which allows the procedural alarms transmission to the train-side systems. These kinds of alarms indicate anomalous situations to the train drivers: primary system failure, signal exceeds authorization to a certain point as a consequence of a failure of any electro-mechanical track component, etc.

5.2 Remote Application Management Service (Remote-AMS)

One of the main objectives of applications running inside the train is to provide information in order to facilitate the work of the train driver. Usually these applications need to use information generated in the ground centre. If this information changes, it needs to be updated remotely. In addition, there are terrestrial applications that use information generated by some on board applications. Therefore, it is necessary to be able to download that information from trains.

In order to resolve these issues, a new service that allows the remote management of on board applications (upgrade, download and deletion of information) will be proposed. This new service is composed by a terrestrial software module and an on board one that communicate each other to permit this remote management. So, this system would be integrated with the communication architecture described before based in heavy communications scheme (voluminous information with no real-time requirements).

5.2.1 Functionality

The main functionality of this service is to control the updating of the information used by applications running on the train terminal (for example track flat information or supporting documentation for the driver generated by the ground information systems) as well as downloading and deleting information generated by some on board applications (for example log files) remotely from the ground centre.

The solution consists of two software applications, one for the "ground" (control centre) and the other to be deployed in all train terminals. These applications are integrated with the previously described connectivity architecture via heavy communication since it involves the exchange of large volumes of information that do not require real-time communications.

Therefore, in this case, the Broadband Communication Manager (BCM) will be responsible for the arbitration of the train-to-earth heavy communication requests which are made by the software application running on the ground centre. This terrestrial application aims to communicate with those applications running inside the train to carry out all tasks related to the described service.

Thus, the terrestrial Remote-AMS application is installed on ground centre, and it will be responsible for managing the status of all applications in each terminal. On the other hand, on board Remote-AMS application will handle update, download and deletion requests made by the terrestrial Remote-AMS application. So, **terrestrial and on board Remote-AMS functionality** involves these issues:

- Knowledge about the configuration information for each application installed on each train terminal at any time (version, creation date, last updated date, etc).
- Knowledge about files and/or documents used by each application that can be updated, downloaded and/or deleted. This information will include: version, creation and last update date, update status (pending or not), etc. This management is done through a repository of information in a database.
- Management of information (files) of all current and future applications installed in the train terminals. For such management the Remote-AMS service will use FTP (File Transfer Protocol) configured to work locally (inside the train). So, the technology used is well known and standardized for remote management information.
- Management of the updating of configuration information of the applications installed on the train terminals.
- Management of the download of the information generated by the applications running on the train terminals.
- Management of the deletion of obsolete information in train terminals.

- Manage queries about the status information of the on board applications.
- Integration with the train-to-earth wireless communication architecture via heavy communications scheme.

5.2.2 Architecture

As mentioned before, this service is integrated with the previously described connectivity architecture via heavy communications scheme. So, when the terrestrial Remote-AMS generate tasks which involve downloading or uploading information from and to trains, it have to communicate with Broadband Communications Manager (BCM), because this is the entity who arbitrates heavy communications between ground and train applications. In this case, BCM arbitrates communications between terrestrial and on board Remote-AMS application. For proper integration with BCM, Remote-AMS (more specifically the terrestrial one) shall be compliant with the protocol of communication established by this management entity.

At this point it is important to remember that BCM does not interfere between final applications communication. The Fig. 7 shows Remote-AMS service architecture and its integration with the train-to-earth wireless communication technology.

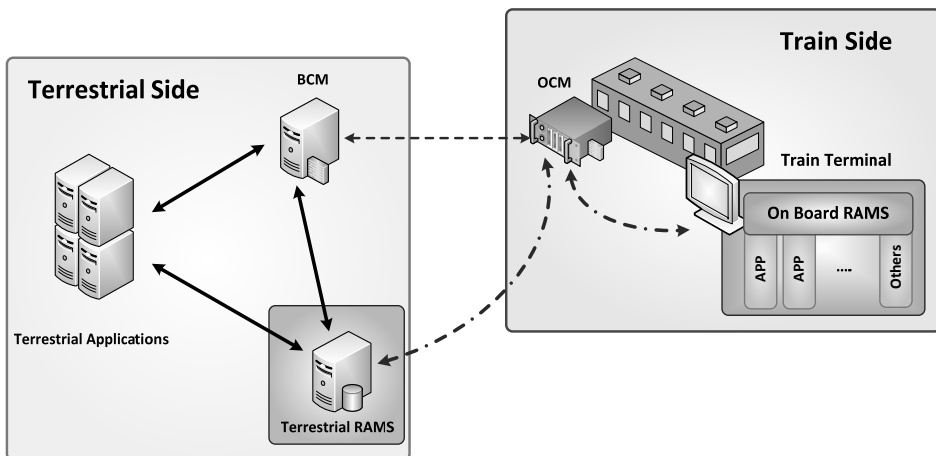


Fig. 7. Remote Application Management Service (Remote-AMS) and how it interoperates with the Broadband Communication Manager (BCM).

So, whenever terrestrial Remote-AMS schedules a task it has to send to the BCM a connection message. Once connected, there will be many communication requests as required. When BCM determines that a Remote-AMS request should be addressed, sends a notification to terrestrial Remote-AMS indicating to perform the service task corresponding to this request. Once the communication is completed, terrestrial Remote-AMS sends a notification to BCM which sets this request as completed and removes it from the corresponding communication prioritization queue. This same pattern is followed for all terrestrial Remote-AMS communication requests.

6. Future work

As future lines of work, the major efforts in train-to-earth wireless communication are focused on the improvement of the capabilities of the communication channel. Concerning with these improvements the followings are two of the hot topic to deal with:

1. **Communication network virtualization.** Where the technology and the terrestrial platform which takes part in the digital contents interchange are selected transparently to the front-end back-end applications depending on which technology suits better the communication features (information volume, nature and priority, communications cost; coverage, and so on).
Nowadays, the wireless technologies available for this kind of communication between the railway and the terrestrial platform are: WiMax, WiFi, GPRS, UMTS, TETRA or GSM-R. Nevertheless, the proposed solutions have to be compatible with any other future communication technology.
2. **Hybrid self-managed and shared communication channel.** The challenge is to design a shared communications channel to be used for all applications, regardless of current or future functionality of those applications. With this new communications scenario, applications will only provide the information to be transmitted, and the destination of communication (being hidden protocols and the complexities of the communication). The channel itself will decide when is the right time to send the information and what is the best technology.

Apart from the improvements of the communications architecture, the design of a **framework for vertical services deployment** could be very interesting. This framework would offer an easy and seamless integration of new applications with the wireless communication infrastructure and with other future horizontal services. This infrastructure is based on standards and technological paradigms which are highly/ long enough proved in different environments. Furthermore, their interoperability and integration benefits are sufficiently contrasted; one example is the SOA (Service Oriented Architecture) paradigm case. It would be interesting to adopt the Software Engineering best practices and standard of interoperability used in other areas in the railway industry.

Finally, the proposed infrastructure would boost the **development of new vertical services** which can be classified in four categories: (1) driver assistance services, (2) services for passengers, (3) freights tracking services (based on RFID technology) and (4) services for train health monitoring. All these services have in common the need for exchange of digital content (often multimedia) between trains and ground control centres. The ubiquitous nature of connectivity that is provided by the new communications scenario will improve existing railway applications. And furthermore, will facilitate the development of new context-aware and customized services for end users. These advanced features result in fundamental improvements in the field of rail services.

7. Conclusion

In the railway industry, communications were born almost exclusively for the purpose of managing and regulating traffic flow, requiring by the mobile nature of this sector two modes of communication: those that occur between fixed elements of the rail infrastructure, which are based mainly in wired systems; and those which participate in the fixed and

mobile elements (communications known as "train-to-earth"), which require a wireless communication channel and traditionally have materialized on the use of analogical communication systems such as traditional phone or radio.

Today, despite the maturity of the railroad industry and the advances in wireless communications technologies, the rail industry continues to base the operation of its priority services in analogical and wired communication technologies, which in fact belong to the past but still are efficient and robust.

New generation of wireless communications technologies, such as those based on conventional mobile technology (GSM, GPRS or UMTS), or broadband solutions (such as WiFi or WiMax), opens countless possibilities of use in the railway industry. As the cost of their deployment is very low, they perfectly complement traditional communication systems, and they have wide bandwidth and wide coverage that enable the deployment of new generation services in this area, some of them directly related to the end user, in order to provide a high quality transport service. On the other hand, the rise of high-speed trains is facilitating EXPANSION GSM-R as a basis for communication between signaling and regulation systems for the railway traffic.

This is precisely the topic on which this chapter is focused. It described a specific architecture of next-generation wireless communications for the rail industry to establish a train-to-earth bidirectional communication channel. This architecture is a single channel of communication between all train applications and those in the control centres. The aim of this communication channel is to standardize the way the data is transmitted between them. Thus, this channel is a resource shared by all the applications that simplifies the complex details related to communications and provides advanced services oriented to communication; services such as the selective treatment of the transmissions based on the nature and volume of the information, the location of the messages destination, the management of priorities and arbitration of communication shifts, attempts management, and so on. Moreover, we illustrate the challenges of bandwidth management in railway wireless broadband communications, and how we have faced to them. We have designed a new system that distributes communication shifts between terrestrial applications and train systems, which require the exchange of large amounts of information.

This chapter summarized the results of the research in train-to-earth wireless communications done during the last five years in collaboration with train manufacturers, railway technology providers and railway operators. Our wireless communications architecture has been incorporated into the manufacturing process of a new series of trains, which is a European-wide revolution since it enables wireless and transparent communication between terrestrial applications and those which are deployed on the trains. Furthermore, this architecture is being the basis for new digital services currently under development which will be in the market in a short time. They have different nature and purpose: from services that control the status of the train, to services for the end-user, and support systems for the train drivers.

8. Acknowledgment

This research was partially supported by Eusko Trenbideak - Ferrocarriles Vascos (a railway company from the north of Spain), Innovate and Transport Engineering (a railway

technology provider), the Ministry of Industry, Tourism and Trade of Spain under the Avanza funding program (Grant TSI-020501-2008-148), and the Basque Country Government under the Euskadi+09 funding program (Grant UE09+/70). This support is gratefully acknowledged.

9. References

- Aguado, M.; Jacob, E.; Saiz, P.; Unzilla, J.J.; Higuero, M.V. & Matias, J. (2005). Railway signaling systems and new trends in wireless data communication, *Proceedings of IEEE 62nd Vehicular Technology Conference*, pp. 1333-1336, ISBN: 0-7803-9152-7, Sept 2005.
- Aguado, M.; Onandi, O.; Agustin, P.S.; Higuero, M. & Jacob Taquet, E. (2008). WiMax on Rails: A Broadband Communication Architecture for CBTC Systems, *IEEE Vehicular Technology Magazine*, Vol. 3, No. 3, pp. 47-56, ISSN: 1556-6072, Nov 2008.
- Ambegoda, A.L.A.T.D.; De Silva, W.T.S.; Hemachandra, K.T.; Samarasinghe, T.N. & Samarasinghe, A.T.L.K. (2008). Centralized Traffic Controlling System for Sri Lanka Railways, in *Proc. of 4th International Conference on Information and Automation for Sustainability (ICIAFS 2008)*, Colombo, Sri Lanka, pp. 145-149, ISBN 978-1-4244-2899-1, Dec 2008.
- Berrios, A. (2007). Las comunicaciones ferroviarias: avances y reto. *Anales de mecánica y electricidad*, Vol. 84, No. 1, pp. 64-69.
- California Software Labs (2008). *Extending end-to-end QoS to WiFi based WLAN*, CSWL whitepaper, <http://www.cswl.com/whitepapers/qos-wireless-lan.html>.
- Carballedo, R., Perallos, A., Salaberria, I. & Gutiérrez, U. Managing 'train-to-earth' heavy communications: A middleware software to manage broadband wireless communications in the railway scope, in *Proc. of International Conference on Wireless Information Networks and Systems (WINSYS 2010) - 7th IEEE International Joint Conference on e-Business and Telecommunications (ICETE)*, Athens, Greece, pp. 1-6, July 2010.
- Carballedo, R., Perallos, A., Salaberria, I., Odriozola, I. & Gutiérrez, U. A backup system based on a decentralized positioning system for managing the railway traffic in emergency situations, in *Proc. of the 13th International IEEE Conference on Intelligent Transportation Systems (ITSC 2010)*, Madeira Island, Portugal, pp. 285-290, Sep. 2010.
- European Telecommunications Standards Institute (ETSI) (2008). *Electromagnetic compatibility and Radio spectrum Matters (ERM); Digital Mobile Radio (DMR) General System Design*, technical report: ETSI TR 102 398 V1.1.2, May 2008.
- European Telecommunications Standards Institute (ETSI) (2011). *TERrestrial Trunked RAdio (TETRA)*, <http://www.etsi.org/website/Technologies/TETRA.aspx>
- Garstenauer, J. & Pocuca, S. (2011). The future of railway communications, *Proceedings of the 34th International Convention MIPRO*, pp. 421-423, ISBN: 978-1-4577-0996-8, Opatija, Croatia, May 2011.
- Gatti, A. (2002). Trains as Mobile devices: the TrainCom Project. *Wireless Design Conference*, London, 2002.
- Gutiérrez, U., Salaberria, I., Perallos, A. & Carballedo, R. Towards a Broadband Communications Manager to regulate train-to-earth communications, in *Proc of 15th IEEE Mediterranean Electrotechnical Conference (MELECON 2010)*, La Valletta, Malta, pp. 1600-1605, May 2010.

- IEEE 802.11 (2007). *IEEE 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*, ISBN: 0-7381-5655-8, June 2007.
- IEEE 802.16.2 (2004). *IEEE Recommended Practice for Local and metropolitan area networks - Coexistence of Fixed Broadband Wireless Access Systems*, ISBN: 0-7381-3986-6, 2004.
- International Union of Railways (UIC) (2011). *GSM-R: the railway system for mobile communications*, <http://www.uic.asso.fr/uic/spip.php?rubrique851>
- Laplante, P.A. & Woolsey, F.C. (2003). IEEE 1473: An open source communications protocol for railway vehicles, *IEEE Computer Society*, Vol. 5, Issue 6, pp. 12-16, ISSN: 1520-9202, Nov. 2003.
- Marrero, D.; Macías, E.M. & Suárez, A. (2008). An admission control and traffic regulation mechanism for infrastructure WiFi networks, *IAENG International Journal of Computer Science*, Vol. 35, Issue 1, pp. 154-160, ISSN: 1819-656X, 2008.
- Noh-sam P. & Gil-Haeng, L. (2005). A framework for policy-based sla management over wireless lan, in *Proceedings of the Second International Conference on e-Business and Telecommunication Networks (ICETE 2005)*, pp. 173-176, INSTICC Press 2005, ISBN 972-8865-32-5, Reading, UK, October 2005.
- Pinto, P.; Bernardo, L. & Sobral, P. (2004). Service integration between wireless systems: A core-level approach to internetworking, in *Proc. of 1st International Conference on E-Business and Telecommunication Networks (ICETE 2004)*, pp. 127-134, INSTICC Press 2004, ISBN 972-8865-15-5, Setúbal, Portugal, August 2004.
- Salaberria, I.; Carballedo, R.; Gutierrez, U. & Perallos, A. (2009). Wireless Communications Architecture for "Train-to-Earth" Communication in the Railway Industry, *Proceedings of 6th International Symposium on Distributed Computing and Artificial Intelligence 2009 (DCAI2009)*, S. Omatu et al. (Eds.): IWANN 2009, Part II, LNCS 5518, Springer-Verlag, ISBN: 3-642-02480-7, pp. 625-632. June 2009.
- Sciutto, G.; Lucchini, M.; Mazzini, D. & Veglia, C. (2007). Technologies to Support the Railway Circulation in Emergency Conditions, in *Proc. of 2nd International Conference Safety and Security Engineering*, Malta, 2007.
- Shafiullah, G.; Gyasi-Agyei, A. & Wolfs, P.J. (2007). Survey of Wireless Communications Applications in the Railway Industry, *Proceedings of 2nd International Conference on Wireless Broadband and Ultra-Wideband Communications (AusWireless)*, pp. 27-30, ISBN: 9780769528465, Sidney, Australie, Aug 2007.
- Shang-Guan, W.; Cai, B-G.; Wang, J. & Liu, J. (2009). Research of Train Control System Special Database and Position Matching Algorithm," in *Proc. of IEEE Intelligent Vehicles Symposium*, Xian, China, pp. 1039-1044, ISBN 978-1-4244-3503-6, June 2009.
- Yaipairoj, S.; Harmantzis, F. & Gunasekaran, V. (2005). A Pricing Model of GPRS Networks with Wi-Fi Integration for "Heavy" Data Users, in *Proceedings of the Second International Conference on e-Business and Telecommunication Networks (ICETE 2005)*, pp. 80-85, INSTICC Press 2005, ISBN 972-8865-32-5, Reading, UK, October 2005.

Super-Broadband Wireless Access Network

Seyed Reza Abdollahi, H.S. Al-Raweshidy and T.J. Owens
*WNCC, School of Eng. and Design,
Brunel University, Uxbridge, London
UK*

1. Introduction

Today's communication network deployment is driven by the requirement to send, receive, hand off, and deliver voice, video, and data communications from one end-user to another. Current deployment strategies result in end-to-end networks composed of the interconnection of networks each of which can be classified as falling into one of three main categories of network: core, metropolitan and access network. Each component network of the end-to-end communication network performs different roles. Nowadays, the increase in the number and size of access networks is the biggest contributor to the rapid expansion of communication networks that transport information such as voice, video and data from one end-user to another one via wired, wireless, or converged wired and wireless technologies. Such services are commonly marketed collectively as a triple play service, a term which typically refers to the provision of high-speed Internet access, cable television, and telephone services over a single broadband connection. The metropolitan networks perform a key role in tripleplay service provision in delivering the service traffic to a multiplicity of access networks that provide service coverage across a clearly defined geographical area such as a city over fiber or wireless technologies infrastructure. The core networks or long haul networks are those parts of the end-to-end communication network that interconnect the metropolitan area networks. The core network infrastructure includes optical routers, switches, multiplexers and demultiplexers, used to deliver triple play service traffic to the metropolitan networks and route traffic from one metropolitan network to another.

Fig. 1, shows a simplified diagram of network connecting tripleplay service providers to end-users of the service. In this network, the uplink traffic from the end-users is input to the network via wireless or wired access network connections in the user's home. The packets associated with this traffic are multiplexed together and forwarded to the local metropolitan network for delivery to a long haul network for transporting to the service providers' access network and hence to the service provider. The downlink traffic from different service providers which is typically traffic corresponding to requested services is input to the network via local access network connections in the service provider premises. The downlink traffic from a particular access network is multiplexed together and delivered to the local metropolitan network for forwarding to a core network (or in some cases another metropolitan network) and hence to the end-users access networks for delivery to the end users. As many access networks are connected to a metropolitan

network the traffic data rates throughout a metropolitan network are significantly higher than those throughout an access network. As many metropolitan networks feed traffic into a core network the traffic handling capabilities of a core network are significantly higher than those of a metropolitan network. The network traffic on core networks is expected to reach the order of hundreds exabytes in the near future, (Laskar et al., 2007). The rapidly changing face of networked communications has seen a continued growth in the need to transfer enormous amounts of information across large distances. A consequence of this is that technologies that are used extensively for transferring information such as coaxial cable, satellite, and microwave radio are rapidly running out of spare capacity, (McDonough, 2007).

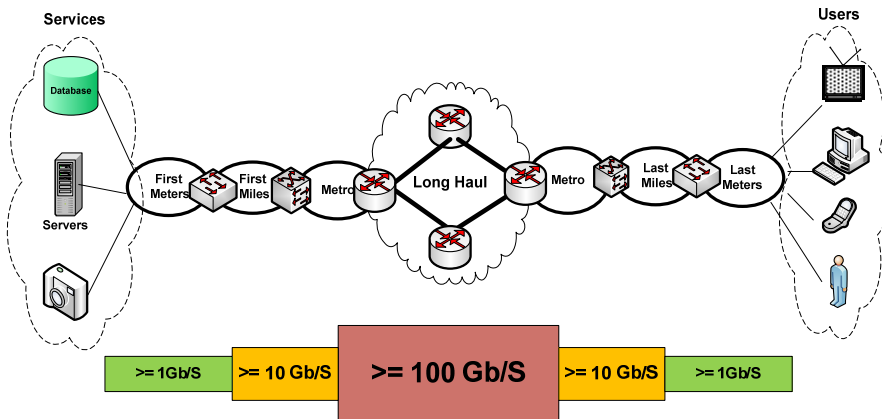


Fig. 1. Near term future network capacity requirements.

Therefore, transportation of the traffic volumes that will be demanded by users in the near future will require significantly greater network transmission bandwidth than that provided by the current infrastructure. Consequently, in the near term each category of component network of existing end-to-end networks will face different and increasingly difficult challenges with respect to transmission speed, cost, interference, reliability, and delivery of the demanded traffic to or from end-users. Currently, super-broadband penetration and the on-going growth in the internet traffic to and from business and home users are placing a huge bandwidth demand on the existing infrastructure.

Broadband wireless sits at the confluence of two of the most remarkable growth stories of the telecommunication industry in recent years. Wireless and broadband have each enjoyed rapid mass-market adoption. Wireless mobile services grew from 11 million subscribers worldwide in 1990 to more than 5 billion by the end of 2010. The world's largest manufacturer of mobile phones has forecast that the number of mobile users accessing the internet via mobile broadband will grow to over 2 billion globally by the end 2014. Fixed broadband subscribers numbered only 57,000 in 1998 and rapidly increased to 555 million subscribers by the end of 2009. The number of fixed broadband subscribers is projected to exceed 720 million by 2015 despite the current economic situation, (OASE, 2010; ITU, 2011). The growth in the numbers of mobile telephone

subscribers, broadband and internet users over the last decade and the projections for the growth in these numbers are depicted in Fig. 2.

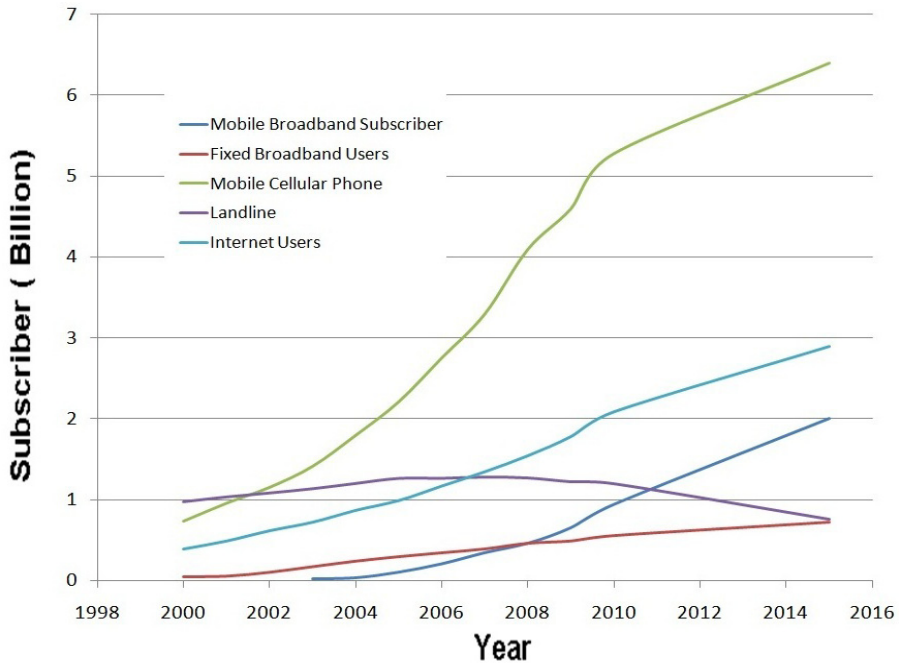


Fig. 2. Worldwide subscriber growth in the numbers of mobile telephony, internet, and broadband access users.

It follows that the demand for use of the available radio spectrum is very high, with terrestrial mobile phone and broadband internet systems being just one of many types of access technology vying for bandwidth. Mobile telephony and internet applications require the systems that support them to operate reliably in non-line-of-sight environments with a propagation distance of 0.5-30 km, and at velocities up to 100 km/h or higher. These operating environment constraints limit the maximum radio frequency the systems can use as operating at very high frequencies, i.e. approaching microwave frequencies, results in excessive channel path loss, and excessive Doppler spread at high velocity. This limits the spectrum suitable for mobile applications making the value of the radio spectrum extremely high. As an example, in Europe auctions of 3G licenses for the use of radio spectrum began in 1999. In the United Kingdom, 90 MHz of bandwidth was auctioned off for £22.5 billion (GBP). In Germany, the result was similar, with 100 MHz of bandwidth raising \$46 billion (US). This represents a value of around \$ 450 million (US) per MHz. The duration of these license agreements is 20 years. Therefore, it is vitally

important that the spectral efficiency of the communication system should be maximized, as this one of the main limitations to providing low cost high data rate services, (OMEGA ICT Project, 2011; Yuen et al., 2004). By deploying converged fiber and wireless communication (Fi-Wi) technologies, network operators and service providers can meet the challenges of providing low cost high data rate services to wireless users. Only the relatively huge bandwidth of a fiber-optic access network can currently support low cost high data rate services for wired and wireless users.

This chapter makes the case for radio over fiber (RoF) networks as a future proof solution for supporting super-broadband services in a reliable, cost-effective, and environmentally friendly way.

This chapter is organized as follows: In Section II, the evolution of Internet traffic driven by the growth in wired and wireless subscribers worldwide is discussed. In Section III, solutions for cost effective transportation of traffic volumes in line with the demand expected as a result of anticipated growth in interactive video, voice communication and data services are presented. In Section IV, the radio over fiber (RoF) network as a future proof solution for supporting super-broadband services is described as a reliable, cost-effective and environmentally friendly technology. Finally, concluding remarks are given in Section V.

2. Evolution of data traffic and future demand

Globally, mobile communication data traffic is expected to increase 26-fold between 2010 and 2015 and reach 6.3 exabytes per month by 2015. Furthermore, the compound annual growth rate (CAGR) of mobile data traffic is expected to reach 92 percent over the period 2010 to 2015. Moreover, during 7 years from 2005 to 2012 mobile data traffic will have increased a thousand-fold. In 2010, about 49.8% of mobile data traffic was video traffic. By deploying a converged fiber and wireless communication (Fi-Wi) technologies, the operators and service providers can meet the challenges they face from the continued dramatic growth in mobile data traffic volumes.

By the end of 2011, video traffic over mobile networks reached about 52.8% of the total traffic on mobile networks. It is expected that almost 67% of the world's total mobile traffic will be video by 2015 and that the volume of video traffic on mobile networks will have doubled every year over the period 2010 to 2015, (FP7, 2010, Cisco Visual Networking Index, 2011). In Fig. 3, the worldwide growth in data traffic rates per month are compared for mobile terminals and other devices. Fig. 3 (a) shows the anticipated growth of data traffic by user terminal type for the following terminal types: tablets, machine-to-machine (M2M), home gateways, smartphones, laptops, non-smartphones, and other portable devices. It is predicted that in 2015 82.4% of all network data traffic, about 5.768 exabytes per month, will be being transported to and from just by two types of portable wireless devices. Specifically, it is predicted that 55.8% and 26.6% of all network data traffic will relate to laptop and smartphone users, respectively. As shown in Fig. 3 (b), the expectation is that the data traffic rate relating to mobile devices will be about 6.3 exabytes per month by the end of 2015, (Cisco Visual Networking Index, 2011).

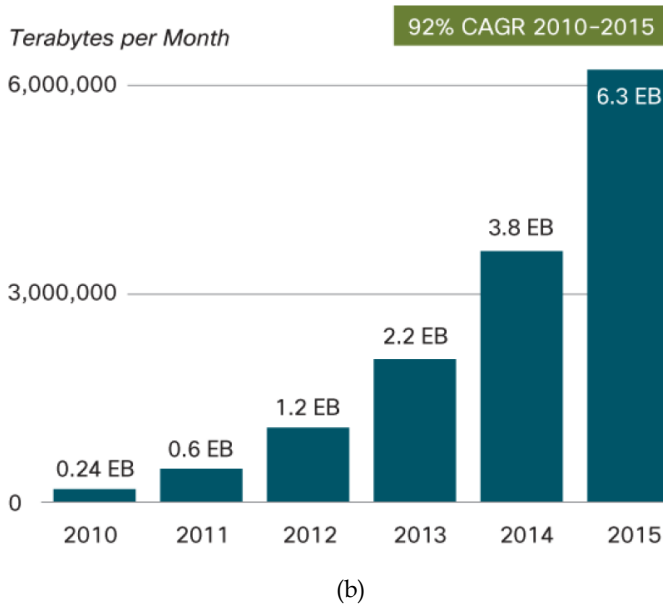
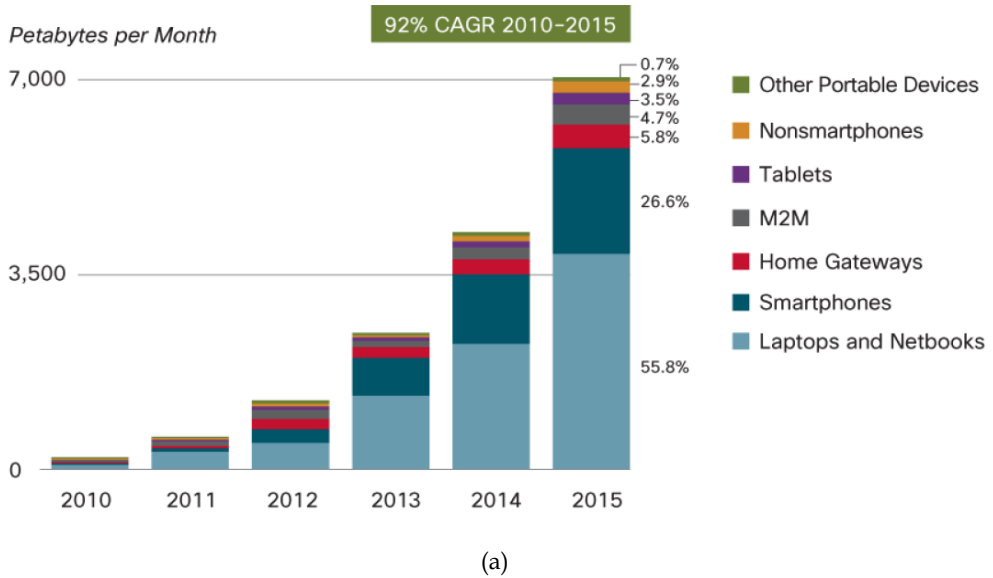


Fig. 3. The anticipated growth of data traffic (a): by user terminal type , (b) forecast of mobile data traffic growth by 2015, (Cisco Visual Networking Index, 2011).

High-Definition Television (HDTV) can now be provided in many countries throughout the world while Ultra High Definition Television (UHDTV) is now being studied in Japan as the most promising candidate for next-generation television beyond HDTV, and Super-High-

Definition Television (SHDTV). UHDTV consists of extremely high-resolution imagery and multi-channel 3D video and sound to give viewers a stronger sensation of presence. The UHDTV project's commercializing outlook is to become available in domestic homes over the period 2016 to 2020. For example, in 2005, NHK demonstrated a live relay of a UHDTV program using dense wavelength division multiplexing (DWDM) with 24 Gbit/s speed over a distance of 260 km on a fiber optic network. In 2006 NHK demonstrated a solution for bandwidth efficient delivery of UHDTV, utilizing a codec developed by NHK the video was compressed from 24 Gbit/s to 180–600 Mbit/s and the audio was compressed from 28 Mbit/s to 7–28 Mbit/s, (Sugawara et al., 2007; Kudo, 2005).

3. Deployment of super-broadband services

Globally the evolution of internet video services will be in the three following phases: 1) experiencing a growth of internet video as viewed on the PC, 2) internet delivery of video to the TV, and 3) interactive video communications, Fig. 4. Considering the future ultra high, super high and high definition resolution of end-user demanded and generated data traffic, each phase will impact on a different aspect of the end-to-end delivery network such as bandwidth, spectral efficiency, cost, power consumption, architecture, and technology. In addition to internet video, there is very high growth in the internet protocol (IP) transport of cable and mobile IPTV, and video on-demand services, (OASE, 2010).

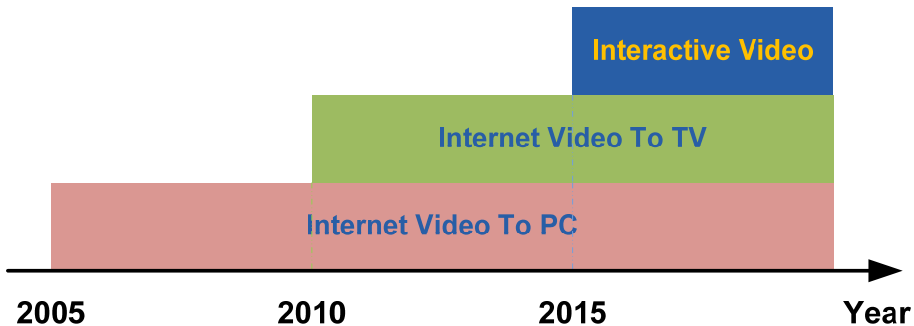


Fig. 4. Three waves of consumer Internet traffic growth.

Mobile voice services are already considered a necessity by many end-users, and mobile data, video, and TV are now becoming an essential part of some end-users' lives. The number of mobile subscribers' is growing rapidly and is expected to reach over 6.2 billion subscribers by 2015. Mobile users' bandwidth demand due to video services is increasing. Therefore, there is an essential need to increase the capacity of delivery networks for mobile broadband, data access, and video services to retain subscribers as well as keep cost in check.

Major considerations in planning the deployment of next-generation mobile networks are an increasing need for service portability and interoperability driven by the proliferation of mobile and portable digital devices and an accompanying need for the networks to enable

such devices, including smartphones, tablets, laptops, and non-smartphones, to connect to them seamlessly. The expansion of wireless ubiquity will result in increasing numbers of consumers depending on mobile networks creating a need for increasing economies of scale to deliver lower cost per-bit. According to a prediction of future combined consumer and advertiser spend on mobile media and associated data, which includes handset browsing, mobile applications, mobile games, mobile music, mobile TV, ringtones, wall papers and alerts, spend will rise from just under \$75 billion at the end of 2010 to \$138 billion by 2015, at a 13.17 CAGR, (MacQueen, 2010). Moreover, it is predicted (RNCOS Industry Research Solution, 2011) that the number of mobile TV subscribers worldwide will grow at a CAGR of around 43% during 2011-2014 to reach about 792.5 million by the end 2014.

In response to this remarkable development, core and metro networks have experienced a tremendous growth in bandwidth and capacity with the widespread deployment of fibre-optic technology over the past decade, (OASE, 2010). Fiber optic transmission has become one of the most exciting and rapidly changing fields in telecommunication engineering. Fiber optic communication systems have many advantages over more conventional transmission systems. They are less affected by noise, are completely unaffected by electromagnetic interference (EMI) and radio frequency interference (RFI), do not conduct electricity and therefore, provide electrical isolation, are completely unaffected by lightning and high voltage switching, and carry extremely high data transmission rates over very long distances, (Guo et al., 2007). As shown in Fig. 1, data speeds in metro and long-haul systems are evolving from 10 Gbps to 40 Gbps transmission. A 100 Gbps per wavelength channel system is taking shape as a next step for core and metro networks, (FP7, 2010). Wavelength division multiplexing (WDM) techniques, such as: dense WDM (DWDM), and highly DWDM (HDWDM) offer the potential for huge bandwidth fiber optic networks with all-optical switching and routing in the future.

In the recent years wireless services have been taking a steadily increasing share of the telecommunications market. End users not only benefit from their main virtue, mobility, but are also demanding ever larger bandwidth. Larger wireless capacity per user requires the reduction of the wireless cell size, i.e. establishing pico-cells. These can be realised using Wi-Fi systems based on the wireless Local Area Network (LAN) IEEE 802.11n standard which offers data rates of up to 600 Mbit/s. Furthermore, the Wi-Fi Alliance and the Wireless Gigabit Alliance (WiGig) announced that they will cooperate on multi-gigabit wireless schemes that are likely to bring robust wireless networking from the 60 GHz frequency band to consumers whose devices are equipped with Wi-Fi. The partnership will pave the way for new wireless devices that will operate in the 2.4, 5 and 60 GHz bands. It is anticipated that data transfer rates up to 7 Gbps can be achieved, although the highest data rates are likely to be available only over short distances within living room-sized areas. Nevertheless, the highest rates will be more than 10 times faster than 802.11n (Anthony, 2011). Furthermore, Worldwide interoperability for Microwave Access second generation (WiMAX 2), the marketing name for systems based on the IEEE 802.16m standard, is expected to expand capacity to 300 Mbps peak rates via advances in antennas, channel stacking and frequency re-use over the period 2012 to 2013, (Schwarz, 2011). Looking further ahead the recently ratified IEEE 802.15.3c standard has been defined for the frequency band of 57.0-66.0 GHz, allocated by regulatory agencies in Europe, Japan, Canada, and the United

States. According to this standard, single carrier mode in millimeter wave PHY supports a variety of modulation and coding schemes (MCSs) that support up to 5 Gb/s, (Guo and Kuo, 2007).

Super-broadband access not only provides faster web surfing and quicker file download, but also enables several multimedia applications such as real-time high definition audio and video streaming, multimedia conferencing, and interactive gaming. Broadband connections are currently being used for voice telephony using Voice-over-Internet-Protocol (VoIP) technology. More advanced broadband access systems, such as fiber to the home (FTTH) and very high data rate digital subscriber line (VDSL), enable applications such as entertainment-quality video, including HDTV, and Video on Demand (VoD) to be provided, but for SHDTV and UHDTV services a super-broadband network is essential. As the broadband market continues to grow, several new applications are likely to emerge and it is difficult to predict which ones will succeed in the future.

Broadband wireless is about bringing the broadband experience to a wireless context, which offers users certain unique benefits and convenience. There are two fundamentally different types of broadband wireless services. The first attempts to provide a set of services similar to that of the traditional fixed-line broadband but using wireless as the medium of transmission. This type, called fixed wireless broadband, can be thought of as a competitive alternative to DSL or cable modem. The second type of broadband wireless, called mobile broadband, offers the additional functionality of connectivity in mobility. Mobile broadband attempts to bring broadband applications to new user experience scenarios and hence can offer the end-user a very different value proposition.

Long Term Evolution (LTE) is a new radio platform technology that will allow operators to achieve even higher peak throughputs than High Speed Packet Access evolution (HSPA+) in higher spectrum bandwidth. Furthermore, the overall objective for LTE is to provide an extremely high performance radio-access technology that offers full vehicular mobility and can readily coexist with HSPA and earlier networks. Because of scalable bandwidth, operators will be able to migrate their networks and users from HSPA to LTE easily over time. LTE assumes a full IP network architecture, (Rysavy Research, 2007).

Fig. 5 shows the evolution of the 3GPP family of standards towards LTE Advanced (Chang et al., 2007; Rodrigo et al., 2009). LTE uses OFDMA (Orthogonal Frequency Division Multiplexing Access) on the downlink and FDMA (Frequency Division Multiple Access) on the uplink for better power performance of the end-user's handset, which is well suited to achieving high peak data rates in high spectrum bandwidth, achieving peak rates in the 1 Gbps range with wider radio channels. However, wider channels would result in highly complex terminals and is not simply achievable with the conventional communication infrastructure. Moreover, access bandwidth requirements for delivering multi-channel HDTV, SHDTV, and UHDTV signals and online gaming services are expected to grow beyond several Gbps in the near future and the current subscriber access networks have not been scaled up commensurately. To avoid being the bottleneck in the last miles and last meters, and exploit the benefits of both wired and wireless technologies, mobile and wireless communication service providers and operators are actively seeking convergent network architecture to deliver multiple super-broadband services to serve both fixed and mobile users, (Nokia, 2009; PIANO+, 2010).

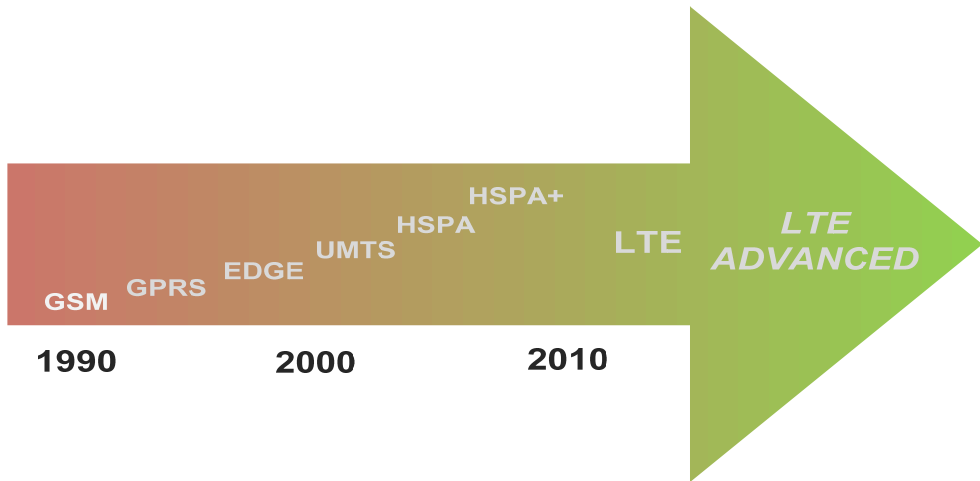


Fig. 5. Evolution of the 3GPP family of standards.

In this regard, optical-wireless access technologies have been considered the most promising solution to increase the capacity, coverage, bandwidth, and mobility in environments such as conference centers, airports, hotels, shopping malls, and ultimately to homes and small offices. As a result, research activity in the field of optical networks and converged optical and wireless communication technologies has grown rapidly and steadily over the last several years. This is because optical communication is a promising choice to fulfill the ever-increasing demand on bandwidth via the vast available capacity of optical fiber and its economic cost. Wireless communication technology on the other hand can provide mobility during communication periods and it is entering a new phase where the focus is shifting from voice to high definition multimedia services. Present consumers are no longer interested in the underlying technology; they simply need reliable and cost effective communication systems that can support end-users' demanded services anytime, anywhere, any media, that they want.

3.1 Core and metro networks

The two main categories of network to be considered from the point of view of establishing super-broadband access networks are core and metro networks. In this subsection, the two main challenges facing core and metro networks are discussed. These challenges are realising the bandwidth potential of fiber optic core networks by appropriate wavelength allocation and switching strategies. Therefore in this subsection, the discussion focussed on optical switching paradigms and dynamic wavelength allocation.

The main barrier to the use of most existing core and metro networks for future traffic transportation arises from their active electrical switching and routing systems which delay packets when processing them for switching in the electrical domain. It takes time to convert a signal from the optical to the electrical domain and vice versa. In addition the synchronization and data retiming processing takes time. Indeed, a great part of the

research into optical networks is dedicated to transparency in optical networks in order to bypass Optical/Electrical/Optical (O/E/O) conversions in the intermediate nodes of the network. Thus, a number of network protocols such as MPLS (Multi Protocol Label Switching, GMPLS, etc. (Larkin, 2005) together with switching strategies (circuit- burst- or packet-switching) are proposed for data transparency in the network. Among the switching strategies, burst switching is the most compatible with the current optoelectronic technologies in terms of data transparency and switching speed. Packet switching is more efficient for data communication, but due to the limited speed of electrical networks compared to the current optical networks and the insufficient evolution of all-optical signal processing alternatives, packet based optical networks are not a practical solution for transparent optical networks. A comparison of the all optical switching schemes, optical circuit switching (OCS), optical burst switching (OBS) and optical packet switching (OPS) is shown in Fig. 6.

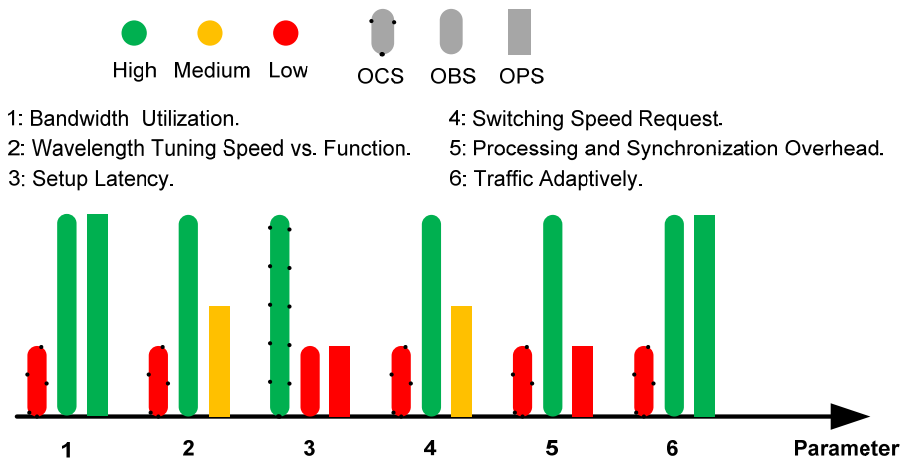


Fig. 6. Comparison of all-optical switching technologies in terms of relative magnitudes of performance measures.

Optical packet switching (OPS) is a viable candidate switching scheme for future networks because it is a purely-connectionless networking solution that is fully compatible with IP-centric data traffic and offers the finest network granularity, the best bandwidth utilization, flexibility, high-speed, and the ability to use the resources available economically.

OPS places more demanding prerequisites on the network than OBS because it processes packets on the fly. The most feasible approach to implementing of OPS involves processing synchronously transmitted packets with fixed lengths. However, in this case the hardware overhead is on the implementation of the packet synchronizer at the input to the switch. Despite their feasibility limitations, OPS demonstrators assisted the development of numerous ultra-fast switching and processing techniques regarding wavelength conversion, header encoding/decoding and processing, label swapping, fast clock extraction, and regeneration.

The main challenges in OPS are the implementation of the optical header processing mechanism, the development of an intelligent switch controller, the realization of ultra fast switching at a nanosecond timescale, and the exploitation on buffering mechanisms to reduce packet blocking (Rodrigo et al, 2009; Raffaelli et al. 2008; Le Rouzic et al., 2005).

Furthermore, the channel allocation and spectral efficiency are other key points for super-broadband network deployment. There are different schemes for channel allocation and multiplexing techniques such as wavelength division multiplexing (WDM), Dense-WDM(DWDM), Highly DWDM, Orthogonal WDM (Goldfarb et al., 2007; Llorente et al., 2005) that are suitable for super-broadband network deployment. The WDM multiplexing based schemes are in addition to multiplexing schemes including time, frequency, and code division multiplexing techniques, which are used in current wired and wireless communication networks and perform well on them. Moreover, cognitive channel and spectrum allocation improves the network's throughput and reduces the cost-over head significantly.

3.2 Access network

Ultra-fast and super-broadband are recognized as becoming increasingly important as demands for bandwidth multiply. Investment in the development of next-generation optical-wireless converged access technologies will enable a future network to be deployed that will radically reduce Fiber-Wireless (FiWi) infrastructure costs by removing local exchanges and potentially much of the metro network. To integrate fiber and wireless technologies, there are important challenges. First, it will be crucial to have mechanism in place to control system load, which will translate into the physical characteristics of the different radio access technologies of wireless systems, the variability of users' requirements and the data rate of on-going wireless connections, complicating the resource management/sharing in FiWi access networks. This raises technical issues such the required protocol interfaces between the resource management entities of tightly coupled networks, and calls for the design of very flexible and effective protocols to allow enhanced routing and link adaptation that makes the best usage of the available resources while dynamically accommodating the users' traffic properties and quality of services requirements.

3.2.1 Passive optical network

There are different topologies for deploying the fiber network from a central exchange station to end-user's premises such as: 1) point-to-point (P-to-P): where individual fibers run from the central station to end-users, 2) point-to-multi-point (P-to-MP) active star architecture: where a single feeder fiber carries all traffic to an remote active node close to the end-users, and from there individual short branching fibers run to the end-users. In this architecture, the fiber network implementation cost is less than that of a point-to-point topology but the main disadvantages of this architecture are a) the bandwidth of the feeder fiber is shared between several end-users and the allocated dedicated bandwidth for each end-user is less than in the point-to-point architecture. b) the requirement for active equipment in a remote node will impose some restrictions on network deployment such as the availability of a reliable and uninterruptable power supply, proper space for installation of active equipment, air conditioning and ventilation, and maintenance costs, 3) point-to-

multi-point passive star architecture: in which the active node of the active star topology is replaced by a passive optical power splitter/combiner that feeds the individual short range fibers to end-users. This topology has become a very popular and is known as the passive optical network (PON). In this topology, in addition to the reduction in installation cost, the active equipment is completely replaced by passive equipment avoiding the powering and related maintenance costs, (Koonen, 2006).

Besides the technical issues of implementation, the maintenance and operation cost overhead should be accounted for as it plays a key role in choosing a particular architecture. In the P-to-P architecture, for each end-user, two dedicated optical line terminations (OLT) are needed, while, in the P-to-MP scheme, for each end-user one dedicated OLT is required at the end-user side, another shared OLT at the central station is interfaced between several end-users. When the number of customers increases, the system costs of the P-to-P architecture grow faster than those of the P-to-MP architecture, as more fibers and more line terminating modules are needed. Therefore, sharing the implemented infrastructures between several operators, service providers, technologies, and end users is an essential solution to reduce the infrastructure network cost overhead. . . As shown in Fig. 7, the initial cost of P-to-P topology ($Cost_{P-to-P}(N_1)$) for N_1 users is lower than initial cost of P-to-MP topology ($Cost_{P-to-MP}(N_1)$), while by increasing the duct length at point L_0 , the $Cost_{P-to-P}(N_1)$ crosses the $Cost_{P-to-MP}(N_1)$ graph and will be greater than it for fibre lengths greater than L_0 . Furthermore, the initial cost of P-to-MP topology for N_2 users, where $N_2 > N_1$, is more cost effective than the initial cost of P-to-P topology for N_2 users.

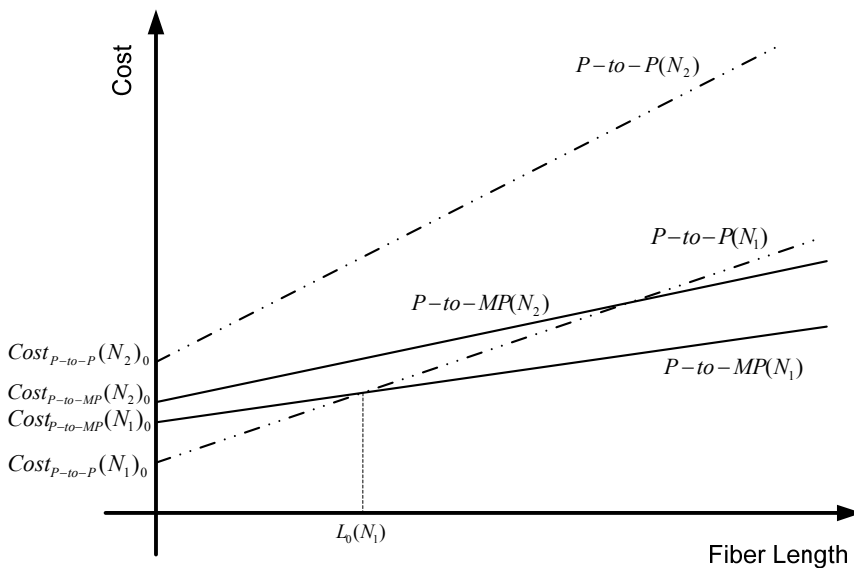


Fig. 7. The comparison of systems cost of FTTH different topology networks versus duct length to end-users premises.

In the P-to-P and P-to-MP active star architectures, each fiber link only carries a data stream between two electro-optic converters, and the traffic streams of the end-users are multiplexed electrically at these terminals. Therefore, there is no risk of collision of optical data streams. Whereas, the traffic multiplexing is done optically in a Passive Optical Network (PON) topology by integration of the data streams at the passive optical power combiner; to avoid collisions between individual data streams it is necessary to implement a well-designed multiplexing technique. A model of WDM PON network is shown in Fig. 8.

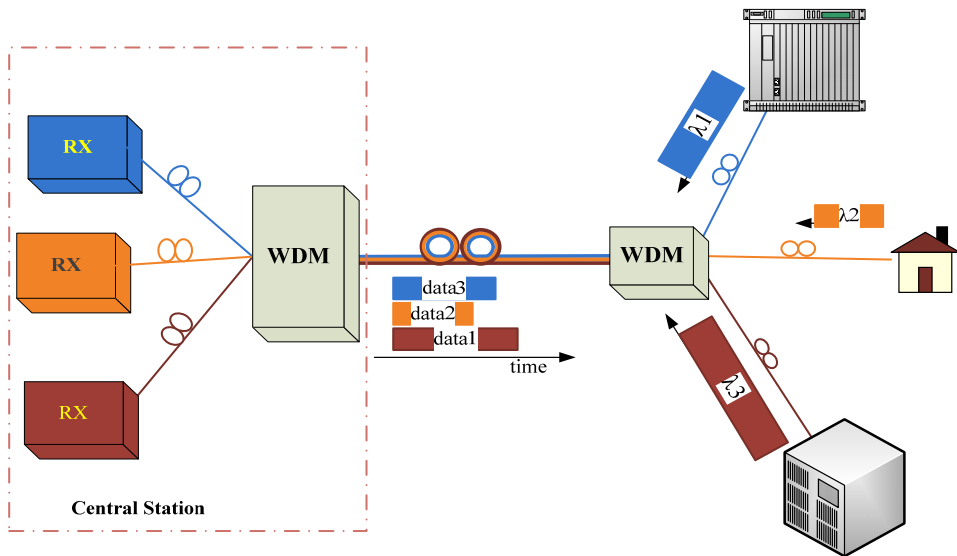


Fig. 8. A model of a point-to-multi-point passive optical network topology.

Several multiplexing techniques are used in PON networks, such as time division multiple access (TDMA), subcarrier multiple access (SCMA), wavelength division multiple access (WDMA), and optical code division multiple access (OCDMA). Excluding the wavelength division multiplexing technique, these multiplexing techniques are available in wireless or wired telecommunication systems. As shown in Fig. 9, in a WDM PON, each optical network unit (ONU) uses a different wavelength channel to send its packets to an OLT in a central office. The wavelength channels can be routed from the OLT to the appropriate ONUs and vice versa by a wavelength demultiplexing/multiplexing device located at the PON splitting point. This wavelength multiplexing technique constitutes independent communication channels and the network could be able to transport different signal formats; even if the channels use different multiplexing techniques no time synchronization between the channels is needed.

Currently Fiber to the home (FTTH) access technologies provide huge bandwidth to users, but are not flexible enough to allow roaming connections. On the other hand, wireless networks offer mobility to users, but do not possess sufficient bandwidth to meet the ultimate demand for multi-channel video services with high definition quality. Therefore, seamless integration of wired and wireless services for future-proof access networks will

lead to a convergence to high bandwidth provision for both fixed and mobile users in a single, low-cost transport platform. This can be accomplished by using the developed hybrid optical and wireless networks, which not only can transmit signals received wirelessly over fiber at the BS, but also simultaneously provide services received over fiber to wireless the end users.

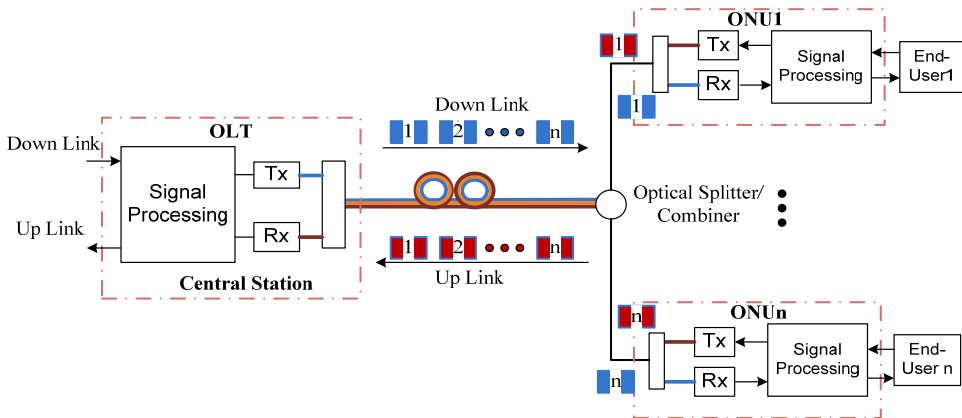


Fig. 9. WDM over a passive optical network.

3.2.2 Dynamic wavelength allocation

By creating multiple wavelengths in a common fiber infrastructure, the capabilities of this infrastructure can be extended into an additional dimension. This wavelength dimension can implement independent communication planes between nodes. For example, interconnections in this plane can be asynchronous, have different quality-of-service requirements, and can transport signals with widely differing characteristics. By using the WDM technique, the access network can: 1) separate services; 2) separate service providers; 3) enable traffic routing; 4) provide higher capacity; 5) improve scalability. For assignment of wavelengths to channels the system may follow different scenarios such as: a) static allocation; b) semi-static allocation; c) dynamic allocation, (Urban et al., 2009).

The static wavelength multiplexing scheme sets a virtual P-to-P topology up between two nodes of the network. However, the rapid growth in access network traffic requires flexible and adaptive planning of the wavelength allocation to each different channel or wired and wireless service to avoid congestion resulting from variable data rates demanded or to guaranteed data traffic transportation or services to/from the end-users. By using adaptive wavelength allocation deliverable services will be more cost-effective on the same network and the vast potential bandwidth of fiber optical networks will be more fully exploited. By assigning the wavelength dynamically at the Optical Network Unit (ONU), with flexible wavelength routing, the access network capabilities can be considerably enhanced. This configuration allows setting up a new wavelength channel before breaking down the old one. Alternatively, it may use wavelength tuneable transmitters and receivers, which can in

principle, address any wavelength in a certain range. The network management and control system commands to which downstream and to which upstream wavelength channel each ONU transceiver is switched. By issuing these commands from a central station, the network operator actually controls the virtual topology of the network, and thus is able to allocate the networks resources in response to the traffic at the various ONU sites. By changing the wavelength selection at the ONUs, the network operator can adjust the system's capacity allocation in order to meet the local traffic demands at the ONU sites.

In this scenario, as soon as the traffic to be sent upstream by an ONU grows and does not fit anymore within its wavelength channel, the network management system can command the ONU to be allocated another wavelength channel, in which sufficient free capacity is available. Obviously, this dynamic wavelength reallocation process reduces the system's blocking probability, i.e. it allows the system to handle more traffic without blocking and thus it can increase the revenue of the operator from a given pool of communication resources at the central station.

4. Radio over fiber network

The deployment of optical and wireless access network infrastructure is starting to proliferate throughout the world. When these heterogeneous access networks converge to a highly integrated network via a common optical feeder network, network operators can reap the benefits of lowering the operating costs of their access networks and meeting the capital costs of future upgrades more easily. In addition, the converged access network will facilitate greater sharing of common network infrastructure between multiple network operators. Signals received wirelessly and transported over optical fiber (RoF) links will be a possible technology for simplifying the architecture of remote base stations (BSs). By relocating key functions of a conventional BS to a central location, BSs could be simplified into remote antenna units that could be inter-connected with the central office (CO) via a high performance optical fiber feeder network.

Wireless networks typically show considerable dynamics in the traffic loads of their radio access points (RAPs) due to the fluctuations in the number and nature of mobile and wireless services demanded by the networks users. Using the traditional RAPs approach this requires all the wireless nodes to be equipped to cater for the highest capacity likely to be demanded of them which results in the inefficient use of network resources. The design of dynamic reconfigurable micro/pico or femo wireless cells increases network complexity but can significantly increase network efficiency. Similarly, within the optical access network layer WDM PONs allow an extra level of reconfiguration as wavelengths can be assigned either by static or dynamic routing.

The numbers of wireless subscribers are increasing and these subscribers are demanding more capacity for ultra-high data rate transfer at speeds of 1Gbps and up while the radio spectrum is limited. This requirement of more bandwidth allocation places a heavy burden on the current operating radio spectrum and causes spectral congestion at lower microwave frequencies. Millimetre Wave (mm-Wave) communication systems offer a unique way to resolve these problems (Ji, et al. 2009). Radio over fiber (RoF) technology is currently receiving a lot of attention due to its ability to provide simple antenna front ends, increased capacity, and multi wireless access coverage.

An analog RoF (ARoF) also known as RoF is the technique of modulating a radio frequency (RF) sub-carrier onto an optical carrier for distribution over a fiber network. An ARoF link includes optical source, modulator, optical amplifier and filters, optical channel and a photodiode as a receiver, electronic amplifiers and filters; a simple ARoF architecture is shown in Fig. 10. In this system, for a downlink at a central station, a signal received wirelessly is modulated onto an optical carrier generated by a laser diode (LD) and the modulated optical signal is transported over a fiber optic cable. The transported optical signal is detected at base station using a photo diode (PD). The received signal, recovered after performing analog signal processing, is fed to an antenna for wireless transmission. For uplink signal transmission from a base station to the central station, the signal received at an antenna is directed to a low noise amplifier (LNA) and modulated onto an optical carrier that is generated by another LD. The generated optical signal is sent back to the central station for any signal processing and detection. In some cases the RF signal is directly modulated by optical source, but as the laser is usually a significant source of noise and distortion in a radio over fiber link, the laser diode normally exhibits nonlinear behavior. When the LD is driven well above its threshold current, its input/output relationship can be modeled by Volterra series of order 3. Therefore, in high data rate links indirect modulation has better performance. However, an ARoF link suffers from the nonlinearity of both microwave and optical components that constitute the optical link (Al-Raweshidy & Komaki, 2002; Cox, 2004; Li & Yu, 2003). Fig. 11, shows an ARoF link architecture with indirect intensity modulation that uses an electro-optical modulator for modulating an electrical signal representing the information in a wireless signal onto a continuous wave laser source.

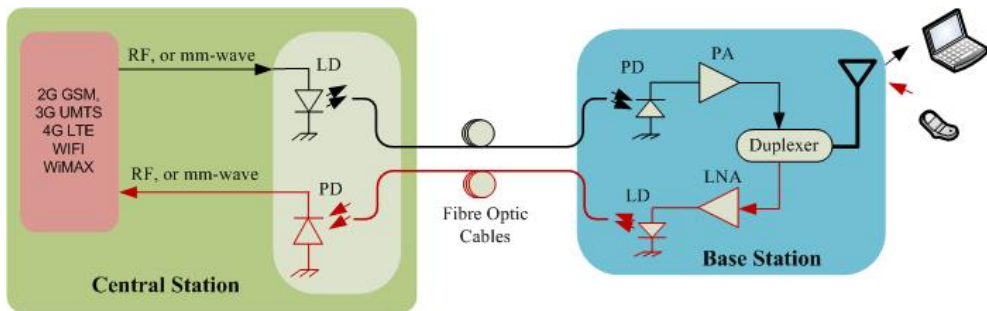


Fig. 10. A direct intensity modulation and detection full-duplex ARoF architecture.

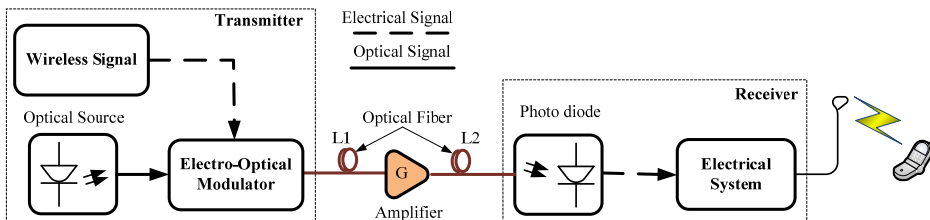


Fig. 11. Downlink architecture of a ARoF link with indirect intensity modulation.

The ARoF technique has been considered a cost-effective and reliable solution for the distribution of future services to wireless devices by using optical fiber with vast transmission bandwidth capacity. An ARoF link is used in remote antenna applications to distribute signals to a Microcell or Picocell base station (BS). The downlink RF signals are distributed from a central station (CS) to a BS known as a Radio Access Point (RAP) through optical fibers. The uplink signals received at RAPs are sent back to the CS for any signal processing. RoF has the following main features: (1) it is transparent to bandwidth or modulation techniques; (2) it only needs simple and small BSs; (3) centralized operation is possible; (4) it supports multiple wired and wireless standards, simultaneously. (5) its power consumption is relatively low. Furthermore, the implementation of the RoF technique faces the following challenges: fiber optic network implementation cost, optical communication components nonlinearity and fiber dispersion. Consequently, in last decade several research projects have sought to develop and discover new solutions to overcome these challenges and broaden the benefits of RoF.

4.1 Radio over fiber's link architecture

The signal that is transmitted over the optical fiber can either be originally an RF, intermediate frequency (IF) or baseband (BB) signal. For the IF and baseband (BB) transmission cases, additional hardware for up converting the signal to the RF band is required at the BS. At the optical transmitter, the RF/IF/BB signal can be modulated onto the optical carrier by using direct or external modulation of the laser light. In an ideal case, the output signal from the optical link will be a copy of the input signal. However, there are some limitations because of non-linearity and frequency response limits in the laser and modulation devices as well as dispersion in the fiber. The transmission of analog signals puts certain requirements on the linearity and dynamic range of the optical link. These demands are different and more exacting than requirements placed on digital transmission systems.

In Fig. 12, typical RoF link configurations are shown, which are classified based on the kinds of frequency bands transmitted over an optical fiber link. In the downlink from the CS to the BS, the information signal from a public switched telephone network (PSTN), an Internet Service Provider (ISP), a mobile telecommunications operator, an Intelligent Transportation System (ITSs) or another CS is fed into the optical network at the CS. The signal that is either RF, IF or BB band modulates an optical signal from a LD. As described earlier, if the RF band is low, it's possible to modulate the LD signal using the RF band signal directly. If the RF band is high, such as the mm-wave band, it's better to use electro-optical modulators (EOMs), like Mach-Zehnder Modulators. The modulated optical signal is transmitted to the BS via optical fibers. At the BS, the RF/IF/BB band signal is recovered by detecting the modulated optical signal by using a PD. The recovered signal, which needs to be upconverted to RF band if an IF or BB signal is to be transmitted to a mobile handset (MHs) via the antenna of the BS.

In the configuration shown in Fig. 12 (a), the modulated signal is generated at the CS in an RF band and directly transmitted to a BS by an EOM, which is called "RF-over-Fiber". At the BS, the modulated signal is recovered by detecting the modulated optical signal with a PD and directly transmitted to a MH. Signal distribution using RF-over-Fiber has the advantage of a simplified BS design but is susceptible to fiber chromatic dispersion that severely limits the transmission distance (Gliese et al., 1996). In the configuration shown in Fig. 12 (b), the

modulated signal is generated at the CS in an IF band and transmitted to a BS by an EOM, which is called "IF-over-Fiber". At each BS, the modulated optical signal is recovered by detecting the modulated optical signal with a PD, up converted to an RF band, and transmitted to a MH. In this scheme, the effect of fiber chromatic dispersion on the distribution of IF signals is much reduced. However, the antennas of the BSs a RoF system incorporating IF-over-Fiber transport require additional electronic hardware such as an mm-wave frequency LO for frequency up- and down conversion.

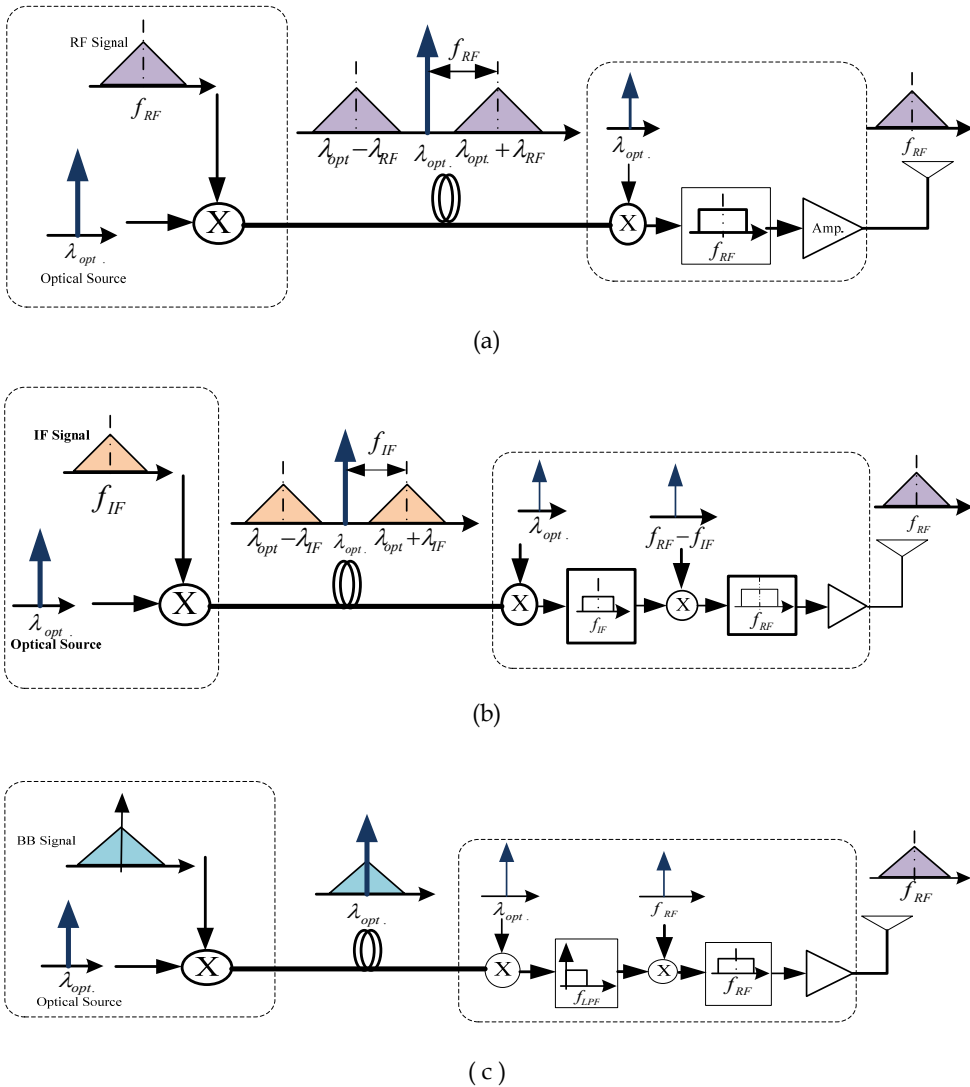


Fig. 12. Different schemes for signal modulation onto an optical carrier for distribution: (a) Radio over Fiber; (b) IF over Fiber; (c) BB over Fiber.

In Fig. 12 (c), the modulated signal is generated at the CS in baseband and transmitted to a BS, which is referred to as “BB-over-Fiber”. At the BS, the modulated signal is recovered by detecting the modulated optical signal with a PD, up converted to an RF band through an IF band or directly, and transmitted to a MH. In baseband transmission, the impact of fiber dispersion effects is negligible, but the BS configuration is the most complex. This is especially important when RoF in mm-wave bands is combined with dense wavelength division multiplexing (DWDM). This increases the amount of equipment at the BSs because an up converter for the downlink and a down converter for the uplink are required. In the RF subcarrier transmission, the BS configuration can be simplified only if an mm-wave optical external modulator and a high-frequency PD are implemented in the electric-to-optic (E/O) convertor and the optic-to-electric (O/E) converter, respectively.

Optical links are mainly transmitting microwave and mm-wave signals by applying an intensity modulation technique onto an optical carrier (Al-Raweshidy & Komaki, 2002). Fundamentally, two methods exist for transmission of the microwave/mm-wave signals over optical links with intensity modulation: (1) direct intensity modulation, (2) external modulation.

In direct intensity modulation an electrical parameter of the light source is modulated by the information RF signal. In practical links, this is the current of the laser diode, serving as the optical transmitter. In Fig. 10, the simplest and most cost-effective architecture of intensity-modulation direct-detection (IMDD) is depicted. In this architecture, the detection is performed using a photo diode (PD). In the direct-modulation process a semiconductor laser directly converts an electrical small-signal modulation (around a bias point set by a dc current) into a corresponding optical small-signal modulation of the intensity of the photons emitted (around the average intensity at the bias point). Thus, a single device serves as the optical source and the RF/optical modulator. An important limitation in this architecture for super broadband access are the restrictions placed on the modulation bandwidth by the laser and the mm-wave band while a simple laser’s linewidth can be modulated to frequencies of several Gigahertz. Furthermore, it is reported that direct intensity modulation lasers can operate at up to 40 GHz or even higher, but, these are expensive and are not cost-effective in the commercial market. Therefore, at frequencies above 10 GHz, external modulation rather than direct modulation is applied.

In the external modulation technique, Fig. 11, an unmodulated light source is modulated with an information RF signal using an electro-optical intensity modulator. Because the number of BSs is high in RoF networks, simple and cost-effective components must be utilized. Therefore, in the uplink of a RoF network system, it is convenient to use direct intensity modulation with cheap lasers; this may require down conversion of the uplink RF signal received at the BS. In the downlink either lasers or external modulators can be used.

4.2 Application of WDM in a radio over fiber system

The application of WDM in RoF networks has many advantages including simplification of the network topology by allocating different wavelengths to individual BSs, enabling easier network and service upgrades and providing simpler network management. Thus, WDM in combination with optical mm-wave transport has been widely studied (Griffin et al., 1999; Toda et al., 2003).

The implementation of WDM in a RoF network is illustrated in Fig. 13, where for simplicity, only downlink transmission is depicted. Optical mm-wave signals from multiple sources are multiplexed and the generated signal is optically amplified, transported over a single fiber and demultiplexed to address each BS concerned separately. A challenging issue is that the optical spectral width of a single optical mm-wave source may approach or exceed WDM channel spacing. Therefore, there have been several reports on dense WDM (DWDM) applied to RoF networks (Griffin et al., 1999; Griffin, 2000; Toda et al., 2003); by utilizing the large number of available wavelengths in the DWDM technique, the lack of free transmission channels for the deployment of more BSs in mm-wave bands can be overcome. Another issue is related to the number of wavelengths required per BS. It is desirable to use one wavelength to support full-duplex operation. In (Nirmalathas et al., 2001), a wavelength reuse technique has been proposed, which is based on recovering the optical carrier used in downstream signal transmission and reusing the same wavelength for upstream signal transmission.

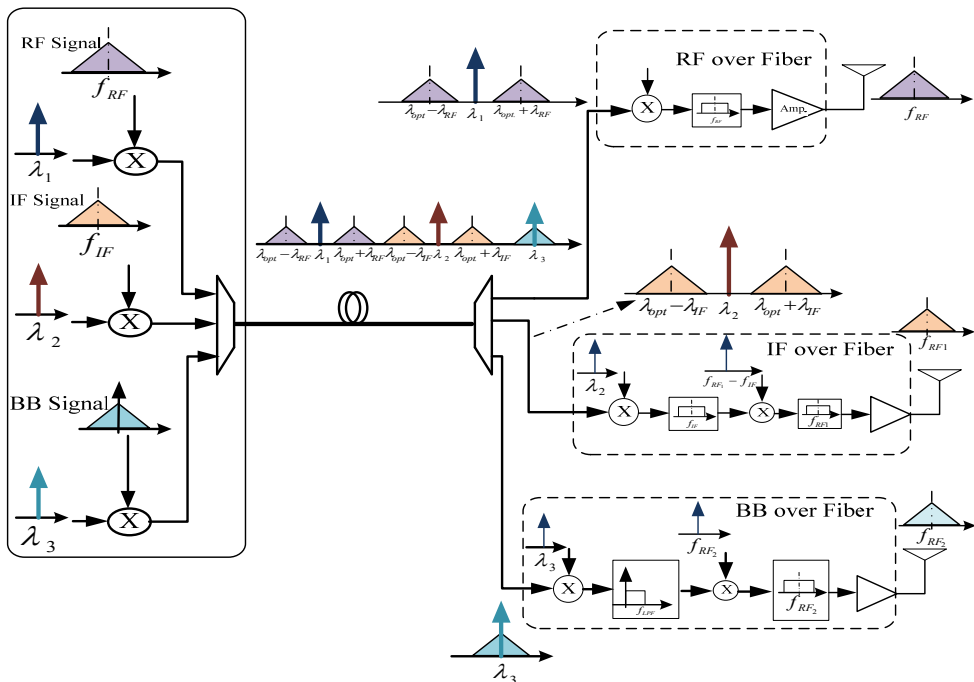


Fig. 13. Schematic illustration of the implementation of WDM in a RoF network.

4.3 Digital radio over fiber

Digital systems are more flexible, more conveniently interface with other systems, are more reliable and robust against additive noise from devices and channels, and achieve a better dynamic range than analog systems. Analog to digital and digital to analog converters (ADC and DAC, respectively) (Walden R. H., 1999) are the link between the analog world and the digital world of signal processing and data handling. In an analog system the bandwidth is limited by devices performance and parasitic components are introduced.

In a Digital RoF (DRoF) system, an electrical RF signal is digitized by using an Electronic ADC (EADC) (Vaughan et al., 1991). Then, the generated digital data is modulated with a continuous coherent optical carrier wave either using a direct modulation technique or by using an external electro-optical modulator as shown in Fig. 14. The modulated optical carrier is transmitted through the fiber. At the base station, after detecting the optical signal using a photo diode, the detected digital data is converted back to the analog domain using an EDAC. Finally, the analog electrical signal is fed to an antenna (Li et al., 2009; Kuwano, 2006, 2008; Lim et al., 2010). Current EDAC systems experience problems such as jitter in the sampling clock (Stephens, 2004; Hancock, 2004), the settling time of the sample and hold circuit, the speed of the comparator, mismatches in the transistor thresholds and passive component values. The limitations imposed by all of these factors become more severe at higher frequencies. Wideband analog to digital conversion is a critical problem encountered in broadband communication and radar systems (Valley, 2007; Kim et al., 2008). For the future beyond Gigabit/s mobile and wireless end-user traffic rates (Abdollahi et al., 2010) due to the limitations of electronic technology for implementing ultra high-speed, high performance EADC, and the resolution of existing EDAC, the deployment of conventional DRoF links (Li et al., 2009; Kuwano, 2006, 2008; Lim et al., 2010) is not simply achievable.

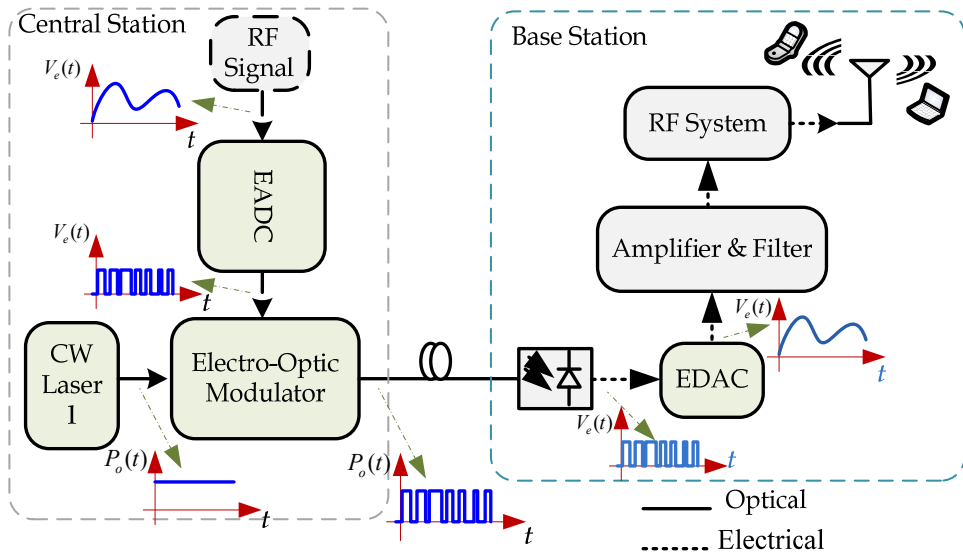


Fig. 14. Conventional DRoF architecture using EADC (downlink) (Li et al., 2009).

Moreover, if a conventional DRoF link could be achieved for Gigabit/s traffic rates the generated digital traffic creates a new challenge, namely, for this architecture to use more electro-optical modulators and photo diodes to implement the wavelength division multiplexing (WDM) technique to diminish the chromatic dispersion caused by the restrictions on the modulation bandwidth for super broadband access by RoF.

4.4 All-photonic digital radio over fiber

An all-photonic DRoF architecture has been proposed (Abdollahi et al. 2011) and is depicted in Fig. 15. This architecture uses an electro-optical modulator, which is simultaneously shared as an optical sampling and modulating device at the CS. A photonic ADC (PADC) by using a mode-locked laser (MLL) and an electro-optical modulator is able to scale the timing jitter of the laser sources to the femtosecond level, (Kim et al. 2007; Bartels et al. 2003), which allows designers to push the resolution bandwidth by many orders of magnitude beyond what electronic sampling systems can currently achieve. The proposed system includes an all-photonic signal processing block for optical quantization and wavelength conversion of the sampled and symmetrically split signal's power. By using the WDM technique to distribute the generated traffic over different wavelengths exceeding the modulation bandwidth of the fiber on a particular wavelength is prevented.

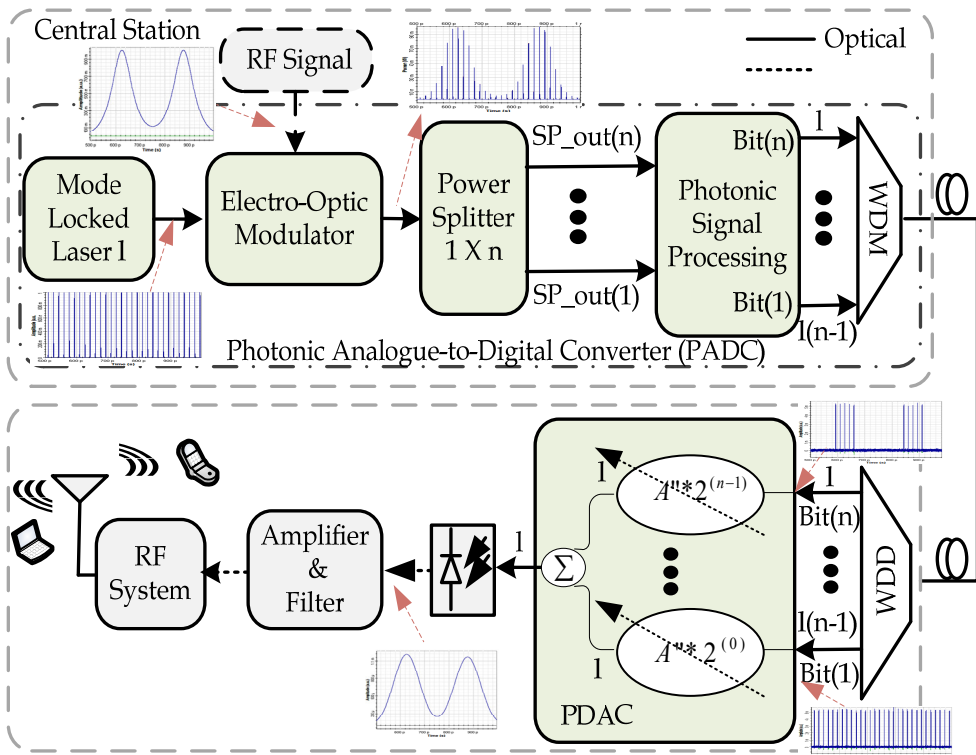


Fig. 15. All-photonic DRoF architecture, (downlink), (Abdollahi et al., 2011).

In Fig. 15, at the CS, the RF signal is sampled and modulated by optical train pulses that are generated using a passive mode-locked laser. The optical power of the sampled pulses is split into n levels using a symmetrical optical splitter, where n denotes the number of quantization bits. Finally, the split signals are fed to a photonic signal processing block for quantization and wavelength conversion operations.

The quantization procedure is performed by the process of Fig. 16 in which A and A' are constant parameters. At the first stage of this process, the stage number is equal to '1' (S=1). In this process entire stages are equal to number of quantization bits, i.e., for each output bit there is a corresponding quantization stage. For quantization of the most significant bit (MSB) the received signal from output number 'n' of the symmetrical splitter SP_out(M) that is defined by the generic number 'M' which is equal to 'n' in this stage. This output optical signal is compared with a reference quantization level equal to $2^{(M-S)} * A'$. If the signal power square is greater than or equal to $2^{(M-S)} * A'$, the output quantization bit is '1'. Otherwise, it is '0'. In this scheme, for performing the pipeline architecture, the quantized bits are converted back into analog domain. Therefore, in stage number '(M-S)', the converted back analog signals from stages 'n' to '(M-S+1)' of the process, are subtracted from the input of the split output signal SP_out(M-S). Then, the given signal is compared with $2^{(M-S)} * A'$. The quantization process is repeated in parallel 'n' times for quantizing each sampled optical signal into 'n' bits.

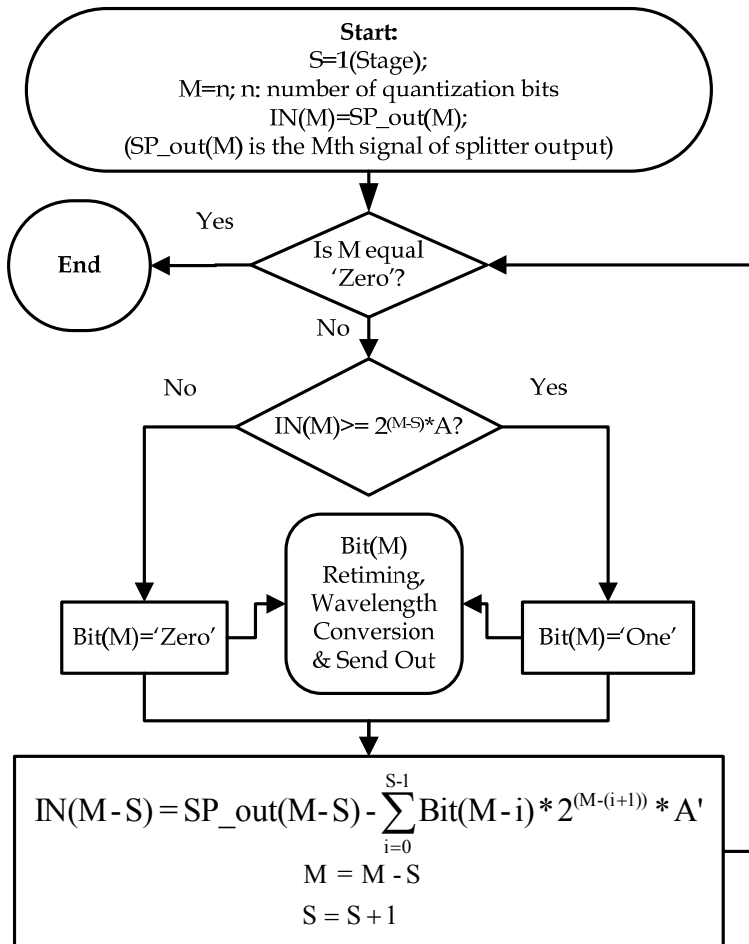


Fig. 16. All-photonic signal processing technique, (Abdollahi et al., 2011).

Subsequent to the wavelength conversion, the digital photonic signals are multiplexed in the wavelength domain by using a WDM and transmitted over a fiber. At the BS, the received signal is demultiplexed by wavelength division demultiplexer (WDD) and fed to the photonic digital-to-analog converter (PDAC). The PDAC subsystem, receives digital optical signals on different wavelengths, and converts them back to the equivalent analog signal at wavelength λ by using a passive PDAC and all-optical wavelength conversion. In the following of this stage, by using a photo diode (PD), the RF signal is recovered and after some RF signal processing it is passed to a multi-band distributed antenna system.

According to results provided (Abdollahi et al., 2011), it is demonstrated that ARoF is more dependent on fiber network impairments and length than DRoF. However, very low phase noise photonic sampling pulses and high speed signal conversion rates can be achieved in an all-photonic DRoF system compared with high-speed electronic circuits generated sampling pulses, signal conversion and processing. Consequently, an all-photonic DRoF system can support a digitized RF signal transmission system for providing super-broadband access to remote distributed wired and wireless access networks. It follows that, compared to the present digital optical communication infrastructures the number of CS would decrease with the introduction of all-photonic DRoF systems and as a result the service providers and network operators cost overheads per bit would be reduced.

5. Conclusions and chapter summery

Mobile and wireless networks generated traffic rates are growing very fast and are expected to double each year. The expectation is for delivering at least 1 Gbps multi-services traffic to each end-user in the near future for personal and multimedia communication services. Therefore, deploying super-broadband networks will be essential for service providers and operators. In this chapter, the convergence of wireless and optical communication technology for deploying future super broadband networks has been discussed.

Fiber optic transmission is rapidly becoming the dominant infrastructure medium for the transportation of fixed and mobile video on the internet. By replacing electronic switching with ultra fast photonic switching fiber optic transmission is expected to meet the need for super-broadband capacity. Radio-over-Fiber is a potential solution for deploying wireless access to broadband and super-broadband seamlessly. It can provide dynamic allocation of resources and can be realised with simple and small BSs with centralized operations. The requirement for more bandwidth allocation places a heavy burden on the current operating radio frequency (RF) spectrum and causes spectral congestion at lower microwave frequencies. Millimeter wave (mm-Wave) communication systems offer a unique way to resolve the bandwidth problems. When heterogeneous access networks converge to a highly integrated network via a common optical feeder network, network operators can reap the benefits of lowering the operating cost of access networks and meeting the capital costs of future upgrades easily. In addition, the converged access network will facilitate greater sharing of common network infrastructure between multiple network operators. Radio signals transportation over optical fiber (RoF) links will be a possible technology for simplifying the architecture of remote base stations (BSs). By relocating key functions of a conventional BS to a central location, BSs could be simplified into remote antenna units that could be inter-connected with a central office (CO) via high performance optical fiber feeder network. Wireless networks typically show considerable dynamics in traffic load of the

radio access points (RAPs), due to fluctuations in the number mobile and wireless service users using them and the services they demand.

On the other hand, using the traditional RAPs approach requires equipping all of the wireless nodes for the highest capacity demanded which results in the inefficient use of resources. The design of dynamic reconfigurable micro/pico or femo wireless cells increases network complexity but also greatly increases network efficiency. Within the optical access network layer WDM PONs allow an extra level of reconfiguration as wavelengths can be assigned to channels as part of static or dynamic routing. Therefore, integrating dynamic wavelength routing with RoF technology facilitates future flexible, low cost and reconfigurable super-broadband wired and wireless access network.

DRoF links are more independent of fiber network impairments and length than ARoF links. By using very low phase noise photonic sampling pulses and high speed signal conversion rates in place of high-speed electronic circuits generated sampling pulses, signal conversion and processing in an all-photonic DRoF system, digitized RF signal transmission for delivering future super-broadband remote distributed wired and wireless access networks traffic can be realised. Consequently, compared to the present digital optical communication infrastructure the number of CS will decrease in an all-photonic DRoF infrastructure and as a result, service providers and operators cost overhead per bit will be significantly reduced.

6. References

- Abdollahi, S. R.; Al-Raweshidy, H.S.; Nilavalan, R. (2011). Fully-Photonic Analogue-to-Digital Conversion Technique for Super-Broadband Digitized-Radio over Fibre Link, Proceedings of 16th European Conference on Networks and Optical Communication, (July 2011, pp. 72-75), UK.
- Abdollahi, S. R.; Al-Raweshidy, H.S.; Nilavalan, R.; Darzi A. (2010). Future Broadband Access Network Challenges, IEEE WOCN, (Sep. 2010), pp. 1-5.
- Al-Raweshidy, H.; Komaki, S. (2002). Radio over Fiber Technology for Mobile Communication Networks, Artech House, 685 Canton Street, MA 02062, (2002), pp. 136-138.
- Anthony, S., (2011). WiGig: 7Gbps Unified Data/Audiovisual Wi-Fi Coming in 2012, Available from <http://www.extremetech.com/computing/89904-wigig-7gbps-data-display-and-audio-mid-range-networking-coming-in-2012>, (July 2011).
- Bartels, A.; Diddams, S. A.; Ramond, T. M.; Holberg, L. (2003). Mode-locked Laser Pulse Trains with Subfemtosecond Timing Jitter Synchronized to an Optical Reference Oscillator, Optic Letters, Vol.28, (2003), pp. 663-665.
- Chang, G.-K.; Chowdhury, A.; Yu, J.; Jia, Z.; Younce, R. (2007). Next generation 100Gbit/s Ethernet Technologies," APOC 2007, Invited Paper, (November 2007), Wuhan China.
- Cisco Visual Networking Index (2011). Global Mobile Data Traffic Forecast Update, (Sep 2011).
- FP7 (2010). A Converged Copper-Optical-Radio OFDMA-Based Access Network with High Capacity and Flexibility, ICT Objective 1.1 The Network of the Future, (Jan. 2010).
- Cox, C. H. (2004). Analog Optical Links, Cambridge University Press, (2004), Cambridge UK.

- Gliese, U; . Norskow, S.; Nielsen, . T. N. (1996). Chromatic Dispersion in Fiber-Optic Microwave and Millimeter-Wave Links, IEEE Transaction of Microwave Theory Technology, Vol. 44, No. 10, (Oct. 1996), pp. 1716-1724.
- Goldfarb, G.; Li, G.; Taylor, M. G. (2007). Orthogonal Wavelength-Division Multiplexing Using Coherent Detection, IEEE Photonics Technology Letters, Vol. 19, No. 24, (Dec., 2007), pp. 2015-2017.
- Griffin, R. A. (2000). DWDM Aspects of Radio-over-Fiber, Proceedings of LEOS 2000 Annual Meeting, Vol.1, (Nov. 2000), pp. 76-77.
- Griffin, R. A.; Lane, P. M.; O'Reilly, J. J.; (1999). Radio-Over-Fiber Distribution Using an Optical Millimeter-Wave/DWDM Overlay, Proceeding of OFC/IOOC 99, Vol.2, (Feb. 1999), pp. 70-72.
- Guo,Y.F.; Kuo, G.S. A novel QoS-guaranteed power-efficient management scheme for IEEE 802.15.3 HR-WPAN, 2007 4th Annual IEEE Consumer Communications and Networking Conference CCNC 2007 (2007) Pages: 634-638
- Hancock, J. (2004), Jitter-Understanding it, Measuring it, Eliminating it, Part 1: Jitter Fundamentals, High Frequency Electronics, Summit Technical Media, LLC, (April, 2004) pp. 44-50.
- ITU (2009). The World in, ICT Facts and Figs, Available from <http://www.itu.int>, (2009).
- Ji, H. C.; Kim, H.; Chung, Y. C. (2009). Full-Duplex Radio-Over-Fiber System Using Phase-Modulated Downlink and Intensity-Modulated Uplink, IEEE Photonics Technology Letters, Vol.21, No.1, (Jan. 2009), pp 9-11.
- Kim, J.; Chen, J.; Cox, J.; Kartnr, F. X. (2007). Attosecond-Resolution Timing Jitter Characterization of Free-Running Mode-Locked lasers, Optics Letters, Vol.32, (2007), pp. 3519-3521.
- Kim, J.; Park, M. J.; Perrott, M. H.; Kartner, F. (2008). Photonic Sub-Sampling Analog-to-Digital Conversion of Microwave Signals at 40-GHz with Higher than 7-ENOB Resolution, Optics Express, Vol.16, No.21, (2008), pp. 16509-16515.
- Koonen, T. (2006), Fiber to the Home/Fiber to the Premises: What, Where, and When? Proceedings of the IEEE , Vol. 94, No. 5, (May 2006), pp. 911-934.
- Kudo, K. (2005). Introduction of High-Definition TV System in NHK News Center, ABU Technical Committee 2005 Annual Meeting, (20-24 November 2005), Hanoi.
- Kuwano, S.; Suzuki, Y.; Yamada, Y.; Fujinio, Y.; Fujiti, T.; Uchida, D.; Watanabe, K. (2008). Diversity Techniques Employing Digitized Radio over Fiber Technology for Wide-Area Ubiquitous Network, IEEE Global Telecommunication Conference, (Globecom) , (Dec. 2008), pp. 1-5.
- Kuwano, S.; Suzuki, Y.; Yamada, Y.; Watanabe, K. (2006). Digitized Radio-over-Fiber (DRoF) System for Wide-Area Ubiquitous Wireless Network", IEEE International Topical Meeting on Microwave Photonics, (2006), pp. 1-4.
- Larkin, N. ASON & GMPLS; The Battle for the Optical Control Plane; Available from <http://www.dataconnection.com/network/download/whitepapers/asongmpls.pdf>.
- Laskar, J.; Pinel, S.; Dawn, D.; Sarkar, S.; Perumana, B.; Sen, P.; (2007). The Next Wireless Wave is a Millimeter Wave. Microwave Journal, Vol.90, No. 8, (August 2007), pp 22-35.

- Le Rouzic, E.; Gosselin, S. (2005). 160 Gb/s Optical Networking: A Prospective Techno-Economic Analysis. *Journal of Lightwave Technology*, Vol.23, No.10, (Oct. 2005), pp 3024-3033.
- Li, G. L.; Yu, P. K. L.; (2003). Optical Intensity Modulators for Digital and Analog Applications, *Journal of Lightwave Technology*. Vol. 21, (2003), pp. 2010-2030.
- Li, T.; Crisp, M.; Penty, R. V.; White, I. H. (2009). Low Bit Rate Digital Radio over Fiber system, *IEEE International Topical Meeting on Microwave Photonic*, (2009), pp. 1-4.
- Lim, C.; Nirmalathas, A.; Bakaul, M.; Gamage, P.; Lee, K. L.; Yang, Y.; Novak, D.; Waterhouse, R. (2010). Fiber-Wireless Networks and Subsystem Technologies, *Journal of Lightwave Technology*, Vol. 28, No. 4, (2010) , pp. 390-405.
- Llorente, R.; Lee, J. H.; Clavero, R. (2005). Orthogonal Wavelength-Division-Multiplexing Technique Feasibility Evaluation, *Journal of Lightwave Technology*, Vol.23, No.3, (March 2005), pp. 1145-1151.
- MacQueen, D., (2010). Global Mobile Forecast 2001-2015, *Mobile Media Strategies*, (March 2010), pp. 10.
- Mcdonough, J. (2007). Moving Standards to 100 Gbps and Beyond, *IEEE Communication Magazine*, Vol. 45, No.11, (Nov. 2007), pp. 6-9.
- Nirmalathas, A.; Novak, D.; Lim, C.; Waterhouse, R. B. (2001). Wavelength Reuse in the WDM Optical Interface of a Millimeter-Wave Fiber-Wireless Antenna Base Station, *IEEE Transaction of Microwave Theory Technology*, Vol.49, No.10, (Oct. 2001), pp. 2006-2012.
- Nokia. (2009). LTE-Delivering the Optimal Upgrade Path for 3G Networks, Available from www.nokia.com/NOKIA.../Nokia.../LTE_Press_Backgrounder.pdf.
- OASE (2010). Optical Access Seamless Evolution. Available from http://cordis.europa.eu/fetch?CALLER=PROJ_ICT&ACTION=D&CAT=PROJ&RCN=93075, (Feb. 2010).
- OMEGA ICT Project. (2011). Gigabit Home Networks, Seven Framework Programme, Available from <http://www.ict-omega.eu>, (2008-2011).
- PIANO+ (2010). Photonic-based Internet Access Networks of the Future. Available from <http://www.trdf.co.il/eng/kolkoreinfo.php?id=448>, (Feb., 2010).
- Raffaelli, C.; Vlachos, K.; Andriolli, N.; Apostolopolus, D. (2008). Photonic in Switching: Architectures, Systems and enabling technologies. *Computer Networks Volume 52, Issue 10*, (July 2008), pp. 1873-1890.
- Rodrigo, M. de V.; Latouche, G., Remiche, M. -A. (2009). Modeling Bufferless Packet-Switching Networks with Packet Dependencies. *Computer Networks 53* (Feb. 2009), pp. 1450-1466.
- RNCOS Industry Research Solution, (2011). Global Mobile TV Forecasting to 2013, (Aug. 2011).
- Rysavy Research, (2007). Edge, HSPA and LTE the Mobile Broadband Advantage, *3G Americas*, (Sep., 2007).
- Schwarz, Y. (2011). WiMAX 2: the future Super Broadband 4G Network, Available from <http://www.goingwimax.com/> (July 5, 2011).
- Stephens, R. (2004). Analyzing Jitter at High Data Rates, *IEEE Optical Communication*, Feb. 2004, pp. 6-10.
- Stephens R. Jitter Analysis: The Dual-Dirac Model, RJ/DJ, and Q-sacle, *Agilent Technical Note*, Dec. 2004.

- Sugawara, M.; Masaoka, K.; Emoto, M.; Matsuo, Y.; Nojiri, Y. (2007). Research on Human Factors in Ultra-high-definition Television to Determine its Specifications, SMPTE Technical Conference, (October 2007).
- Toda, H.; Yamashita, T.; Kuri, T.; Kitayama, K. (2003). Demultiplexing Using an Arrayed-Waveguide Grating for Frequency-Interleaved DWDM Millimeter-Wave Radio-on-Fiber Systems, *Journal of Lightwave Technology*, Vol. 21, No. 8, (Aug. 2003), pp. 1735-1741.
- Urban, P. J.; Huiszoon, B.; Roy, R.; de Laat, M. M.; Huijskens, F. M.; Klein, E. J.; Khoe, G. D.; Koonen, A. M. J.; Waardt, H. D. (2009), High-Bit-Rate Dynamically Reconfigurable WDM-TDM Access Network, *Journal of Optical Communication Network* Vol.1, No. 2, (July 2009), pp. 143-159.
- Valley, G. C. (2007). Photonic analog-to-digital converters, *Optics Express*. Vol.15, No.5, (2007), pp. 1955-1982.
- Vaughan, R. G.; Scott, N. L.; White, D. R., (1991). The Theory of Bandpass Sampling, *IEEE Transaction on Signal Processing*, Vol.39, No.9, (Sep. 1991), pp.1973-1984.
- Walden R. H. (1999). Analog-to-Digital Converter Survey and Analysis, *IEEE Journal of Selected Areas in Communications*, Vol.17, No.4, (1999), pp. 539-550.
- Yuen, R.; Fernando, X. N.; Krishnan, S. (2004). Radio Over Multimode Fiber for Wireless Access, *IEEE Canadian Conference on Electrical and Computer Engineering*, Vol.3, (May 2004) pp. 1715-1718.

Part 6

Biological Effects of Wireless Communication Technologies

Evaluations of International Expert Group Reports on the Biological Effects of Radiofrequency Fields

Luc Verschaeve

*Scientific Institute of Public Health, O.D. Public Health and Surveillance,
Brussels and University of Antwerp, Department of Biomedical Sciences
Belgium*

1. Introduction

Electromagnetic fields, in particular so-called radiofrequencies are used by mobile or wireless communication systems as for example GSM mobile telephones, DECT telephones, wifi etc. Recent years were characterized by a tremendous increase in applications and types of wireless communication systems and this is responsible for an important increase in human exposure to radiofrequency radiation. Discussions on alleged adverse health effects are going on for years and so far no consensus agreement has been reached. These discussions are held amongst scientists as well as amongst laymen from the general public and authorities. Radio, TV, newspapers and magazines often bring erroneous information to the public. But also scientists do not agree. The scientific literature is full of papers showing that these fields can be dangerous and others showing that they are not. This holds true for virtually all possible endpoints and scientific disciplines that were studied, going from *in vitro* studies on cell proliferation, genetic and immunological effects, over animal experimental data on cancer and non cancer issues and human epidemiological investigations. It is not uncommon that controversial results are reported by the same laboratory. This results in claims of 'danger' when reference is made to essentially 'positive' papers (showing adverse biological effects) or claims of innocuity when only papers showing no effects are emphasized. It is clear that all (peer reviewed) scientific data should be considered and carefully analysed in order to come to a best possible 'weight of evidence' evaluation of risk. According to the WHO (World health Organisation) and ICNIRP (International Committee on Non Ionizing Radiation Protection) a single study does not provide the basis for hazard identification. It can at the best form the basis of a hypothesis. Confirmation of the results of any study is needed through replication and/or supportive studies. Only the resulting body of evidence forms the basis for science-based judgments by defining exposure levels for adverse health effects and no observable adverse effects.

This is recognized by most scientists all over the world and this explains why there were and still are many expert groups issued from the scientific community that evaluate(d) the alleged adverse health effects of radiofrequency fields in general, and mobile telephone frequencies in particular. It should be noted that radiofrequencies pose the additional problem (not encountered with other agents) that effects can be thermal or non thermal. At

high exposure levels cells or tissues can heat and thermal effects can be observed that are not obtained by normal environmental exposure levels as for example. when exposure is to radiation from a mobile phone base station antenna or when using the handset. Thermal effects are well known but experiments where thermal exposure levels were studied are not relevant in the discussion of “mobile phones and health”. Yet, often thermal exposure levels were used, even when the authors of the study claimed that they investigated non thermal exposure levels (wrong experimental set up and dosimetry). It is therefore also important to evaluate not only the biology but also the dosimetric aspects of an investigation.

The purpose of the present chapter is to give an overview of the conclusions of different (inter)national expert groups based on their analyses.

2. Evaluation of different expert group reports (2009-2011)

We found 33 expert group reports that were devoted to health effects of radiofrequency fields and that were published in the period 2009-2011.

2.1 ICNIRP reports (2009)

Statement on the “Guidelines for limiting exposure to time-varying electric, magnetic and electromagnetic fields (up to 300 GHz)”. The International Commission on Non-Ionizing Radiation Protection (ICNIRP). Health Physics 97(3):257-259 (2009).

Juutilainen J, Lagroye I, Miyakoshi J, van Rongen E, Saunders R, de Seze R, Tenforde T, Verschaeve L, Veyret B and Xu Z (2009) Exposure to high frequency electromagnetic fields, biological effects and health consequences (100 kHz – 300 GHz). In: Vecchia P., Matthes R., Ziegelberger G., Lin J., Saunders R., Swerdlow A., eds., Review of Experimental Studies of RF Biological Effects (100 kHz – 300 GHz), ICNIRP 16/2009, ISBN 978-3-934994-10-2 pp. 94-319.

The International Committee on Non Ionizing Radiation Protection (ICNIRP) consists of a main commission (12 members) and 4 subcommittee's: epidemiology (5 members), biology (8 members), physics (7 members) and Optics (7 members). Information on ICNIRP can be obtained at <http://www.icnirp.de>. ICNIRP works in close collaboration with WHO and publishes guidelines and statements (see above) as well as literature reviews that are prepared by their (subcommittee) members. The most recent review on biological effects of radiofrequency radiation is from 2009 (see above). It is a consensus report that was approved by all (sub) committee members and peer reviewed by other experts that do not belong to ICNIRP. The report took all peer-reviewed publications into consideration. It was later on updated and published as single review papers in the scientific literature (van Rongen et al., 2009; Verschaeve et al., 2010; Juutilainen et al., 2011). Recommendations (guidelines) are exclusively based on scientific grounds. Although many countries in the world do adopt the ICNIRP recommendations they are sometimes criticized for insufficient implementation of the precautionary principle. Yet, on pure scientific grounds the ICNIRP papers, recommendations and reviews may be considered of high quality.

Above mentioned ICNIRP documents indicate that it is not possible to deny the existence of non thermal effects following RF-exposure but they consider evidence in favour of such (adverse) effects very weak. Recent *in vitro* and *in vivo* cancer studies show that these effects are unlikely. Also recent epidemiological investigations (e.g., in 2009 already available results from the interphone study) were considered as being indicative for the absence of

cancer risk from mobile phones. Other studies that allowed a sufficient 'weight of evidence' evaluation did also not show any indication of health-related biological effects. ICNIRP therefore concluded that there are no indications of non thermal adverse health effects and that their recommendations from 1998 (ICNIRP, 1998) do not need to be adapted.

2.2 Scientific Committee on Emerging and newly Identified Health Risks (SCENIHR), EU, January 2009

Health Effects of Exposure to EMF, Directorate general Health & Consumers, European Commission. January 2009, pp. 83.

SCENIHR produces reports and advises on new technologies which may constitute a health risk for humans. Examples are nanoparticles, but also radiofrequency radiation as those applied in wireless communication systems. A detailed report on health effects of electromagnetic fields was published in 2007 and updated in 2009.

SCENIHR expert group members are selected following a call. Apart from 3 permanent members there were 6 nominated members, all well known in the field and covering different scientific disciplines. They discussed all peer reviewed (English) papers. When other papers were considered the reason for doing so was explained. Evaluation was done according to criteria that were well defined in advance. They included a particular attention to the reported study methods, the number of participants in a study (test and control population), the number of cells or animals that were analysed in the study, possible bias and confounders and dosimetry. Therefore not all papers were given the same weight or importance. Explanations were given when some studies were excluded from the discussion or where given less attention. The focus was on papers that were published after the 2007 report.

The summary and conclusions of the SCENIHR (2009) report were that it is unlikely that radiofrequency radiation is carcinogenic although further studies on long-term cancer effects are needed due to the long latency period for most brain tumours. Some investigations showed non reproducible associations between RF-exposure and self-reported symptoms. Most studies were negative. Overall, recent investigations did not show effects of RF-exposure on reproduction and development, whereas findings of effects on the nervous system (e.g., cognitive effects) were not consistent. Effects on EEG should be further investigated.

SCENIHR concludes that it is still not possible to exclude a small risk from RF-exposure. Therefore uncertainties that were identified in the 2007 report were still present. The weight of evidence analysis is nevertheless rather reassuring. There were no minority opinions.

SCENIHR recommends further research, especially long-term prospective studies, including studies on children.

2.3 Reports from the Dutch health council

Health Council of the Netherlands. Electromagnetic Fields: Annual Update 2008. The Hague: Health Council of the Netherlands, 2008; publication no.2009/02.

<http://www.gezondheidsraad.nl/sites/default/files/200902.pdf>

The Health Council is an independent scientific advisory body. Its task is to provide the government and parliament with advice in the field of public health and health/healthcare

research. The Standing Committee on Radiation and Health deals with questions relating to the health effects of exposure to radiation and questions surrounding the use of medical imaging techniques. Following the rise of technologies such as mobile telephony, attention has in recent years mainly focused on the risks of non-ionizing radiation. Applications, such as high-voltage power lines, also give rise to queries from time to time. The standing committee also monitors scientific developments in the field of ionizing radiation, ultraviolet radiation and ultrasound. Members of the standing committees are carefully selected so as to form a multidisciplinary group of independent experts.

The annual update 2008 (published in 2009) considered two different aspects of RF-bio effects: RF-effects on brain function and 'Electromagnetic Hypersensitivity'. It was prepared by the members of the "electromagnetic field committee" and discussed and approved by the standing committee "Radiation". The report includes a description of the criteria used in the evaluation process. These were inclusion of peer reviewed scientific papers of 'sufficient' quality only, attention for dose-effect relationships and reproducible or consistent results that were supported by quantitative and statistical analyses. Possible working mechanisms were also taken into consideration although absence of such mechanisms did not necessarily exclude plausibility of a causal relationship between exposure and effect. For human studies further attention was paid to 'double blind studies', the constitution of the control populations and other methodological aspects of the study (exposure regimes etc.). Minority opinions were allowed.

The Health Council's conclusion was that effects on brain function were described in some papers but that there were no indications that they might be hazardous. They also concluded that good quality papers do not support the existence of a causal relationship between RF-exposure and symptoms like headache, migraine, fatigue, itching, insomnia etc. But there was a relationship between supposed RF-exposure and subjective symptoms indicating the presence of a placebo effect. No advises were formulated.

The Dutch health council also published other reports or advises on the subject that we do not consider here (see <http://www.gezondheidsraad.nl/en>).

2.4 Statens strålskyddsinstitut (SSI = Swedish radiation protection agency)

Recent Research on EMF and Health Risks; Sixth annual report from the Independent Expert Group on Electromagnetic Fields, 2009

<http://www.stralsakerhetsmyndigheten.se/Allmanhet/>

The Swedish Radiation Protection Agency has appointed an independent international expert group for the evaluation of scientific developments and in order to provide advises on the possible health effects of electromagnetic fields. This working group takes into consideration other expert group reports as a basis for its discussions and reports that should be updated each year. The report from 2009 is the 6th and latest report that was published so far. It concerns *in vitro* and *in vivo* effects of radiofrequencies, in particular genotoxic and non genotoxic endpoints, effects on reproduction, neurodegenerative effects, immunological effects, behavioural effects, cancer etc. Also human studies were evaluated including investigations on brain activity, cognitive functions, sleep disorders, subjective complaints and epidemiological (cancer) studies. The working group consisted of 9 internationally renowned experts.

The report does not give an extensive description of the used methodology but it is clear that peer reviewed scientific papers were carefully evaluated. The conclusion of the report was that "...there are no new positive findings from cellular studies that have been well established in terms of experimental quality and replication." It also stated that "...recent animal studies have not identified any clear effects on a variety of different biological endpoints following exposure to RF-radiation typical of mobile phone use, generally at levels too low to induce significant heating." The SSI furthermore concluded that there are no indications of an increased cancer risk in mobile phone users (up to 10 years of exposure to mobile phone radiation). Absence of cancer risks (as by 2009) is consistent with the results from laboratory investigations in animals as well as with *in vitro* studies that did not identify a possible working mechanism. The working group also considered two studies on children that did not found any effect. In their evaluation of "electromagnetic hypersensitivity" the conclusion was that there were no indications other than the presence of a placebo effect. The self-declared hypersensitivity is however considered a real health problem (but not caused by the radiation) that should receive sufficient attention.

The SSI did not formulate particular advises but it emphasised the need of further studies, especially on children.

2.5 EFHRAN reports

Report on the analysis of risks associated to exposure to EMF: in vitro and in vivo (animals) studies, July 2010

http://efhran.polimi.it/docs/IMS-EFHRAN_09072010.pdf

Risk analysis of human exposure to electromagnetic fields, July 2010

http://efhran.polimi.it/docs/EFHRAN_D2_final.pdf

Members of the "European Health Risk Assessment Network on Electromagnetic Fields Exposure" (EFHRAN) belong to research institutes from 7 different European countries and are supported by external collaborators from 12 countries. All are international experts in research on non ionizing radiation. Some industrial groups, as for example the European 'consumer voice' in standardisation - ANEC and the GSM Association (GSMA) or the Network Operators' Association AISBL (ETNO) were associated to EFHRAN. The working group evaluated investigations on animals and humans. The role played by EFHRAN members and associated groups in the realisation of the report was not made very clear. The evaluation of effects were done according to a scoring method that is similar to the one used by IARC (International Agency for Research on Cancer). For each endpoint the evidence was evaluated as being "sufficient", "limited", "inadequate" or "inexistent" (= lack of evidence). A critical evaluation was performed of the relevant scientific literature which was based on the data provided by the SCENIHR (2009) report and on data that were published afterwards. The EFHRAN report was devoted to different kinds of non ionising radiation but we will here only consider the evaluation of studies on radiofrequency radiation.

The EFHRAN conclusions were as follows:

Cancer related studies:

- Limited evidence *in vitro* and lack of evidence with respect to *in vivo* investigations
- Inadequate evidence for non genotoxic effects
- Inadequate evidence from cancer studies in humans

Effect on the nervous system:

- Lack of evidence for effects on the blood brain barrier
- Limited evidence of effects on stress response genes and gene expression
- Lack of evidence with respect to behavioural effects
- Limited evidence from *in vitro* investigations
- Inadequate evidence in humans related to neurodegenerative diseases and RF-exposure

Effects on reproduction and development:

- Inadequate evidence concerning development and teratology
- Inadequate evidence for reproductive effects in animals and *in vitro* studies
- Inadequate evidence for effects in humans (e.g., behavioural effects in children from RF-exposed mothers)

Other effects:

- Lack of evidence for auditory effects
- Inadequate evidence of *in vivo* immunological effects
- Inadequate evidence for cardiovascular effects in humans
- No indications of electromagnetic hypersensitivity

2.6 Latin American expert committee on high frequency electromagnetic fields and human health, June 2010

Latin American Expert Committee on High Frequency Electromagnetic Fields and Human Health. Scientific review: Non Ionizing electromagnetic radiation in the radiofrequency spectrum and its effects on human health.

www.wireless-health.org.br/downloads/LatinAmericanScienceReviewReport.pdf

The goal of this study was to comply with the increasing anxiety of the population from Latin American countries with regard to their exposure to non ionizing radiations, especially from wireless communication systems (mobile phones, handset and base station antennas). The report was written by an expert panel which consisted of 5 scientists from different South American countries and a number of renowned international experts. The study was performed on request of the Eduled Institute for Medicine and Health which is a non-profit research- and development institute at Campinas, Sao Paulo (Brazil).

The study reviewed some 350 scientific investigations that were published since February 2010, with emphasize on studies that were performed in South America. Special attention was devoted to Risk Communication and application of the precautionary principle (which are usually not considered in other expert group reports). Attention was also given to regional and international exposure standards and recommendations from international bodies such as ICNIRP (International Committee on Non Ionizing Radiation Protection), IEEE (Institute of Electrical and Electronics Engineers), ITU (International Telecommunication Union) and the FCC (Federal Communication Commission, USA).

This is a well done study but it should be stressed that it is written by a limited number of persons that were assisted by an advisory group with obvious ICNIRP/WHO signature. It is therefore not surprising that the conclusions were similar to those of ICNIRP and WHO. The

conclusions were that there is insufficient evidence and lack of consistent data in favour of a causal relationship between low intensity radiofrequency radiation and short term adverse biological effects. The report acknowledge the existence of some alarming studies, e.g., on the blood brain barrier, but they were interpreted as due to thermal effects that are not relevant with respect to public exposures. Provocation studies in humans did not support the presence of health effects below thermal exposure levels. There were no indications of effects from mobile phone radiation on well being and no consistent indications of effects on cognitive functions, neurophysiologic and other physiologic or behavioural disorders. Epidemiological evidence is so far reassuring but it was acknowledged that we should await more studies on long term RF-exposures before any definite conclusion can be reached. The authors also stressed that it is not only important to investigate adverse health effects but that attention should also be paid to the benefits of wireless communication devices. They emphasize the need of correct information of the public via, for example, a central Latin American information centre for the general public and stakeholders. Not only biological effect studies are needed but also studies on socio-economic aspects of the mobile phone technology.

2.7 The Bioinitiative report (2007 – updated 2010)

BioInitiative: A Rationale for a Biologically-based Exposure Standard for Electromagnetic Radiation
www.bioinitiative.org/report/index.htm

This report was written by a number of individual scientists and public health and public policy workers who believe that existing public exposure standards for as well extreme low frequency fields (power lines) as radiofrequency radiation (mobile phones) are inadequate. Notably, not all authors were scientists and not all can be considered 'independent'. Possible conflicts of interest were not assessed. The purpose of this report was to assess scientific evidence on health impacts from electromagnetic radiation below current public exposure limits and to evaluate what changes in these limits are warranted now to reduce possible public health risks in the future. The report is a collection of a number of chapters, called 'sections', written by the individual authors. The sections were not written in a standardised way and there was apparently no consultation or discussion on these sections between the authors. The methods used to collect literature data were not defined. In most cases a selection of the available scientific material has been made in favour of those reporting alarming data (also from the non peer-reviewed literature) whereas negative (reassuring) data were often not reported. The selection criteria for inclusion or rejection of papers were not stated. The report is not a consensus report and the overall summary is often an over exaggeration that does not always comply with the content of the sections.

According to the report it is obvious that exposure to the electromagnetic fields, even at environmental exposure levels, constitute an important health risk for humans and that positive (alarming) data are reported (and considered very likely if not proven) for almost all biological endpoints that were investigated. The report therefore contains recommendations on establishing limits for exposure to electromagnetic fields that are much lower than the limits that are currently applied in many countries all over the world.

The report certainly has some merits but as stated above there are many shortcomings. A detailed evaluation of the Bioinitiative report and its shortcomings is for example given on the website of the Dutch Health Council and will therefore not be further detailed in this paper (http://www.gezondheidsraad.nl/sites/default/files/200817E_0.pdf).

2.8 The AFSSET report (2010)

Agence française de sécurité sanitaire de l'environnement et du travail (Afsset), Comité d'Experts Spécialisés liés à l'évaluation des risques liés aux agents physiques, aux nouvelles technologies et aux grands aménagements, Octobre 2009. Groupe de Travail Radiofréquences, mise à jour de l'expertise relative aux radiofréquences (Saisine n°2007/007) (2009).

www.afsset.fr/index_2010.php

The AFSSET became since 2010 the “French agency for Food, Environment and Occupational Health and Safety (now ANSES)”. It was asked by the French government to provide an overview and evaluation of the scientific knowledge on biological effects from mobile phone frequencies. The request was especially focussed on alleged effects on the blood brain barrier and epidemiological investigations on brain cancer in relation with mobile phone and other wireless applications of radiofrequency radiation. A working group was constituted according to strict criteria following a call for experts. Members were experts in the different relevant area of the subject, including medical doctors, biologists, biophysicists, epidemiologists, engineers (dosimetry) and human and social sciences (1 chairman and 12 members). The working group produced a report that was submitted to another expert committee (CES) of 26 members comprising 4 members of the AFSSET working group. There were also approximately 30 external auditors.

The report was written following several (13) meetings that were held between September 2008 and October 2009 comprising 19 auditions. A large database of publications was used essentially including peer reviewed (English) papers. Other reports (SCENIHR, Bioinitiative, etc.) were also consulted in order to identify publications that might have been overlooked.

It may be interesting to know that many of the members were not the usual players involved in research on non ionizing radiation bio effects and no members of other expert groups on the subject. They all possessed of course the necessary expertise to fulfil their tasks.

According to the AFSSET report there are no indications for short or long term adverse health effects as a result of exposure to radiofrequency radiation. Epidemiological investigations were reassuring but nothing can be said about long term effects that were not yet (sufficiently) investigated. Upon receipt of the report from the working group AFSSET concluded a little bit more mitigated. Due to the presence of some studies showing effects and hence remaining uncertainties further research is encouraged.

2.9 IARC (2011)

IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Volume 102: Non-Ionizing Radiation, Part II: Radiofrequency Electromagnetic Fields [includes mobile telephones, microwaves, and radar] – in press (2011)

www.iarc.fr/en/media-centre/pr/2011/pdfs/pr208_E.pdf

In May 2011, 30 scientists from 14 countries met at the international Agency for Research on Cancer (IARC) in Lyon, France, to assess the carcinogenicity of radiofrequency electromagnetic fields. The results of this meeting will be published in the IARC Monographs (nr. 102; *in press*). This monograph will contain information on (1) exposure data, (2) studies of cancer in humans, (3) studies of cancer in experimental animals, (4) mechanistic and other relevant data, together with a summary and final evaluation and rationale. A summary report is already

published (Baan et al., 2011). As for all other evaluations performed by IARC the evaluation of carcinogenic risks to humans of radiofrequency electromagnetic fields resulted from discussions that were held in different working groups (human cancer studies, animal cancer studies and other relevant topics + supporting group related to dosimetry) and in plenary sessions. Working group members were essentially chosen by IARC staff members based on their scientific merits as judged by their peer reviewed publications.

The general principles and procedures as well as the scientific review and evaluation process is well described in the IARC preamble document which can be found on the IARC website (<http://monographs.iarc.fr/ENG/Preamble/CurrentPreamble.pdf>). All participants have carefully filled in a conflict of interest document well in advance of the meeting as well as at the start of the meeting. Discussions were based on scientific reviews that were written before the meeting by some of the experts on subjects that belong to their field of expertise.

The evaluation of the carcinogenic risks to humans of radiofrequency fields results in a classification in one out of 5 categories (group 1, 2A, 2B, 3 or 4) as indicated in table 1. The decision is based on the human evidence and evidence in experimental animals where the designation “sufficient” evidence, “limited” evidence, “inadequate” evidence or “evidence suggesting lack of carcinogenicity” is given by voting. This results in an overall classification of the carcinogenic risk as indicated in figure 1. The overall evaluation can be changed (e.g., from group 2B to 2A, or 2B to 3) according to the arguments (evaluations) provided by the working group on mechanistic and other relevant data.

Group 1	<i>Carcinogenic to humans</i>	107 agents
Group 2A	<i>Probably carcinogenic to humans</i>	59
Group 2B	<i>Possibly carcinogenic to humans</i>	267
Group 3	<i>Not classifiable as to its carcinogenicity to humans</i>	508
Group 4	<i>Probably not carcinogenic to humans</i>	1

Table 1. Agents Classified by the IARC Monographs, Volumes 1-102 (<http://monographs.iarc.fr/ENG/Classification/index.php>)

IARC EVALUATION

		EVIDENCE IN EXPERIMENTAL ANIMALS			
		Sufficient	Limited	Inadequate	ESLC
Evidence In Humans	Sufficient	Group 1			
	Limited	Group 2A	Group 2B		
	Inadequate	Group 2B	Group 3		
	ESLC	Group 3			Group 4

Mechanistic data can be pivotable when the human data are not conclusive

Fig. 1. IARC evaluation based on evidence from human and animal data (figure provided by IARC).

According to IARC useful information was available regarding associations between the use of wireless phones and glioma, and to a lesser extent acoustic neuroma. The international Interphone study and studies from a Swedish research group (dr. Hardell) were found of most importance in the evaluation process. Both studies were found to be susceptible to bias – due to recall errors and selection for participation- but the working group nevertheless concluded that the findings of an increased risk at the highest exposed groups could not be dismissed as reflecting bias alone. A causal interpretation between exposure to mobile phone radiation and glioma and acoustic neuroma was therefore considered possible. The working group therefore decided that there is limited evidence in humans for the carcinogenicity of radiofrequency radiation. The working group also concluded that there is limited evidence in experimental animals for the carcinogenicity of RF-radiations. Although there was evidence of an effect of RF-radiation on some of the ‘other relevant endpoints’ the working group reached the overall conclusion that these results provided only weak mechanistic evidence relevant to RF-induced cancer in humans. Therefore, the conclusion is that radiofrequency fields should be classified in group 2B (possible carcinogenic; see Figure 1).

Radiofrequency radiation is thus classified in the same group (2B) than extreme low frequency magnetic fields, coffee and styrene. This raises some questions. Are their effects really comparable? Maybe the classification is not discriminative enough to allow differentiation in the overall EMF frequency range nor does it allow to sufficiently [account] for different qualities of underlying data. According to Leitgeb (2011a,b) other classification systems, e.g., the system developed in 2001 by the German Commission on Radiation Protection (SSK), allows categorization of evidence in other and more classes. Using this system Leitgeb assigned microwave radiation to class E0: “Lack of/or insufficient evidence for causality”. This illustrates that a classification in the IARC group 2B should not be interpreted by the public as proof of carcinogenicity at the same level as group 2A and 1. This is of course not correct but very often done.

2.10 French national academy of medicine (2009)

The academy stated that the precautionary principle may not be ‘misused’ to impose unscientific opinions. Scientific data are needed, not a subjective interpretation of the precautionary principle. According to the Academy “ *No mechanism is known through which electromagnetic fields in the range of energies and frequencies used for mobile communication could have a negative effect on health.*”

2.11 French academy of medicine, academy of sciences en academy of technologies (2009)

The National Academy of Medicine, the Academy of Science and the Academy of Technologies deplore the conclusions drawn by AFSSET from their experts’ report. The three Academies congratulate the experts for their work but roundly criticize the Agency’s recommendations. It does not understand why the presentation of the report does not insist on the reassuring aspects that are much more important than the few studies reporting effects. The latter are not to be considered credible alert signals. The academies also do not agree with the AFSSET recommendation to reduce exposure to cellular antennas that they consider scientifically not justified.

2.12 French health ministry (2009)

The website (www.sante.gouv.fr/effets-sur-la-sante.html) of the French Health Ministry was updated in August 2009. It states that the hypothesis that radiation from mobile phone base station antennas can be hazardous to man is no longer valid. It also stated that there are no indications so far that radiation from the handset poses a health risk but did not exclude that this may be the case. The Ministry proposed a number of simple measures to reduce the radiation exposure, especially for children.

2.13 French Parliamentary Office for the Evaluation of Scientific and Technological Choices (OPECST; 2009)

According to the report of this parliamentary organisation one cannot be completely sure that mobile phone radiation is absolutely safe but there are no proven effects so far. For this reason the report states that the ICNIRP guidelines remain valid.

2.14 Report from the Belgian superior health council (2009)

<http://www.health.belgium.be/eportal/Aboutus/relatedinstitutions/SuperiorHealthCouncil/index.htm?fodnlang=en>

The Belgian Superior Health Council (SHC) was founded in 1849. It is the scientific advisory body of the Federal Public Service "Health, Food Chain Safety and Environment". In order to guarantee and enhance public health, the council draws up scientific advisory reports that aim at providing guidance to political decision-makers and health professionals. The working group on Non Ionizing radiation of the SHC already made several reports/advises on topics related to wireless communication devices.

This advisory report nr. 8519 on standards for mobile phone masts is one of these. It follows previous advises on this topic and was issued in response to a request from the Minister of Public Health to supply the necessary elements for answering a letter sent by the GSM Operators' Forum (GOF) concerning masts that emit radio waves. In this letter, the GOF claims that the proposed standard of 3 V/m (at 900 MHz) is too rigid.

The SHC stresses that it takes the view that, on account of the scientific uncertainties, the precautionary principle must be applied in this case in order to protect the population and therefore it maintained its proposal of 3V/m. The SHC recommends once again that there should be a policy that favours independent measurements and research (biological effects, epidemiological studies, etc.). This should be done with the assistance of an administration that is competent in this matter and has sufficient staff at its disposal. Advise nr. 8519 (and previous ones) were promulgated before election of the new working group members who do not all agree with the conclusions and advises of the former working group. A revision of the advise in the light of new developments may be envisaged.

2.15 Bundestag (Germany, 2009)

This federal German authority confirmed the validity of the German radiofrequency exposure limits. This is based on the results of German research programmes on mobile

phones. According to the Bundestag the exposure limits in force indeed offer sufficient protection against mobile phone radiation.

2.16 The German Mobile Telecommunication Research Programme (DMF, 2009)

The “German Mobile Telecommunication Research Programme” (<http://www.emf-forschungsprogramm.de/>) started in 2002 and came to an end in 2008. It contained 54 research projects on mobile telecommunication including many different topics (laboratory research, epidemiology, dosimetry) but also aspects of risk communication. The general conclusion was that there is *no reason to question the protective effect of current limit values. Yet, because of the remaining question on health risks from long-term exposure for adults and children and the existence of some studies showing effects one should remain careful with wireless communication technologies.*

2.17 Commission on Radiological Protection (SSK, Germany, August 2009)

The German Commission on Radiological Protection (SSK) has issued a statement in which they reaffirm that there is no scientific evidence of a genotoxic effect (effects on the DNA) of radiofrequency fields or of an influence on gene regulation.

2.18 The Bundesamt für Strahlenschutz (BfS, German, 2009)

According to the German Federal office for radiation protection (BfS) recent studies have failed to demonstrate effects of mobile phone radiation on human fertility. No adverse effects were found on testes and sperm cells. The few papers that showed such effect(s) were considered of low or no scientific value. Experiments on animals have not shown relevant effects whereas *in vitro* studies only showed effects in case of thermal exposure conditions.

2.19 German expert group on children by the Jülich research institute (2009)

http://juwel.fz-juelich.de:8080/dspace/bitstream/2128/3683/1/Gesundheit_16.pdf

This report should be seen as an opinion document written by a limited number of international experts. It gives essentially a summary of different workshops that were held on mobile phones and children. The purpose of the report was to inform the public and authorities about the risks for children from the mobile phone technology. In the report on “Children’s Health and RF EMF Exposure” the expert group concluded that *the review of the existing scientific literature does not support the assumption that children’s health is affected by RF EMF exposure from mobile phones or base stations.* It is not very clear on what grounds this expert group was constituted. This study was supported by the telecom industry.

2.20 Radiation and Nuclear Safety Authority (STUK, Finland, 2009)

The Finnish Radiation and Nuclear Safety Authority stated in its 2009-report that there are no indications so far for long-term adverse health effects from radiofrequency radiation. However, everybody can reduce its own exposure easily if this is found useful.

2.21 Radiation authority of the five nordic countries (2009)

Five Northern European countries (Denmark, Finland, Iceland, Norway and Sweden) have joined to form the "Radiation Authority of the five Nordic Countries". They have issued a common statement which says that *"the Nordic authorities agree that there is no scientific evidence for adverse health effects caused by radiofrequency field strengths in the normal living environment at present. [...] The Nordic authorities therefore at present see no need for a common recommendation for further actions to reduce these radiofrequency fields."* *"Furthermore, in terms of overall public exposure, mobile phones are a much more significant source of radiofrequency radiation than fixed antennas. If the number of fixed antennas is reduced, mobile phones will need to use higher power to maintain their connection, thereby the exposure of the general public may increase."*

The authorities emphasize the need of further well conducted research on the alleged effects of radiofrequency fields on health.

2.22 CCARS scientific committee (Spain, 2009)

The "Comité Científico Asesor en Radio-frecuencias y Salud" (CCARS) published a literature survey and opinion on mobile phones and health. This was essentially based on the most recent reports and opinions from national and international authorities. They concluded that recent scientific/technical breakthroughs do not justify changes in the present RF benchmark levels and exposure limits for the public and workers.

2.23 Council of ministers of the isle of man (United Kingdom, 2009)

According to a working group report there is no general risk to the health of people living near mobile phone base station antenna. The exposures are limited and well below the guidelines. The group also stated that there is no proven relationship between self reported electromagnetic hypersensitivity and electromagnetic fields. At least some of the symptoms may be related to anxiety about the presence of the new technologies. They finally consider that the precautionary principle can be applied yet, especially with respect to children.

2.24 Institute of Engineering and Technology (IET, 2010)

Position statement on low level electromagnetic fields up to 300 GHz.
www.theiet.org/factfiles/bioeffects/postat02final.clin?type=pdf

This is an update of a previous position statement on *"The possible Harmful effects of low-level electromagnetic fields of frequencies up to 300 GHz"*. It claims that there are still no data in favour of adverse health effects from low level (normal) exposure to the radiofrequency fields. The IET has formulated its statement after consultation of the scientific literature using scientific databases (Medline, biosis, inspec) which provided a total of 813 relevant publications over the period 2008-2009. About half of them were on radiofrequency fields. They included cancer studies (e.g., Interphone study results), laboratory investigations in animals and cells, studies on non thermal working mechanisms and others. The statement also emphasize the need of independent replication studies and asks scientific journals to publish results from well sound scientific research only, whatever the results are. Scientists were encouraged to perform good science and to publish only when their work is of excellent quality.

2.25 Reports from the Health Protection Agency (HPA)

<http://www.hpa.org.uk/>

The Health Protection Agency (formerly National Radiological Protection Board) issues different reports and information booklets on different aspects of (amongst others) mobile phone effects. According to their 2010 statement “there are thousands of published scientific papers covering research about the effects of various types of radio waves on cells, tissues, animals and people. The scientific consensus is that, apart from the increased risk of a road accident due to mobile phone use when driving, there is no clear evidence of adverse health effects from the use of mobile phones or from phone masts”.

2.26 The Austrian ministry of health (2009)

The ministry states in a brochure that there is no scientific evidence that cellular phones are hazardous to man. The brochure yet recommends a reasonable use of a mobile phone and limited use by children.

2.27 Australian Radiation Protection and Nuclear Safety Agency (ARPANSA, 2009)

www.arpansa.gov.au/
www.arpansa.gov.au/pubs/eme/fact1.pdf

In an update of its fact sheet on mobile telephony and health ARPANSA says that “*there is essentially no evidence that microwave exposure from mobile telephones causes cancer, and no clear evidence that such exposure accelerates the growth of an already-existing cancer. More research on this issue has been recommended. “Users concerned about the possibility of health effects can minimize their exposure to the microwave emissions by limiting the duration of mobile telephone calls, using a mobile telephone which does not have the antenna in the handset or using a 'hands-free' attachment. “There is no clear evidence in the existing scientific literature that the use of mobile telephones poses a long-term public health hazard (although the possibility of a small risk cannot be ruled out).”*

2.28 Health Canada, July (2009)

<http://www.hc-sc.gc.ca/ewh-semt/radiation/cons/stations/index-eng.php>
<http://www.hc-sc.gc.ca/ewh-semt/radiation/cons/radiofreq/index-eng.php>
http://www.hc-sc.gc.ca/ewh-semt/pubs/radiation/radio_guide-lignes_direct-eng.php

Health Canada is the Federal department responsible for helping Canadians maintain and improve their health, while respecting individual choices and circumstances. It publishes different documents and fact sheets (see for example website addresses given above). According to these the consensus of the scientific community is that RF energy from cell phone towers is too low to cause adverse health effects in humans. In fact, worst-case RF exposure levels emitted from cell phone towers are typically thousands of times below those specified by science-based exposure standards. The RF energy from cell phones also poses no confirmed health risk but it is acknowledged that cell phone use is not entirely risk-free due to distraction, possible interference with some (medical) devices or other sensitive electronic equipment.

2.29 Food and Drug Administration (FDA, USA, 2009 – 2010)

<http://www.fda.gov/>

<http://www.fda.gov/Radiation-EmittingProducts/RadiationEmittingProductsandProcedures/HomeBusinessandEntertainment/CellPhones/ucm116282.htm>

<http://www.fda.gov/Radiation-EmittingProducts/RadiationEmittingProductsandProcedures/HomeBusinessandEntertainment/CellPhones/ucm116331.htm>

The FDA updated its pages on cellular telephones and health. It states that the weight of scientific evidence has not linked cell phones with any health problems. The steps adults can take to reduce RF exposure apply to children and teenagers as well.

2.30 National Cancer Institute (NCI, USA, September 2009)

<http://www.cancer.gov/cancertopics/factsheet/Risk/cellphones>

A fact sheet from the National Cancer Institute stated that studies thus far have not shown a consistent link between cell phone use and cancers of the brain, nerves, or other tissues of the head or neck. More research is however needed because cell phone technology and how people use cell phones have been changing rapidly.

2.31 US Health Physics Society (2010)

<http://hps.org/>

http://hps.org/documents/Mobile_Telephone_Fact_Sheet_update_May_2010.pdf

This society also publishes different fact sheets on mobile phones and wireless communication technologies. A recent one on mobile telephones does not deflect from previous ones as it still stated that the available evidence does not show that use of mobile phones or exposure to emissions from their base stations (cell towers) causes brain cancer or any other health effect.

2.32 Committee on Man and Radiation (COMAR, 2009)

<http://ewh.ieee.org/soc/embs/comar/>

This committee is a technical committee of the “Engineering in Medicine and Biology Society” (EMBS) of the “Institute of Electrical and Electronics Engineers” (IEEE). This committee is particularly interested in the biological effects of non ionizing radiations, including radiofrequency fields. The conclusions from their scientific evaluation stated that the scientific evidence is absolutely not in accordance with what the Bioinitiative project asserted. Indeed the weight of evidence does not support the safety limits recommended by the Bioinitiative group. COMAR recommends on the contrary that the public health officials continue to base their policies on RF safety limits recommended by established and sanctioned international organisations such as ICNIRP, IEEE etc.

2.33 WHO reports

<http://www.who.int/en/>

<http://www.who.int/peh-emf/publications/facts/factsheets/en/>

<http://www.who.int/mediacentre/factsheets/fs193/en/index.html>

WHO published different fact sheets on electromagnetic fields and their effects on human health. An update of the “mobile phone fact sheet 193 (June 2011) is not very different from the previous version(s). It still states that to date, no adverse health effects have been established as being caused by mobile phone use. It also says that it is still too early to fully assess long term effects in humans but that results of animal studies consistently show no increased cancer risk for long-term exposure to radiofrequency fields.

2.34 Council of Europe’s Committee on the Environment, Agriculture and Local and Regional Affairs (2011)

Committee on the Environment, Agriculture, and Local and Regional Affairs of the Council of Europe. The potential dangers of electromagnetic fields and their effect on the environment. 2011 May 6.

<http://assembly.coe.int/main.asp?Link=/documents/workingdocs/doc11/edoc12608.htm>

Jowitt T. *GSMA slams Euro call for ban on wireless in schools. eWeek Europe. 2011 May 16.*

www.eweekurope.co.uk/news/gsma-slams-euro-call-for-ban-on-wireless-in-schools-29363

<http://assembly.coe.int/Mainf.asp?link=/Documents/AdoptedText/ta11/ERES1815.htm>

This committee referred to the precautionary principle in order to ask for a reconsideration of the existing guidelines or exposure standards.

This committee consists of 47 members. It can influence decisions of the European Union but is not entitled to adapt existing regulations or to adopt new ones. According to the committee several measures should be taken. These include (1) adoption of reasonable measures to reduce exposure of children to electromagnetic fields, (2) a reconsideration of the ICNIRP guidelines and advises, (3) adoption of campaigns to alert the public, especially concerning health effects on children and adolescents, (4) adoption of measures to protect hypersensitive subjects, (5) encourage new scientific research to develop new less hazardous technologies, (6) A 0.6 V/m exposure limit for radiofrequency technologies such as wifi, WLAN, wiMAX, DECT and mobile phones and indication of SAR-values on the appliances, (7) increasing public information to protect children and a ban on RF-sources in schools (DECT, mobile phones, wifi, WLAN, WiMAX), (8) siting of antenna for wireless communication devices only after a public consultation and all antennas should be at a reasonable distance from dwellings, (9) creation of risk assessment procedures and protection of “early warning scientists”, and (10) research in biological effect studies should be encouraged by increasing research funds.

The report does not take into consideration the many other reassuring reports. Its conclusions are not based on a weight of evidence evaluation. The report has the merit that it brings forward the concerns of the public and that it proposes a number of measures that can be taken into consideration. Some of the proposed measures are however not very realistic, especially on the short run.

3. Summary of expert group evaluations

Table 2 gives a summary of the different expert group evaluations together with the main topics to which this evaluation refers and eventually formulated advises. The main result is

formulated as “-“ when the group concluded that there is no strong or insufficient evidence in favour of adverse health effects, or “+“ when in their opinion evidence is sufficient to conclude that there is a real health risk.

	EXPERT REPORT	CONCLUSION	ADVISES	+/-
1.	ICNIRP (2009) (all topics covered, advises/exposure standards)	No changes needed compared to previous advises	Recommandations (1998) remain valid	-
2.	SCENIHR (2009) (all topics covered, <i>in vitro</i> , <i>in vivo</i> , epidemiological investigations)	-no cancer risk identified -insufficient evidence for electromagnetic hypersensitivity, cognitive effects and reproductive and developmental disorders -Uncertainties remain	-Need for more long-term investigations - Further research needed on effects on EEG during sleep	-
3.	HEALTH COUNCIL OF THE NETHERLANDS (2008-2009) (electromagnetic hypersensitivity and effects on brain activity)	-No indications of effects on brain activity -No causal relationship between RF-exposure and complaints (hypersensitivity)	-	-
4.	SSI (2009) (epidemiological investigations, <i>in vitro</i> , <i>in vivo</i> studies)	-No strong indications of effects on health	-More research on children needed	-
5.	EFHRAN (2010) (human, <i>in vitro</i> and <i>in vivo</i> studies)	-No strong indications of effects on health. - <i>In vitro</i> studies show at the most some 'limited evidence'	-	-
6.	LATIN AMERICAN EXPERT GROUP (2010) (all topics covered, includes exposure standards and risk communication)	-Insufficient evidence for adverse health effects from <i>in vitro</i> and <i>in vivo</i> studies -Epidemiological investigations are reassuring but uncertainty remains regarding long-term effects -Also advantages of mobile phones are highlighted	-Need to continue research -Attention to and funds for socio-economical studies are also needed	-

	EXPERT REPORT	CONCLUSION	ADVISES	+/-
7.	BIOINITIATIVE REPORT (2007-2010) (all topics covered)	-RF-radiation is hazardous to humans, even at low (daily life) exposure levels (= below the current exposure standards). Hazards were identified for virtually all possible endpoints	-Much stronger exposure standards than the current ones are needed	+
8.	BELGIAN SUPERIOR HEALTH COUNCIL (2009-2010) (exposure standards for fixed antennas for mobile communication)	-Previous advises (3V/m at 900 MHz) remain valid	Exposure standards should be 3V/m based on the precautionary principle	(+)
9.	AFSSET (2010) (Effects of mobile phones, especially on the blood-brain-barrier and brain cancer)	-So far no indications of short-term and long-term effects -Long-term effects remain uncertain yet	-Further research needed -Exposure levels can be reduced	-
10.	FRENCH ACADEMY OF SCIENCES (2009) (all topics covered)	-No risks identified	-	-
11.	FRENCH ACADEMY OF SCIENCES AND TECHNOLOGIES (2009) (all topics covered)	-No risks identified	-Reassuring results should also be highlighted	-
12.	FRENCH MINISTRY OF HEALTH (2009) (all topics covered)	-No risks from base station antennas -No indications for risks from mobile phones (but still uncertainty)	-	-
13.	OPEST (F) (2009) (all topics covered)	-Adverse effects from mobile phone technology are not proven yet	-	-
14.	BUNDESTAG (D) (2009) (all topics covered)	-No risks -Adequacy of current German exposure standards is confirmed	-	-
15.	SSK (D) (2009) (Genetic effects)	-No scientific evidence in favour of genotoxicity of RF-radiation	-Existing exposure limits should not be adapted	-

	EXPERT REPORT	CONCLUSION	ADVISES	+/-
16.	BfS (D) (2009) (Fertility)	-No significant effects on testes and sperm	-	-
17.	GERMAN EXPERT GROUP ON CHILDREN (Jülich Research Institute) (2009) (risks for children)	-No indications of adverse health effects in children	-	-
18.	DMF (D) (2009) (general)	-No reasons to lower current exposure limits	-Further attention needed	-
19.	STUK (FIN) (2009) (general)	-No indications of long term effects	-	-
20.	RADIATION SAFETY AUTHORITY OF 5 NORDIC COUNTRIES (Scandinavia) (2009) (all topics covered)	-There is no scientific base to conclude that RF-radiation at "normal exposure levels" is hazardous to humans - There is no reason to lower existing exposure standards	-Further research is needed	-
21.	SSM (S) (2009) (<i>in vitro</i> , <i>in vivo</i> , human studies)	-No significant evolution in research data -No evidence for increased cancer risk	-	-
22.	CCARS (E) (2009) (general)	-No increased incidence of brain cancer -Uncertainties remain with respect to long-term effects -No reasons to lower existing exposure limits	-	-
23.	COUNCIL OF MINISTERS OF ISLE OF MAN (UK) (2009) (Antennas)	-No health risks for humans -Electromagnetic hypersensitivity related to mobile phones is not proven	-	-
24.	INSTITUTE OF ENGINEERING & TECHNOLOGY (IET) (UK) (2010) (all topics covered)	-No indications of health risks	-	-
25.	HEALTH PROTECTION AGENCY (HPA) (UK) (2010) (all topics covered)	-No danger from mobile phones (except traffic accidents)	-	-

	EXPERT REPORT	CONCLUSION	ADVISES	+/-
26.	AUSTRIAN MINISTRY OF HEALTH ((2009) (all topics covered)	- No danger from mobile phones	-Reasonable use of a mobile phone should be recommended, in particular by children	-
27.	ARPANSA (AUS) (2009) (all topics covered)	-No evidence for an increased cancer risk from mobile phone radiation	-Advises for reduction of exposure levels for those who wish to do so	-
28.	HEALTH CANADA (CAN) (2009) (all topics covered)	-No risks -Current exposure limits remain valid	-	-
29.	FDA (USA) (2010) (general)	-No risks from mobile phones (also in children)	-	-
30.	NCI (USA) (2009) (all topics covered)	-No adverse effects from a mobile phone -Uncertainty related to long-term effects warrants some care	-	-
31.	COMAR (INT) (2009) (all topics covered)	-Scientific data are not at all in accordance with the conclusions and assertions of the Bioinitiative report -Exposure limits (IEEE and other) are certainly adequate	-	-
32.	WHO (INT) (2010) (all topics covered)	-Adverse effects from mobile phones are not proven	-	-
33.	IARC/WHO (2011) (cancer)	RF-radiation is possibly carcinogenic in humans (group 2B in IARC classification)	-	(+)

Table 2. Summary of the expert group reports (scientific disciplines, conclusions and advises; -/+ : overall conclusion in terms of respectively absence of sufficient evidence for adverse health effects (-), or sufficient evidence for adverse health effects (+).

It can be seen from the table that the vast majority of the reports *do not* consider that radiofrequency fields at current exposure levels (especially from mobile phone base-station antennas and handsets) pose a serious health risk to humans. The only exception comes from the Bioinitiative report. All reports, except the Bioinitiative report, conclude that there is so far no clear indication of adverse health effects from RF-exposure from applications for wireless communication purposes. They usually remain prudent with regard to long-term bio-effects, not because of strong indications that such effects might occur, but only because

there are so far not enough data available to draw a sound conclusion. The same holds true for the IARC evaluation on carcinogenicity where the conclusion "possible carcinogenic" (group 2B) only means that, despite overall reassuring data, there is some limited evidence for carcinogenicity at long term exposures that cannot be ruled out so far. The Belgian Superior Health Council recommended more severe exposure limits (compared to most limits in application) but this recommendation is based on the precautionary principle rather than on solid arguments in favour of hazard or risk.

4. Evaluation of expert group reports based on 10 criteria

An evaluation of the different reports should take into account a great number of aspects. Amongst them the composition of the working group, the topics that were taken into account and the methods that were used are certainly some of the important aspects. We therefore tried to identify the members or participants in the working group activities and tried to see whether they constituted a *multidisciplinary* and *independent* group of experts. Did they evaluate all scientific (peer reviewed) publications, or did they make a selection of papers, and if so, what was the rationale for doing so? Was this satisfactory? Was the report a consensus report? Where minority opinions mentioned?

An evaluation of the reports bases on the answer to these questions can for example be done according to 10 criteria as indicated in Table 3. It is obvious that such an evaluation is always to a certain extent subjective. However, the purpose was not to make a ranking of the expert group reports according to their quality but especially to try to explain why they may (eventually) come to divergent conclusions on radiofrequency induced health effects. Because it is not possible to give in this chapter detailed answers to all the questions for each of the working groups the reports were given a score based on the answers and criteria indicated in table 3 (score of 0 when not a single criterion was met, up to 10 when all criteria were met).

Expert group:

- selection procedure of members and presence or absence of declarations of interest
- composition, complementarity and expertise of expert group members
- possibility to include minority statements

Methods used in the evaluation of the scientific data:

- peer reviewed publications, transparent procedure for selection of data
- method employed

Criteria for evaluation of scientific data:

- transparant and clearly described criteria
- attention to the number of participants/animals/cells considered in the studies
- attention to potential bias and confounding factors
- attention to dosimetry
- evaluation of used study methods and experimental set up in the studies under consideration

Table 3. Evaluation of expert group reports based on 10 criteria.

We did not make a full evaluation of all reports because some did not provide sufficient information or were not expert group reports *as such* as they were for example only opinion papers or short evaluations or advises as formulated in leaflets or fact sheets from certain organisations. In such cases a (re)examination of all available scientific data was not necessary and hence not attempted. Here, a “quality comparison” with the “bigger” reports would not be fair. The results of the evaluation are therefore only given as an example for a number of important reports (based on the criteria in Table 3, and summarized in Table 4).

It can be seen that most expert group reports got an excellent score, except the Bioinitiative report. This report certainly has merits and individual sections were often written by well renowned scientists, but overall it was deficient against most of the criteria as indicated before (see also http://www.gezondheidsraad.nl/sites/default/files/200817E_0.pdf). As mentioned before the purpose of the Bioinitiative report was to demonstrate that RF-radiation (at low-exposure levels as from mobile phones and their base station antennas) may be hazardous to humans. The purpose was to indicate that exposure limits should be considerably revised. The report was written in such a way that the outcome was in accordance with these goals.

As indicate above any such evaluation is always subjective to a certain extent, also because it is not always possible to fully appreciate the work that was done. Reports may mention that all peer reviewed papers were consulted but obviously this cannot be verified. They can mention that particular attention was paid to “conflict of interests” of the participating members, or report that literature data was carefully analysed and that particular attention was paid to, for example, aspects of biological dosimetry, but it was also not always possible to understand how this was done. Table 4 nevertheless can be useful as a general appraisal. It shows that most reports got a good to excellent ‘score’. Reports from ICNIRP, SCENIHR or the Dutch Health council got a maximal score of ‘10’ as they all fulfilled satisfactorily the 10 criteria of Table 3. All ICNIRP members are experts in non ionizing radiations bio-effects and/or dosimetry. Some questions can be raised on how the members were elected and in how far they constitute a balanced representation of opinions, but the methodology, through literature evaluations by subcommittee members and a careful and strong ‘peer reviewed’ process of their work can be seen as sufficient guarantee of quality. This justifies a high score although this does not automatically imply that ICNIRP opinions should be accepted without questions. ICNIRP was for example often accused of insufficiently applying the precautionary principle and hence of being not careful enough in its advises. This opinion can be defended. The same holds true for the reports from the Dutch Health Council. All criteria were met (= high score) which does not mean that the council is never criticized or criticisable. It is indeed often criticized, again for not applying the precautionary principle and insisting on absence of proof and lack of convincing data, hence not taking the few alarming data sufficiently into account. The Belgian Superior Health Council is on the contrary often criticized for emphasizing too much on the precautionary principle and providing advises that are scientifically not well sound. We have not extensively described their reports as they were *only* advises from a working group which did not perform a complete literature search and evaluation. The report from the IARC working group on Radiofrequency Electromagnetic Fields (including mobile telephones; Baan et al., 2011, and Monograph Volume 102, *in preparation*) also received a maximum score as it is based on an extensive evaluation of the scientific literature performed by a great number of experts and according to a well described and rigid procedure (see also

<http://monographs.iarc.fr/ENG/Preamble/index.php>). Special attention was also taken to conflict of interests.

We already mentioned that most of the reports express the same opinion. This is not surprising knowing that they are all based on the same scientific data and evidence and usually also similar and well defined criteria. Another reason for fairly concordant conclusions may be yet that different expert groups were often partly composed of the same scientists. The Swedish SSI report was written following constitution of an expert group from which some members were also members from ICNIRP. The same holds true for EFRAN and EDUMED. It is not surprising then that these expert groups expressed the same opinion or did not substantially deviate from the ICNIRP position.

Study	Subject	Expert group	Method	Quality	Score
ICNIRP, 2009	RF- Epidemiology, animals & in vitro studies	+++	++	+++++	10
SCENIHR, 2009	RF-ELF-IF-Static fields ; Epidemiology, <i>in vitro</i> & <i>in vivo</i> studies	+++	++	+++++	10
Dutch Health Council (2009)	RF- Epidemiology and experimental human studies	+++	++	+++++	10
SSI (IEG), 2009	RF - Epidemiology and <i>in vitro</i> & <i>in vivo</i> studies	++	+	+++++	8
EFRAN, 2010	RF - Epidemiology and <i>in vitro</i> & <i>in vivo</i> studies	++	++	+++++	9
EDUMED, Latin American Expert Group, 2010	RF- epidemiology, experimental human, <i>in</i> <i>vitro</i> & <i>in vivo</i> studies	+	++	+++++	8
BIOINITIATIVE, 2007/2010	RF-ELF- epidemiology, experimental human, <i>in</i> <i>vitro</i> & <i>in vivo</i> studies	+	+	+	3
AFSSET, 2010	RF- especially blood- brain-barrier, epidemiology and psychosocial and cultural aspects	+++	++	++++	9
IARC, 2011	RF- studies on cancer and cancer related aspects (epidemiology, <i>in vitro</i> & <i>in vivo</i> studies)	+++	++	+++++	10

Table 4. Evaluation of a number of important expert Group reports based on well defined criteria (cf. Table 3).

5. Conclusion

From the more than 30 expert group opinions that were published during the 2009-2011 period the vast majority did not consider that there is a demonstrated health risk from RF-exposure from mobile telephones and other wireless communication devices. Because of remaining uncertainties, especially with respect to long-term exposures, some caution is still expressed. This is the reason why IARC recently classified RF-electromagnetic fields as 2B-carcinogens (= possibly carcinogenic).

6. References

- Baan R., Lauby-Secretan B., El Ghissassi F. et al. on behalf of the WHO International Agency for Research on Cancer Monograph Working Group (2011) Carcinogenicity of radiofrequency electromagnetic fields. *Lancet Oncology*, 12, 624-626.
- ICNIRP (1998) Guidelines for limiting exposure to time varying electric, magnetic, and electromagnetic fields up to 300 GHz. *Health Phys.* 74, 494-522.
- Juutilainen J., Lagroye I., Miyakoshi J., van Rongen E., Saunders R., de Seze R., Tenforde T., Verschaeve L., Veyret B., Xu Z. (2011) Experimental studies on carcinogenicity of radiofrequency radiation. *Crit. Rev. Environ. Sci. Technol.* 41, 1664-1695.
- Leitgeb N. (2011a) Editorial. *Wien Med. Wochenschr.* 161,225.
- Leitgeb N. (2011b) Comparative health risk assessment of electromagnetic fields. *Wien Med. Wochenschr.* 161,251-262.
- van Rongen E., Saunders R., Croft R., Juutilainen J., Lagroye I., Miyakoshi J., de Seze R., Tenforde T., Verschaeve L., Veyret B., Xu Z. (2009) Effects of radiofrequency electromagnetic fields on the human nervous system. *J. Toxicol. Environ. Health B Crit. Rev.* 12, 572-597.
- Verschaeve L., Juutilainen J., Lagroye I., Miyakoshi J., van Rongen E., Saunders R., de Seze R., Tenforde T., Veyret B., Xu Z. (2010) In vitro and in vivo genotoxicity of radiofrequency fields. *Mutation Res.* 705, 252-268

Part 6

Biological Effects of Wireless Communication Technologies

Power Management in Sensing Subsystem of Wireless Multimedia Sensor Networks

Mohammad Alaei and Jose Maria Barcelo-Ordinas
*Computer Architecture Department, Universitat Politecnica de Catalunya, Barcelona
Spain*

1. Introduction

A wireless sensor network consists of sensor nodes deployed over a geographical area for monitoring physical phenomena like temperature, humidity, vibrations, seismic events, and so on. Typically, a sensor node is a tiny device that includes three basic components: a sensing subsystem for data acquisition from the physical surrounding environment, a processing subsystem for local data processing and storage, and a wireless communication subsystem for data transmission. In addition, a power source supplies the energy needed by the device to perform the programmed task. This power source often consists of a battery with a limited energy budget. In addition, it is usually impossible or inconvenient to recharge the battery, because nodes are deployed in a hostile or unpractical environment. On the other hand, the sensor network should have a lifetime long enough to fulfill the application requirements. Accordingly, energy conservation in nodes and maximization of network lifetime are commonly recognized as a key challenge in the design and implementation of WSNs.

Experimental measurements have shown that generally data transmission is very expensive in terms of energy consumption, while data processing consumes significantly less (Raghuathan et al., 2002). The energy cost of transmitting a single bit of information is approximately the same as that needed for processing a thousand operations in a typical sensor node (Pottie & Kaiser, 2000). The energy consumption of the sensing subsystem depends on the specific sensor type. In some cases of scalar sensors, it is negligible with respect to the energy consumed by the processing and, above all, the communication subsystems. In other cases, the energy expenditure for data sensing may be comparable to, or even greater (in the case of multimedia sensing) than the energy needed for data transmission. In general, energy-saving techniques focus on two subsystems: the communication subsystem (i.e., energy management is taken into account in the operations of each single node, as well as in the design of networking protocols), and the sensing subsystem (i.e., techniques are used to reduce the amount or frequency of energy-expensive samples).

1.1 Power consumption in sensing subsystem

In fact, the energy consumption of the sensing subsystem not only may be relevant, but it can also be greater than the energy consumption of the radio or even greater than the energy

consumption of the rest of the sensor node (Alippi et al., 2007). This can be due to many different factors (Raghunathan et al., 2006):

- Power hungry transducers. Some sensors intrinsically require high power resources to perform their sampling task. For example, sensing arrays such as CCDs or multimedia sensors (Akyildiz et al., 2007) such as CMOS image sensors generally require a lot of power. Also chemical or biological sensors (Diamond, 2006) can be power hungry as well.
- Long acquisition time. The acquisition time may be in the order of hundreds of milliseconds or even seconds, especially in the case of multimedia sensors. Hence the energy consumed by the sensing subsystem may be high, even if the sensor power consumption is moderate. In this case reducing communications may be not enough, but energy conservation schemes have to actually reduce the number of acquisitions (i.e. data samples). It should also be pointed out that energy-efficient data acquisition techniques are not exclusively aimed at reducing the energy consumption of the sensing subsystem. By reducing the data sampled by source nodes, they decrease the number of communications as well. Actually, many energy-efficient data-acquisition techniques have been conceived for minimizing the radio energy consumption, under the assumption that the sensor consumption is negligible.
- Power hungry A/D converters. Sensors like acoustic and seismic transducers generally require high-rate and high-resolution A/D converters. The power consumption of the converters can account for the most significant power consumption of the sensing subsystem, as in (Schott et al., 2005).

1.2 Multimedia sensing subsystem

One of the main differences between multimedia sensor networks and other types of sensor networks lies in the nature of how the image sensors perceive information from the environment. Most scalar sensors provide measurements as 1-dimensional data signals. However, image sensors are composed of a large number of photosensitive cells. One measurement of the image sensor provides a 2-dimensional set of data points, which we see as an image. The additional dimensionality of the data set results in richer information content as well as in a higher complexity of data processing and analysis. In addition, a camera's sensing model is inherently different from the sensing model of any other type of sensor. Typically, a scalar sensor collects data from its vicinity, as determined by its sensing range. Multimedia nodes are characterized by a directional sensing model, called Field of View (FoV, see Figure 1), and can capture images of distant/vicinal objects/scenes within its FoV from a certain direction. The object covered by the camera can be distant from the camera and the captured images will depend on the relative positions and orientation of the cameras towards the observed object (Soro & Heinzelman, 2005; Tezcan & Wang, 2008; Adriaens et al., 2006). Because of non-coincidence between neighborhood and sensed region by multimedia nodes, coverage-based techniques in WSN do not satisfy WMSN requirements.

Accordingly, the amount of power consumed in the sensing subsystem of a multimedia sensor node is considerably more than of a scalar ordinary sensor. For example, a

temperature sensor (texas instrument, 2011) as a scalar sensor consumes $6\mu\text{W}$ for sensing the environment. To have a view of multimedia sensors power consumption, table 1 shows the power consumed by four classes of cameras that are available today either as prototypes or as commercial products. At the lowest end of the spectrum is tiny Cyclops (Rahimi et al., 2005) that consumes a mere 46mW and can capture low resolution video. CMU-Cams (Rowe et al., 2002) are cell-phone class cameras with on-board processing for motion detection, histogram computation, etc. At the high-end, web-cams can capture high-resolution video at full frame rate while consuming 200mW , whereas Pan-Tilt-Zoom cameras are re-targetable sensors that produce high quality video while consuming 1W . It is noticeable that the mentioned power amounts are the power consumed by the camera sensors without considering the power consumed by the host motes, see (Tavli et al, 2011) for a survey of visual network platforms.

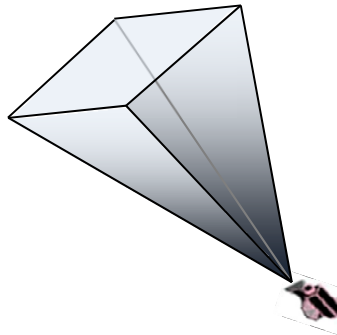


Fig. 1. The Field of View (FoV) of a multimedia sensor node.

Multimedia Sensor	Power of image capturing	Capability in image capturing
Cyclops	42 mW	Fixed angle lens, 352×288 at 10 fps
CMU-Cam	200 mW	Fixed angle lens, 352×288 up to 60 fps
Web-Cam	200 mW	Auto focus lens, 640×480 at 30 fps
High-end PTZ Camera	1 W	Pan-tilt-zoom lens, 1024×768 up to 30fps

Table 1. Power consumption and capabilities of four classes of camera sensors.

On the other hand, given the large amount of data generated by the multimedia nodes, both processing and transmitting image data are quite costly in terms of energy, much more so than for other types of sensor networks. Furthermore, visual sensor networks require large bandwidth for transmitting image data. Thus both energy and bandwidth are even more constrained than in other types of wireless sensor networks.

In this chapter, we describe a power efficient mechanism for managing the sensing subsystem of multimedia sensor nodes for surveillance in WMSNs. For this purpose, the deployed multimedia nodes are clustered according to their common covering regions and the clusters are managed to schedule the members to collaboratively survey the sensing area in a duty-cycled manner. With avoiding acquisition of redundant and correlated data, not only the sensing subsystem of nodes save its energy, but also the transmission and processing subsystems meet an optimized amount of data to be transmitted/processed and thus can conserve their residual energy. Therefore, the network lifetime is considerably prolonged.

The chapter is organized as follows. In section 2 we present an overview of work related to sensor management and scheduling policies. A surveillance mechanism with its details in grouping, management and scheduling multimedia nodes to be energy efficient is explained in section 3. Finally, the future work and conclusions are derived.

2. Sensor management and scheduling policies

In redundantly deployed multimedia sensor networks a subset of cameras can perform continuous monitoring and provide information with a desired quality. This subset of active cameras can be changed over time, which enables balancing of the cameras energy consumption, while spreading the monitoring task among the cameras. In such a scenario the decision about the camera nodes activity and the duration of their activity is based on sensor management policies. *Sensor management policies* define the selection and scheduling (that determines the activity duration) of the camera nodes activity in such a way that the visual information from selected cameras satisfies the application specified requirements while the use of camera resources is minimized. Various quality metrics are used in the evaluation of sensor management policies, such as the energy-efficiency of the selection method or the quality of the gathered image data from the selected cameras. In addition, camera management policies are directed by the application; for example, target tracking usually requires selection of cameras that cover only a part of the scene that contains the non-occluded object, while monitoring of large areas requires the selection of cameras with the largest combined FoV. While energy-efficient organization of camera nodes is oftentimes addressed by camera management policies, the quality of the data produced by the network is the main concern of the application.

The problem of finding the best camera candidates is investigated in (Soro & Heinzelman, 2007). In this work, the authors propose several cost metrics for the selection of a set of camera nodes that provide images used for reconstructing a view from a user-specified view point. Two types of metrics are considered: coverage aware cost metrics and quality-aware cost metrics. The *coverage-aware cost metrics* consider the remaining energy of the camera nodes and the coverage of the indoor space, and favor the selection of the cameras with higher remaining energy and more redundant coverage. The *quality-aware cost metrics* favor the selection of the cameras that provide images from a similar view point as the user's view point. Thus, these camera selection methods provide a trade-off between network lifetime and the quality of the reconstructed images.

Monitoring of large areas (such as parking lots, public areas, large stores, etc.) requires complete coverage of the area at every point in time. Such an application is analyzed in (Dagher et al., 2006), where the authors provide an optimal strategy for allocating parts of the monitored region to the cameras while maximizing the lifetime of the camera nodes. The optimal fractions of regions covered by every camera are found in a centralized way at the base station. The cameras use JPEG2000 to encode the allocated region such that the cost per bit transmission is reduced according to the fraction received from the base station.

Oftentimes, the quality of a reconstructed view from a set of selected cameras is used as a criterion for the evaluation of camera selection policies. In the work (Park et al., 2006)

distributed look-up tables are used to rank the cameras according to how well they image a specific location, and based on this, they choose the best candidates that provide images of the desired location. Their selection criterion is based on the fact that the error in the captured image increases as the object gets further away from the center of the viewing frustum. Thus, they divide the frustum of each camera into smaller unit volumes (subfrustums). Then, based on the Euclidian distance of each 3D point to the centers of subfrustums that contain this 3D point, they sort the cameras and find the most favorable camera that contains this point in its field of view. The look-up table entries for each 3D location are propagated through the network in order to build a sorted list of favorable cameras. Thus, camera selection is based exclusively on the quality of the image data provided by the selected cameras, while the resource constraints are not considered.

In order to reduce the energy consumption of cameras, the work (Zamora & Marculescu, 2007) explores distributed power management of camera nodes based on coordinated node wake-ups. The proposed policy assumes that each camera node is awake for a certain period of time, after which the camera node decides whether it should enter the low-power state based on the timeout statuses of its neighboring nodes. Alternatively, camera nodes can decide whether to enter the low-power state based on voting from other neighboring cameras.

Selection of the best cameras for target tracking has been discussed often (Pahalawatta et al., 2004; Ercan et al., 2006). Pahalawatta et al. present a camera selection method for target tracking applications used in energy-constrained visual sensor networks. The camera nodes are selected by minimizing an information utility function (obtained as the uncertainty of the estimated posterior distribution of a target) subject to energy constraints. However, the information obtained from the selected cameras can be lost in the case of object occlusions. This occlusion problem is further discussed by Ercan et al. where they propose a method for camera selection in the case when the tracked object becomes occluded by static or moving occluders. Finding the best camera set for object tracking involves minimizing the MSE of the object position's estimates. Such a greedy heuristic for camera selection shows results close to optimal and outperforms naive heuristics, such as selection of the closest set of cameras to the target, or uniformly spaced cameras. The authors here assume that some information about the scene is known in advance, such as the positions of static occluders, and the object and dynamic occluders prior probabilities for location estimates.

As a conclusion, in multimedia sensor networks, sensor management policies are needed to assure balance between the opposite requirements imposed by the wireless networking and vision processing tasks. While reducing energy consumption by limiting data transmissions is the primary challenge of energy-constrained visual sensor networks, the quality of the image data and application, QoS, improve as the network provides more data. In such an environment, the optimization methods for sensor management developed for wireless sensor networks are hard to directly apply to multimedia sensor networks. Such sensor management policies usually do not consider the event-driven nature of multimedia sensor networks, nor do they consider the unpredictability of data traffic caused by a monitoring procedure. Thus, more research is needed to further explore sensor management for multimedia sensor networks. Since sensor management policies depend on the underlying networking policies and vision processing, future research lies in the intersection of finding

the best trade-offs between these two aspects of visual sensor networks. Additional work is needed to compare the performance of different camera node scheduling sensor policies, including asynchronous (where every camera follows its own on-off schedule) and synchronous (where cameras are divided into different sets, so that in each moment one set of cameras is active) policies. From an application perspective, it would be interesting to explore sensor management policies for supporting multiple applications utilizing a single visual sensor network.

The presented mechanism in the following section groups multimedia nodes in clusters based on their common sensing region of the whole deployment region. The clusters monitor the environment independently but in each cluster the members collaborate in data acquisition in an intermittent manner. The scheduling and activity times in each cluster are determined based on the cluster population and the scale of overlapping between FoV of cluster members. So, the data transmissions are not limited in this kind of sensor management but the volume of sensed data is reduced by management in only sensing subsystem and applying coordination among cluster members to optimize capturing image times and to avoid redundant sensing of the same data in the overlapped FoVs. On the other hand, the sensing region is divided between clusters and each cluster monitors its domain with its exclusive frequency and member scheduling. Thus, clusters are not synchronized for sensing the region whiles each point of the sensing region is monitored frequently according to the number of nodes that cover that point by their sensing subsystem.

3. The surveillance mechanism

3.1 Preliminary

We assume wireless sensor nodes with fixed lenses providing a θ angle FoV, densely deployed in a random manner. The assumption of fixed lenses is based on the current WMSN platforms (Tavli et al, 2011). Almost all of them (SenseEye, MicrelEye, CITRIC, Panoptes, Meerkats) (Kulkarni et al., 2005; Kerhat et al., 2007; Chen et al., 2008; Feng et al., 2005; Margi et al., 2006) have fixed lenses and only high powered PTZ cameras have movement capabilities. We consider a monitor area with N wireless multimedia sensors, represented by the set $S = \{S_1, S_2, \dots, S_N\}$ randomly deployed. Each sensor node is equipped to learn its location coordinates and orientation information via any lightweight localization technique for wireless sensor networks. It is not the purpose of this chapter to define mechanisms to find this location. Without loss of generality, let us assume that nodes in the set S belong to a single-tier network or the same tier of a multitier architecture.

Our policy in order to applying collaboration among multimedia sensor nodes in the surveillance mechanism is clustering the network nodes based on their similarity in sensing the environment. The criterion applied in this purpose is the clustering scale of FoVs of nodes. The nodes having a large region of common area in their FoV, have a similar view of the sensing area then can cooperate in a established group, (Alaei & Barcelo, 2010).

3.2 Cluster formation and cluster membership

Now, let us consider the set $S = \{S_1, S_2, \dots, S_N\}$ of wireless multimedia nodes belonging to the same tier of a network randomly deployed. The cluster formation algorithm is executed in

a centralized manner by the sink after deploying the network. The main reasons in choosing a central architecture are the following: (i) for a distributed architecture, each node should notify to the rest of the nodes about its location A_i and its orientation α_i ($i = 1, \dots, N$). In a centralized architecture the nodes should notify to the sink their location and orientation. Note that this notification can be done using any energy efficient sensor routing protocol and only is necessary at bootstrap phase. All phases of the clustering algorithm are executed only one time, right after node deployment. (ii) In many WSN applications, the sink has ample resources (storage, power supply, communication and computation) availability and capacity which make it suitable to play such a role. (iii) Collecting information by a sink node is more power efficient compared to spreading this information to each and every other node within the network. (iv) Having the global view of the network at the sink node facilitates provision algorithms for closer-to-optimal cluster determination; the global knowledge can be updated at the sink when new nodes are added or some nodes die. Such maintenance tasks can be regarded as a normal routine for the sink. (v) Finally, using a centralized scheme can relieve processing load from the sensors in the field and help in extending the overall network lifetime by reducing energy consumption at individual nodes. The following phases are performed to establish and form clusters, (Figure 2):

- *Bootstrap*: At node bootstrap, each sensor $\{S_i, i = 1, \dots, N\}$ transmits its position (x_i, y_i) and orientation α_i to the sink. To accomplish this step any efficient sensor routing algorithm can be used. Thus, the clustering algorithm is not bound to how the sink receives this information. If there is an un-connected node in the network, it cannot announce itself and thus will not be considered in the algorithm.
- *Cluster Formation*: (i) Initially, the sink creates an empty cluster associated with an un-clustered multimedia node of S . Thus, that node will be clustered as the first member (*i.e.*, Cluster-Head (CH)) of the established cluster. (ii) Then, the sink finds the qualified un-clustered nodes for joining to the CH by computing the area of overlapped polygons of their FoV. From position and orientation of nodes, the sink computes the overlapped region between each un-clustered multimedia node and the CH of the established cluster. For calculating the FoV overlapping area of two nodes, we first survey the intersection of their FoVs. Second, if they intersect each other, we find the intersection region and at last, compute the area of the polygon. For this purpose, in the first step, we define the equations of the sides of FoVs using the vertex coordinates. Then, the intersection of each side of each FoV to all sides of the other is calculated. A decomposition approach is used for calculating the area of the overlapping region of FoVs. If the computed overlapped region is equal or greater than the threshold considered as the *Clustering Scale* (γ) -the minimum region that has to be overlapped between two node FoVs to be grouped in a cluster-, the un-clustered node will be clustered as a member of the established cluster. (iii) When no more nodes can be added to the cluster, the sink takes a new un-clustered node, begins a new cluster and goes to step (ii).
- *Membership notification*: we assume that the sink uses any energy-efficient sensor routing algorithm to notify to each first-member of every cluster about its cluster-ID and what are the members of the cluster. Then, each first-member sends a packet to the members of his cluster notifying them about the cluster which they belong to.

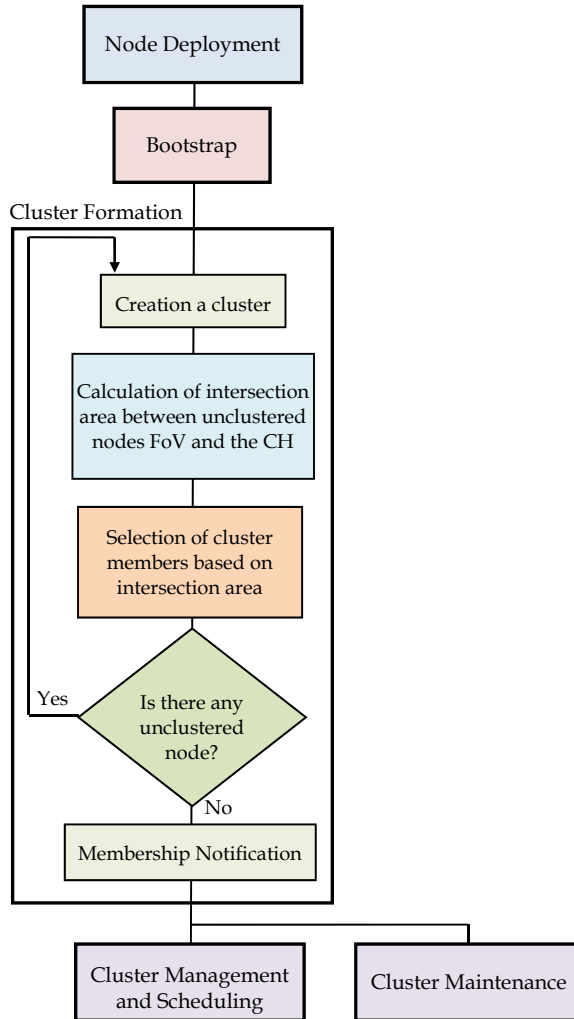


Fig. 2. Clustering Procedure.

The algorithm is executed by the sink once upon deployment and thus all nodes will become clustered. If a node joins to the network hereinafter, it has to send its position and orientation to the sink for announcing itself as a new node. The sink computes the FoV of the new node and finds the first cluster that can accept it as a new member. For this purpose, the sink computes the overlapping regions between FoV of the new node and the CH of each cluster and checks whether he is satisfying the cluster membership test. Then, the sink sends a message to the CH in order that this node re-organizes the cluster with the new member. Depending on the application, this notification may suppose a new reconfiguration in the monitoring task (*i.e.*, a new duty-cycle period). On the other hand, each node periodically sends a Hello message to the CH notifying its current residual

energy. When a node dies, the CH will notify the rest of the members about the new cluster set and will reconfigure any parameter related to the cluster. The CH also periodically compares the residual energy of cluster members and its residual energy to select the new CH with the maximum residual energy in the cluster. If the CH decides to entrust CH role to another cluster member, notifies to the cluster members about the new CH. Note that the beaconing among cluster members implies low overhead since clusters have few nodes and hello periods can be on the order of duty-cycle sensing periods.

3.2.1 Intra-cluster collaboration

Let us see the potential of cooperative node monitoring in clusters in terms of sensor area coverage. We define the Maximum Cluster Coverage Domain (MCCD) parameter for a cluster as the maximum monitoring area which is covered by that cluster. Since each cluster is established considering the clustering scale equal to γ , the MCCD can be computed as follows (C_{size} is the size of the cluster):

$$MCCD = \gamma \cdot A_{FoV} + (1 - \gamma) \cdot A_{FoV} \cdot C_{size} = (C_{size} - \gamma \cdot (C_{size} - 1)) \cdot A_{FoV} = \beta \cdot A_{FoV} \quad (1)$$

where:

$$1 \leq \beta = C_{size} - \gamma \cdot (C_{size} - 1) \quad (2)$$

The effective cluster covering domain can be inferior to the MCCD calculated by Equation (1) since some nodes can overlap more than the region determined by γ . Since MCCD gives us an upper bound on the area covered by the cluster, using MCCD will allow us worst-case dimensioning. Factor β represents the increment of area that the cluster senses with respect to an individual sensor. When each node of a cluster obtains an image from its FoV, a part of the related MCCD with a ratio at least equal to $1/\beta$ respect to the MCCD is captured whereas this part includes overlapped areas of other nodes in the cluster. Sensing the environment by each member delivers information not only from the FoV of the active node but also from some overlapped parts of other nodes in the same cluster: at least $\gamma \cdot A_{FoV}$ of the area is common to the first-member and more than $1/\beta$ of the MCCD is monitored. For example, in a cluster consisting of just 2 members, assuming a clustering scale of $\gamma = 0.5$, the MCCD is $1.5 \cdot A_{FoV}$. Thus, when each of the two members of the cluster is activated and monitors the environment, an area of one FoV is captured that is at least $2/3$ of the whole MCCD of the cluster. Consequently, scheduling and coordination among members in order to sense the field in a collaborative manner may yield a gain in energy saving and performance efficiency even with a low number of members in the cluster.

3.2.2 Cluster formation evaluation

All sensor nodes have been configured with a FoV vertex angle of $\theta = 60^\circ$ and R_s of 20 m. A sensing field spanning an area of $120\text{m} \times 120\text{m}$ has been used. Sensor densities were varied to study the cluster formation from sparse to dense random deployments. Figures illustrate the average results of 50 independent running tests whereas each test corresponds to a different random deployment. Once a random deployment is defined, cluster formation is obtained from node location, angle of orientation and FoVs of nodes, using the described method whose complexity is $O(N \cdot \log N)$. Furthermore, as it was mentioned before, each

node sends a packet to the sink in the bootstrap phase, then the sink notifies each CH via one packet his membership set for that cluster (phase 3) and then the CHs notify cluster nodes about their cluster membership and any related parameter. Thus, the average overhead of the algorithm is forwarding N packets from the nodes to the sink and forwarding N_C packets from the sink to first-members and forwarding $N_C (\mu_{Csize}-1)$ packets from CHs to cluster nodes; where N is the number of nodes, N_C is the average number of clusters and μ_{Csize} is the average cluster size. So the total overhead will be: $N + N_C + N_C (\mu_{Csize}-1)$ packets. The maintenance overhead is $N_C (\mu_{Csize}-1)$ beacons every keep-alive period, where the keep-alive period can be a multiple of the sensing duty-cycle period.

3.2.2.1 Number of clusters and cluster-size

The average number of clusters, μ_{NC} , and the average cluster-size (μ_{Csize}) in a tier/network for different node densities with several clustering scales are shown in Figures 3 and 4. Increasing the node density does not only cause an increment in the number of clusters but also yields more overlapping areas among FoVs and thus raises the cluster-size. However, the clustering scale (γ) also impacts in the cluster membership selection process. The clustering scale determines the minimum region that is required to be overlapped between the FoV of each node belonging to a given cluster and the FoV of the CH of that cluster. So, γ determines the minimum intersection part of FoV of each member with the CH of an established cluster. Lower clustering scales oblige less overlapping areas for cluster membership and increase the domain covered by a given cluster since more nodes will be conforming to the membership rule. Increasing the clustering scale restricts node membership because of higher required overlapping areas between FoVs of nodes. Thus, higher clustering scales result in lower cluster-sizes, less MCCD and thus higher number of clusters.

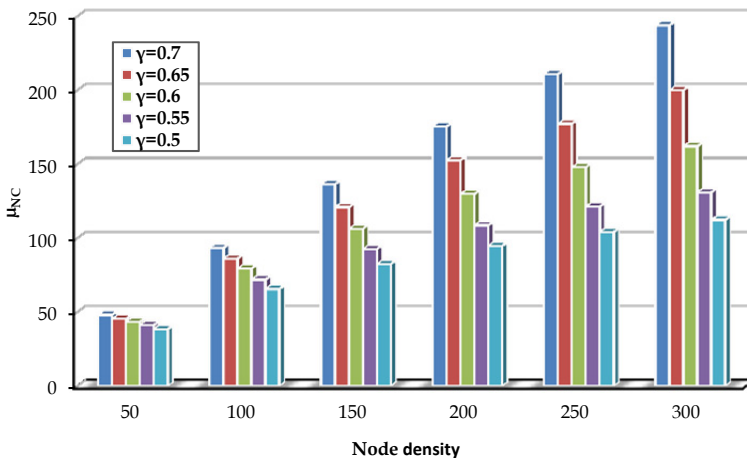


Fig. 3. Average number of established clusters.

Sparse networks have low average cluster-size, μ_{Csize} , because sparse deployments result in low overlapping areas. Moreover, high values of γ also will produce low μ_{Csize} . The result

will be lower potential for node coordination. On the other hand, dense wireless multimedia sensor networks can particularly benefit from higher cluster sizes and thus more potential for node coordination.

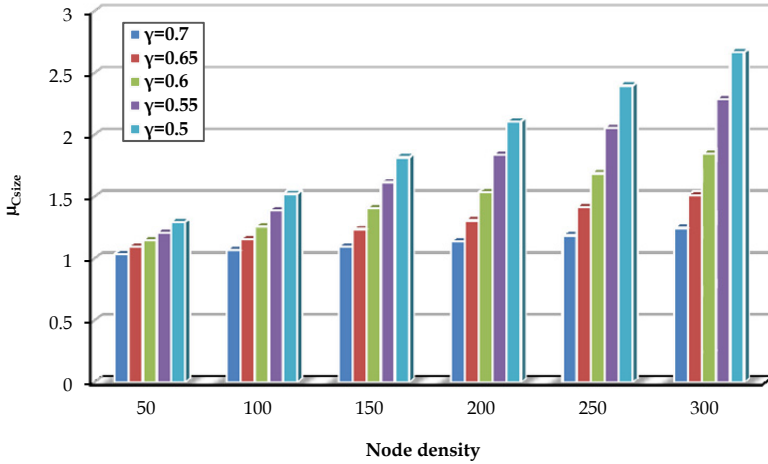


Fig. 4. Average size of established clusters.

Finally, Figure 5 shows the cumulative probability function for the cluster-size in the network for different node densities assuming a clustering scale of $\gamma = 0.5$. For example, in a network consisting of 250 nodes, 28% of clusters have a single member which does not have enough overlapping with others to satisfy the clustering scale, 32% of clusters have a cluster size of 2, 21% of 3, 12% of 4 and 7% of them consisting of more than four members.

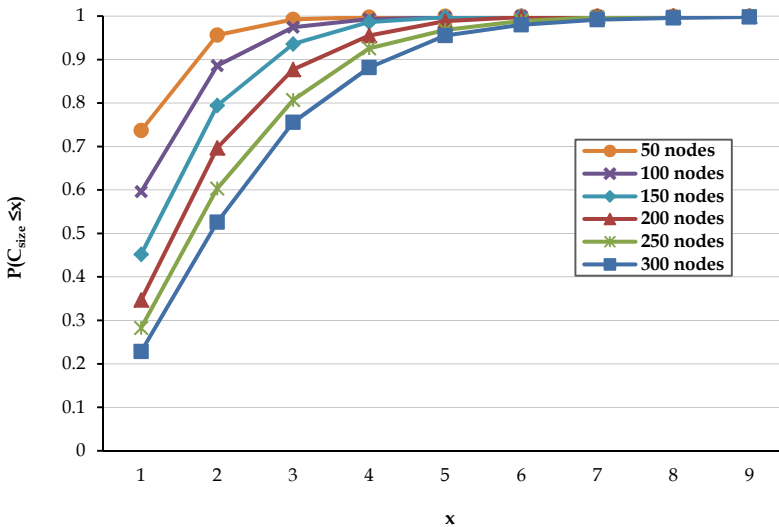


Fig. 5. The cluster size cumulative distribution function ($\gamma = 0.5$).

3.2.2.2 Coverage

Figure 6 illustrates the percentage of area that is covered by the random deployment in terms of node density. As it is shown in the figure, for covering 95% of the area, a dense deployment of 300 nodes is required. As the figure shows, the rate of increment of the covered area for low node densities is faster than for high node densities. This indicates that after a new node is added in a dense deployment, low new coverage area is obtained.

For example, the first 100 nodes cover 75% of the field, but the next 100 nodes will only cover 15% of new area. The conclusion is that dense networks are able to cover high areas at the cost of high overlapping and sensing redundancy, but this overlapping can be used for improving reliability if nodes belonging to the same cluster work in a coordinated manner. Furthermore, the existence of obstacles produces a reduction of the sensing area because of FoV occlusion effect, (Tezcan & Wang, 2008). So, employing dense networks of low-cost, low-resolution and low-power multimedia sensor nodes instead of sparse networks of high-power, high-resolution sensors (*e.g.*, PTZ) will be more beneficial.

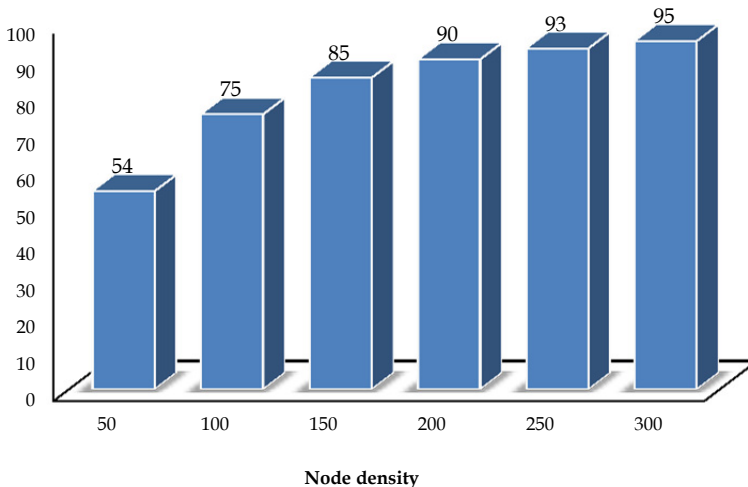


Fig. 6. Percentage of the covered area with respect to the whole deployment area.

Applications that are interested in multiple views will also benefit from this situation, since there will be several nodes monitoring the same area from several perspectives. Applications that are interested in detecting objects and are not interested in having an instantaneous multiple-view of the object may benefit from collaborative node processing in terms of energy savings. For the first set of applications clustering of nodes may serve as an indicator of triggering simultaneous multi-perspective pictures. For the second set of applications, clustering may serve as a baseline framework for collaborative node scheduling avoiding redundant sensing and processing and thus increasing network lifetime. Other applications that are interested in correlated data (*e.g.*, Distributed Video Coding, DVC) may use clustering in order to exploit multi-view correlations to build joint encoders (Pereira et al., 2008).

3.3 Cooperative node selection and scheduling

In monitoring mechanisms, usually cameras should perform duty-cycled monitoring over the area that they sense. That means that every T (Figure 7.a) seconds the sensors in the monitored area will awake and monitor the area. This is the situation for a planned network in which every sensor is placed in such a position that there is no overlapping among sensors. Nevertheless, this duty-cycle scheduling will produce high power consumption in those situations in which there are overlapping sensors, since camera nodes with overlapping areas do not cooperate to sense the area and thus they redundantly monitor the area.

In this section, we explain a cooperative mechanism based on the clustering method that coordinates nodes belonging to the same cluster to work in a collaborative manner to monitor the sensing area. The objective of this mechanism is to increase power conservation by avoiding similar sensing and redundant processing at the same time. Also, collaborative sensing by nodes that have FoVs intersecting each other yields to more reliability: cluster members will monitor the region sequentially and if a moving object is not detected in one image capturing, it will be in the vicinal FoVs at the next capturing times. Thus, the other members in the same cluster may detect the object.

Let us divide the environment in domains covered by clusters of nodes (MCCD, Section 3.2). All clusters concurrently sense their domains. In each cluster, members are awakened sequentially in an intermittent manner by the CH with a time interval related to the cluster-size and the scale of clustering (see Figure 7.b); (i.e., T_{interval} is the time between awakening two consecutive members of a cluster). In this way, each node of a given cluster periodically participates in capturing an image from its unique perspective and surveillance the environment and finally sleeps again with a cluster-based period called T_p . Formulas for these periods are derived in Section 3.3.1.

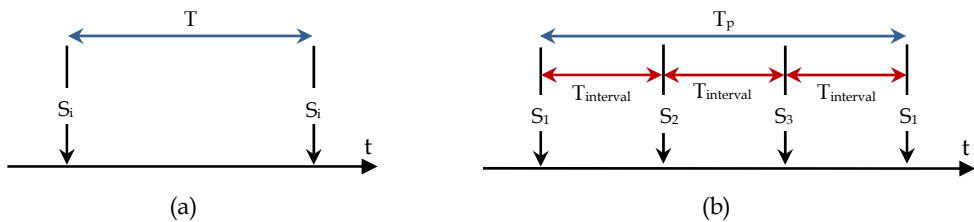


Fig. 7. (a) Period of awakening a given node in the un-cooperative scheduling. (b) Scheduling for a cluster consisting of three members (S_1 , S_2 , S_3).

3.3.1 Cluster-based T_p and T_{interval} computation

Let us consider as baseline mechanism a non-collaborative duty-cycled scheme in which every node awakes with an interval period of time T and monitor the area (i.e., takes a picture and performs object detection) as tier 1 in (Kulkarni et al., 2005). The objective of the collaborative mechanism is to produce a cluster-based duty-cycled scheduler in which: (i) Each node is awakened and senses the area with a reliable period of $T_p > T$ taking advantage of the overlapping among nodes in the cluster, thus, saving energy and increasing network

lifetime. Each cluster will have its own T_p interval, determined according to the cluster-size and the clustering scale. (ii) During the sleeping period of each member of a given cluster, other nodes belonging to the cluster are awakened with intervals of $T_{interval} < T$ (that is equal to: T_p / C_{size}) in a sequential manner.

The area sensed by each cluster is related to the MCCD area. In order to compute T_p we will consider the MCCD area. By awaking each member of a given cluster, in average, a part of the related MCCD with a ratio equal to $1/\beta$ is captured (Equation (2)). Note that the MCCD is an area of $\beta \cdot A_{FoV}$ and is sensed by C_{size} overlapping members, thus sensing the environment by each node delivers information not only from the FoV of the awakened node but also from some overlapped parts of the FoV of other nodes in the same cluster. Then, we may define the node interval duty-cycle period as:

$$T_p = T \cdot \frac{C_{size}}{\beta} = T \cdot \frac{C_{size}}{C_{size} - \gamma \cdot (C_{size} - 1)} \quad (3)$$

Note that the T_p is proprietary for each cluster in terms of its cluster-size and clustering scale. As it was mentioned before, the MCCD calculated by Equation (1) is the maximum covering domain of a cluster while the effective cluster covering domain may be less than MCCD since some members of a given cluster may overlap more than the region determined by γ . Consequently, a given cluster can cover an area less than $\beta \cdot A_{FoV}$. Thus, using β gives us the lowest interval T_p and thus the most reliable one since lower values of β would increase the interval T_p . On the other hand, members of a cluster are awakened sequentially to sense their environment in an intermittent way with time intervals equal to $T_{interval}$:

$$T_{interval} = \frac{T_p}{C_{size}} = \frac{T}{C_{size} - \gamma \cdot (C_{size} - 1)} \leq T \quad (4)$$

Let us consider Figure 6.b and for example a cluster with three members, $C = \{S_1, S_2, S_3\}$, cluster-head S_1 and $\gamma = 0.5$. Every node will be awakened every $T_p = 1.5 \cdot T$ seconds and the area will be monitored every $T_{interval} = 0.5 \cdot T$ seconds. As can be observed, every sensor is awakened with a period higher than the non-collaborative scheme but the area is monitored more times. Then, the area duty-cycled frequency is increased while the sensor duty-cycled frequency is reduced.

Table 2 shows the evolution respects of T_p and $T_{interval}$ to T as a function of γ for several C_{size} . We first have to notice that for a clustering scale factor $\gamma = 1$, $T_p = T$, while for $\gamma < 1$, $T \leq T_p \leq T / (1 - \gamma)$. Then, the duty-cycle frequency at which a specific node is awakened is decreased by a factor that at least is $(1 - \gamma)$ times the frequency of the non-collaborative scheme. On the other hand some sensor of the cluster will be on duty every $T_{interval}$ seconds. Note that $T_{interval}$ will be lower than T and will be smaller as C_{size} increases. This means that the area is monitored more frequently although every specific sensor monitors with less frequency. The reason is justified in how clusters are formed. Any sensor of the cluster overlaps with the first-member by at least an area of $\gamma \cdot A_{FoV}$. Thus, when a sensor enters in duty, he will monitor an area equal to $\gamma \cdot A_{FoV}$ overlapped with the first-member and an area equal to $(1 - \gamma) \cdot A_{FoV}$ that in the worst case does not overlap with any other member of the cluster. Sensing the whole cluster area with $T_{interval}$ equal to T would result in that an area equivalent

to $(1-\gamma) \cdot A_{\text{FOV}}$ would be monitored every $C_{\text{size}} \cdot T$, a value that can be very high. However, using Equation (3), monitoring of the area equivalent to $(1-\gamma) \cdot A_{\text{FOV}}$ is guaranteed by a monitoring interval that is not superior to $T/(1-\gamma)$, that is much lower than $C_{\text{size}} \cdot T$.

$\gamma \backslash C_{\text{size}}$	5	4	3	2
0.5	1.67	1.60	1.5	1.33
0.55	1.79	1.70	1.58	1.38
0.6	1.92	1.82	1.67	1.43
0.65	2.08	1.95	1.77	1.48
0.7	2.27	2.11	1.88	1.54

(a)

$\gamma \backslash C_{\text{size}}$	5	4	3	2
0.5	0.334	0.4	0.5	0.665
0.55	0.358	0.425	0.527	0.690
0.6	0.384	0.455	0.557	0.715
0.65	0.416	0.488	0.590	0.740
0.7	0.454	0.528	0.627	0.770

(b)

Table 2. (a) T_p/T , (b) T_{interval}/T for different cluster sizes and clustering scales.

Sleep/wake up protocols has extensively been studied in the area of wireless sensor networks, mainly for the radio subsystem, (Anastasi et al., 2009). Our clustering algorithm works on the sensing subsystem. It is important to notice that executing object detection does not imply sending packets to the sink. Thus, the sleep/wake up algorithm can be decoupled with the radio subsystem. Sleep/wake up can be based on periodic duty-cycle synchronized by the first-member: every T_p period, the sensing subsystem wakes up and performs object detection. However, clock drifts can cause cluster de-synchronization. To handle resynchronization, the system makes use of the beaconing scheme for cluster maintenance: nodes receive periodical beacons from the first-member and vice versa in order to detect new members or to detect members that have died. Beaconing duty-cycling belongs to the radio subsystem and it is independent of the sensing subsystem. That means that waking up the sensor to send a beacon is independent of waking up the sensor to take a picture and perform object detection. Thus, the cluster-head may resynchronize cluster members without need of waking up the sensing subsystem.

3.4 Lifetime prolongation evaluation

To evaluate the scheduling scheme in terms of power conservation, we compare the cooperative scheduled scheme with a single-tier network or a tier of a multi-tier architecture consisting of N nodes monitoring without coordination among them as (Rahimi et al., 2005; Kulkarni et al., 2005; Feng et al., 2005), in which, nodes are awakened with a time period of T . We note that the evaluation is over the sensing subsystem and that the radio subsystem (*i.e.*; transmission and reception of packets) is not taken into account.

The energy consumed in the network for object detection by N nodes during a duty-cycle interval of T in the non-collaborative scheduling is:

$$E = N \cdot (T_{sleep} \cdot P_{sleep} + E_{w_up} + E_{cap} + E_{detect}) \tag{5}$$

where T_{sleep} and P_{sleep} are the period and power consumption for a node in sleep mode. E_{w_up} , E_{cap} and E_{detect} respectively are the energies consumed in waking up a node, capturing a picture and performing object detection.

Let us now consider the cooperative scheduling algorithm in a clustered tier/network. Both, the interval between waking up consecutive nodes in the same cluster and the period of waking up a given node are functions of the cluster-size of the cluster which the nodes belong to. In one hand, in clusters with high cluster-size, $T_{interval}$ is small and thus cluster duty-cycle frequency is increased. On the other hand, higher number of nodes in the cluster causes longer periods T_P for awaking a given node of the cluster and thus yields an enhancement for power conservation in cluster's members. Assuming average cluster-size for all clusters in the tier/network, T_P will be:

$$T_P = \frac{T \cdot \mu_{C_{size}}}{\mu_{C_{size}} - \gamma \cdot (\mu_{C_{size}} - 1)} \tag{6}$$

where T is the base period for waking nodes in the base un-coordinated tier. Figure 8 shows the evolution of T_P normalized by T (i.e., $\mu_{C_{size}}/\beta$) for several node densities and clustering scales, γ . We may observe that the node average duty-cycle frequency is reduced by factors that are, for example, on the order of 0.78 for a 200 node network and a scale factor of $\gamma = 0.6$.

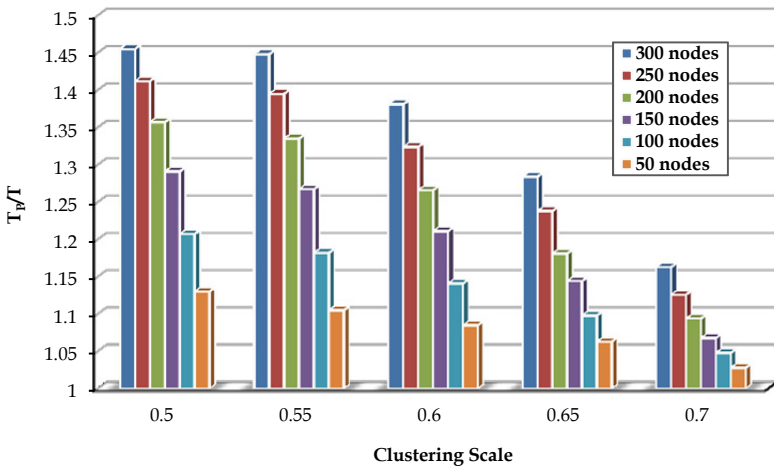


Fig. 8. T_P/T for several node densities and clustering scales.

Consequently, the total amount of averaged consumed energy by nodes for object detection in the coordinated tier during T_P will be:

$$E_p = E + N \cdot P_{sleep} \cdot (T_p - T) \quad (7)$$

From (6) and (7):

$$E_p = E + \frac{\gamma \cdot T \cdot (\mu_{C_{size}} - 1)}{\mu_{C_{size}} - \gamma \cdot (\mu_{C_{size}} - 1)} \cdot N \cdot P_{sleep} \quad (8)$$

So:

$$\frac{E_p}{T_p} = \frac{E \cdot (\mu_{C_{size}} - \gamma \cdot (\mu_{C_{size}} - 1))}{T \cdot \mu_{C_{size}}} + \frac{\gamma \cdot (\mu_{C_{size}} - 1) \cdot N \cdot P_{sleep}}{\mu_{C_{size}}}$$

$$\frac{E_p}{T_p} = \left(1 - \frac{\mu_{C_{size}} - 1}{\mu_{C_{size}}} \cdot \gamma\right) \cdot \frac{E}{T} + \frac{N \cdot \gamma \cdot (\mu_{C_{size}} - 1)}{\mu_{C_{size}}} \cdot P_{sleep} \quad \text{where } (0 < \gamma < 1) \text{ and } (\mu_{C_{size}} > 1)$$

Therefore, the consumed power is:

$$P_p = \lambda \cdot P + \sigma \cdot P_{sleep} \quad (9)$$

where:

$$\lambda = \left(1 - \frac{\mu_{C_{size}} - 1}{\mu_{C_{size}}} \cdot \gamma\right) \quad , \quad 0 < \lambda < 1$$

$$\sigma = \frac{N \cdot \gamma \cdot (\mu_{C_{size}} - 1)}{\mu_{C_{size}}} \quad , \quad 0 < \sigma < \gamma \cdot N$$

Parameter P in Equation (9) is the power consumed in the network with the base un-coordinated mechanism. The consumed power in our scheme (P_p) is reduced by a factor λ with respect to P. The λ factor depends on the average cluster-size and the clustering scale factor. As can be observed from Equation (9) increasing $\mu_{C_{size}}$ produces lower values of λ , and thus a saving in energy with respect to the uncoordinated system. For example a $\mu_{C_{size}}=1.5$ (100 nodes with $\gamma=0.5$) produces a $\lambda = 1 - \gamma/3 = 0.83$ while a $\mu_{C_{size}} = 2.15$ (200 nodes with $\gamma = 0.5$) produces a $\lambda = 1 - 0.53 \gamma = 0.73$. The other term ($\sigma \cdot P_{sleep}$) in Equation (9) is due to the fact of taking nodes to sleep mode in intervals of duration ($T_p > T$) and then nodes sleep $T_p - T$ more time than in the un-clustered scheme.

Figure 9 illustrates the impact of factor λ in Equation (9) in terms of node densities for several clustering scales. From this figure we can see that in high node density tiers, the factor λ is more beneficial since $\mu_{C_{size}}$ is higher and thus there is more potential of cooperation among nodes.

Figure 10 shows the consumed power (P) in the base un-coordinated tier for object detection in four cases of period of duty-cycle for different node densities. The consumed power has been computed for nodes consisting of Cyclops as camera sensor embedded in the host MICA II, similar to the tier 1 in (Kulkarni et al., 2005).

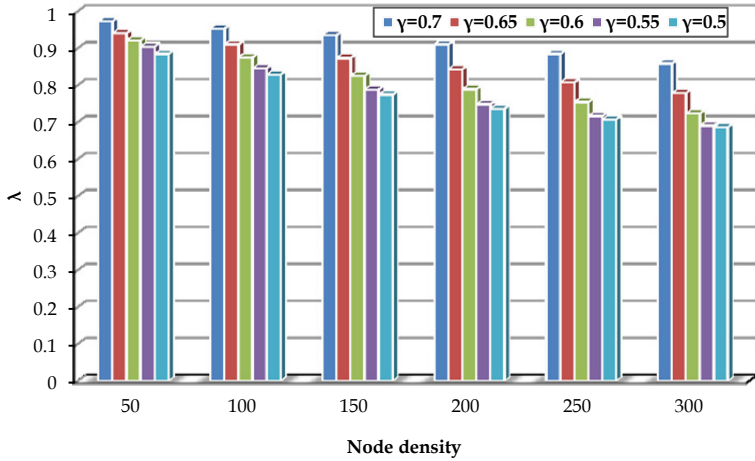


Fig. 9. Factor λ in cooperative scheduling for several clustering scales.

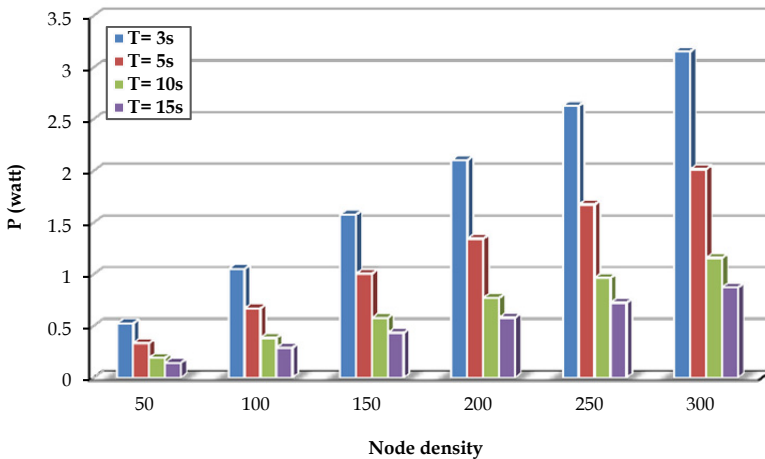
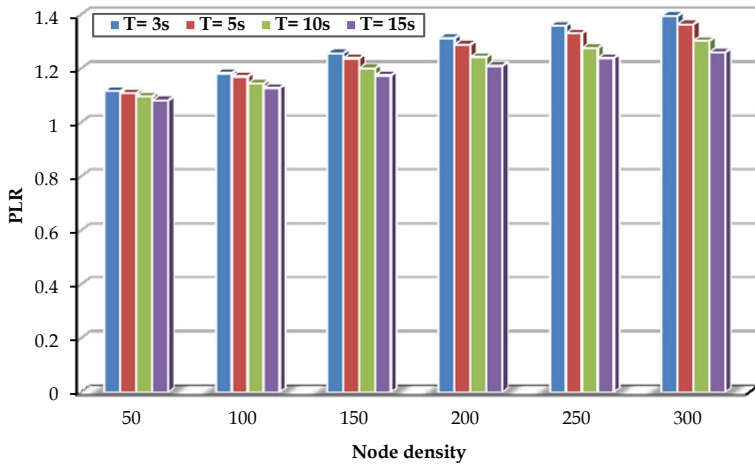


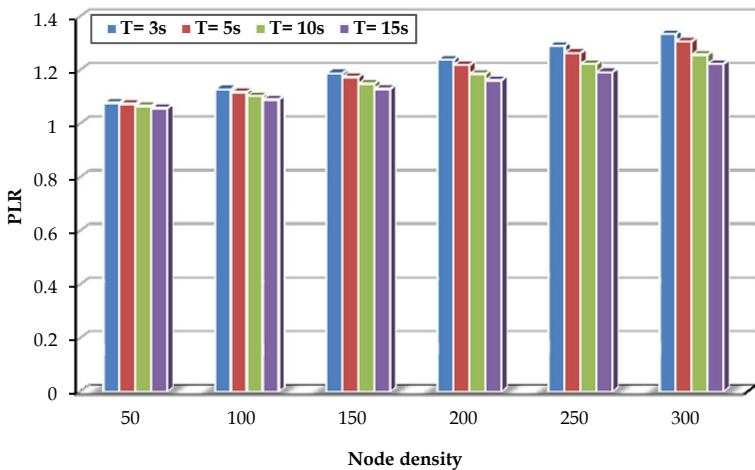
Fig. 10. Consumed power (P) for a non-cooperative tier/network of nodes consisting of Cyclops.

For instance, in the case without coordination, the power consumed in a tier consisting of 200 nodes that performs monitoring with a duty cycle of $T=5$ second, is 1.344 watts. In the coordinated network with the same number of nodes and a clustering scale of 0.5, the power consumed by the network would be reduced by a factor λ of 0.737 (see Figure 9) at the cost of increasing 52.60 mW, ($\sigma \cdot P_{\text{sleep}}$). This means a tier power consumption of $1.344 \cdot 0.737 + 0.0526 = 1.043$ Watts implying a reduction of 22.39%. Thus, in this case, the Prolongation Lifetime Ratio (PLR) would be of $1.344/1.043 = 1.289$. Figure 11.a,b shows the prolongation lifetime ratio assuming a clustering scale of 0.5 and 0.6 for different node densities in four cases of duty-cycle (T). Tiers with high number of nodes have higher capability for cooperation and thus their nodes can conserve considerable amount of energy comparing to

sparse networks and consequently, have longer prolonged lifetime. The figure indicates the more prolongation lifetime for dense tiers.



(a)



(b)

Fig. 11. Prolongation Lifetime Ratio (PLR) for different node densities in the clustered tier with a clustering scale equal to (a) 0.5. (b) 0.6, in four states of base awakening period.

4. Future work

In the clusters established by the depicted mechanism, each cluster member has a common sensing region with the CH. The clusters do not have any intersection and each cluster monitors its covering domain with only intra-cluster collaboration. Clustering with the capability of intersection and cooperation among clusters can increase the scale of efficiency

of monitoring performance and power conservation of cluster members. In a monitoring mechanism utilizing intra and inter cluster cooperation, sensing regions are allocated to intersected clusters thus can be monitored with a higher frequency and/or consuming less amount of energy although the node selection and scheduling procedure will be more complicated. Some initial work has been done in (Alaei & Barcelo, 2010).

5. Conclusion

In this chapter a mechanism for management the wireless multimedia sensor nodes, was described. The mechanism, first, clusters nodes according to their scale of similarity in covering the environment; second, selects and schedules members of established clusters to monitor the sensing region which is divided among clusters. The members of each cluster are scheduled with an exclusive frequency based on the number of members in the cluster and the scale of overlapping among fields of view of the cluster members and thus the monitoring efficiency is increased. Moreover, because of the established intra cluster coordination and collaboration, sensing subsystem of multimedia nodes are optimized to avoid redundant and overlapped sensing. Thus, the capability of energy saving is considerably enhanced with respect to ordinary duty-cycling manners of environment monitoring by WMSNs. On the other hand, optimizing the data sensed by sensing subsystem results in conservation of energy in the transmission and processing subsystems since they meet less amounts of multimedia data to be transmitted and/or processed by the network nodes. Results show how this mechanism prolongs the network lifetime along with a better monitoring performance.

6. Acknowledgment

This work is partially supported by the EuroNF NoE and grants TIN2010-21378-C02-01 and SGR2009-1167.

7. References

- Adriaens, J.; Megerian, S. & Potkonjak, M. (2006). Optimal worst-case coverage of directional field-of-view sensor networks, *Proceedings of the 3rd IEEE Communication Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (IEEE SECON)*, pp. 336-345, ISBN 1-4244-0626-9, Reston, VA USA, September 25-28, 2006
- Akyildiz, I. F.; Melodia, T. & Chowdhury, K. R. (2007). A Survey on Wireless Multimedia Sensor Networks. *Computer Networks*, Vol. 51, Issue 4, (March 2007), pp. 921-960, ISSN 1389-1286
- Alaei, M. & Barcelo-Ordinas, J.M. (2010). A method for clustering and cooperation in Wireless Multimedia Sensor Networks. *Sensors*, Vol. 10, No. 4, (March 2010), pp. 3145-3169, ISSN 1424-8220
- Alaei, M. & Barcelo-Ordinas, J.M. (2010). MCM: multi-cluster-membership approach for FoV-based cluster formation in wireless multimedia sensor networks, *Proceedings of The 6th International Wireless Communications and Mobile Computing Conference (IWCMC 2010)*, pp. 1161-1165, ISBN 978-1-4503-0062-9, Caen, France, June 28-July 2, 2010

- Alippi, C.; Anastasi, G.; Galperti, C.; Mancini, F. & Roveri, M. (2007). Adaptive Sampling for Energy Conservation in Wireless Sensor Networks for Snow Monitoring Applications, *Proceedings of IEEE International Workshop on Mobile Ad-hoc and Sensor Systems for Global and Homeland Security (MASS-GHS07)*, Pisa, Italy, October 8, 2007
- Anastasi, G.; Conti, M.; Francesco, M. & Passarella, A. (2009). Energy conservation in wireless sensor networks: A survey. *Ad Hoc Networks*, Vol. 7, Issue 3, (May 2009), pp. 537-568, ISSN 1570-8705
- Chen, P.; Ahammed, P.; Boyer, C.; Huang, S.; Lin L.; Lobaton, E.; Meingast, M.; Oh, S.; Wang, S.; Yan, P.; Yang, A.Y.; Yeo, C.; Chang, L.C.; Tygar, D. & Sastry, S.S. (2008). CITRIC: A low-bandwidth wireless camera network platform. *Proceedings of the 2nd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC 2008)*, pp. 1-10, ISBN 978-1-4244-2665-2, Palo Alto, CA, USA, September 7-11, 2008
- Dagher, J. C.; Marcellin, M. W. & Neifeld, M. A. (2006). A method for coordinating the distributed transmission of imagery, *IEEE Transactions on Image Processing*, Vol. 15, No. 7, (July 2006), pp. 1705-1717, ISSN 1057-7149
- Diamond, D. (2006). Energy Consumption Issues in Chemo/Biosensing using WSNs, *Energy and Materials: Critical Issues for Wireless Sensor Networks Workshop*, June 30, 2006.
- Ercan, A.; Gamal, A. E. & Guibas, L. (2006). Camera network node selection for target localization in the presence of occlusions, *Proceedings of the ACM SenSys Workshop on Distributed Smart Cameras*, 2006.
- Feng, W.C.; Kaiser, E.; Shea, M.; Feng, W.C & Baillif, L. (2005). Panoptes: scalable low-power video sensor networking technologies. *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 1, Issue 2, (May 2005), pp. 151-167, ISSN 1551-6857
- Kerhet, A.; Magno, M.; Leonardi, F.; Boni, A. & Benini, L. (2007). A low-power wireless video sensor node for distributed object detection. *Journal of Real-Time Image Processing*, Vol. 2, No. 4, October 2007, pp. 331-342, ISSN 1861-8219
- Kulkarni, P.; Ganesan, D.; Shenoy, P. & Lu, Q. (2005). SensEye: A multi tier camera sensor network, *Proceedings of the 13th ACM International Conference on Multimedia (ACM MM 2005)*, pp. 229-238, ISBN 1-59593-044-2, Singapore, November 6-11, 2005
- Margi, C.B.; Lu, X.; Zhang, G.; Stanek, G.; Manduchi, R. & Obraczka, K. (2006). Meerkats: A power-aware, self-managing wireless camera network for wide area monitoring, *Proceedings of International Workshop on Distributed Smart Cameras (DSC 06) in conjunction with SenSys06*, ISBN 1-59593-343-3, Boulder, CO, USA, October 31, 2006
- Pahalawatta, P. V.; Pappas, T. N. & Katsaggelos, A. K. (2004). Optimal sensor selection for video-based target tracking in a wireless sensor network, *Proceedings of the International Conference on Image Processing (ICIP '04)*, pp. 3073- 3076, ISBN 0-7803-8554-3, Singapore, October 24-27, 2004
- Park, J.; Bhat, P. & Kak, A. (2006). A look-up table based approach for solving the camera selection problem in large camera networks, *Proceedings of the International Workshop on Distributed Smart Cameras (DCS '06) in conjunction with SenSys06*, ISBN 1-59593-343-3, Boulder, CO, USA, October 31, 2006
- Pereira, F.; Torres, L.; Guillemot, C.; Ebrahimi, T.; Leonardi, R. & Klomp, S. (2008). Distributed video coding: Selecting the most promising application scenarios. *Signal Processing: Image Communication*, Vol 23, Issue 5, (June 2008), pp. 339-352, ISSN 0923-5965

- Pottie, G. & Kaiser, W. (2000). Wireless Integrated Network Sensors. *Communication of the ACM*, Vol. 43, N. 5, (May 2000), pp. 51-58
- Raghunathan, V.; Ganeriwal, S. & Srivastava, M. (2006). Emerging techniques for long lived wireless sensor networks. *IEEE Communications Magazine*, Vol. 44, Issue 4, (April 2006), pp. 108- 114, ISSN 0163-6804
- Raghunathan, V.; Schurghers, C.; Park, S. & Srivastava, M. (2002). Energy-aware Wireless Microsensor Networks. *IEEE Signal Processing Magazine*, Vol. 19, Issue 2, (March 2002), pp. 40-50, ISSN 1053-5888
- Rahimi, M.; Baer, R.; Iroezi, O.I.; Garcia, J.C.; Warrior, J.; Estrin, D. & Srivastava, M. (2005). Cyclops: in situ image sensing and interpretation in wireless sensor networks, *Proceeding of the 3rd ACM Conference on Embedded Networked Sensor Systems (SenSys 05)*, pp.192–204, ISBN 1-59593-054-X, San Diego, CA, USA, November 2–4, 2005
- Rowe, A.; Rosenberg, C. & Nourbakhsh. I. (2002). A Low Cost Embedded Color Vision System. *Proceedings of the international IEEE/RSJ Conference on Intelligent Robots and Systems (IROS 2002)*, pp. 208-213, ISBN 0-7803-7398-7, Lausanne, Switzerland, Sep.30-Oct.4, 2002
- Schott, B.; Bajura, M.; Czarnaski, J.; Flidr, J.; Tho, T. & Wang, L. (2005). A modular power-aware microsensor with >1000X dynamic power range, *Proceedings of the Fourth International Symposium on Information Processing in Sensor Networks (IPSN 2005)*, pp.469-474, ISBN 0-7803-9201-9, UCLA, Los Angeles, California, USA ,April 25-27 2005
- Soro, S. & Heinzelman, W. (2007). Camera selection in visual sensor networks, *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS '07)*, pp. 81–86, ISBN 978-1-4244-1696-7, London, UK, September 5- 7, 2007.
- Soro, S. & Heinzelman, W. (2005). On the coverage problem in video-based wireless sensor networks. *Proceedings of the 2nd IEEE International Conference on Broadband Communications and Systems (BroadNets)*, pp. 932–939, ISBN 0-7803-9276-0, Boston, MA, USA, October 3–7, 2005
- Tavli, B.; Bicakci, K.; Zilan, R. & Barcelo-Ordinas, J.M. (2011). A Survey of Visual Sensor Platforms. *Journal on Multimedia Tools and Applications*, ISSN 1573-7721, June 2011
- Tezcan, N. & Wang, W. (2008). Self-orienting wireless multimedia sensor networks for occlusion-free viewpoints. *Computer Networks*, vol. 52, issue 13, (September 2008), pp. 2558–2567, ISSN 1389-1286
- Zamora, N. H. & Marculescu, R. (2007). Coordinated distributed power management with video sensor networks: analysis, simulation, and prototyping, *Proceedings of the 1st ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC '07)*, pp. 4–11, ISBN 978-1-4244-1354-6, Vienna, Austria, September 26-28, 2007.
- http://www.ti.com/product/tmp103?DCMP=analog_signalchain_mr&HQS=Other%252bPR%252btmp103-pr

Multimedia Applications for MANETs over Homogeneous and Heterogeneous Mobile Devices

Saleh Ali Alomari and Putra Sumari
*Universiti Sains Malaysia
Malaysia*

1. Introduction

Mobile Ad Hoc Networks (MANETs) are considered a vital part in beyond third generation wireless networks (Nicolaitidis et al., 2003). In the matter of fact, they present a new wireless networking paradigm. Any sort of fixed infrastructure is not used by MANETs. They are important sorts of WLANs, therefore, in a distributed and a cooperative environment, MANETs do efficiently function (Murthy and Mano, 2004) (Sarkar et al., 2008). MANETs are networks of self-creating since there is a lack of routers, configuration prior to the network setup, Access Points (APs) and predetermined topology (Wu et al., 2007). MANETs are as well networks of self-administering and self-organizing. This is because in the network creation process, there is no application for central control. On MANETs, it is extremely hard to apply any of the central administration types, for instance, congestion control due to the dynamic nature of the network topology in MANETs, authentication or central routing. In short, several important applications benefited from MANETs, for example, in military, ubiquitous, emergency and collaboration computing.

In this chapter, describe the necessary background for the MANETs over homogeneous and heterogeneous mobile devices. The researcher begin this chapter to introduce the related background and main concepts of the Mobile Ad Hoc Network (MANETs) in Section 1.2, and explained briefly about the existing wireless mobile network approaches, wireless ad hoc networks, wireless mobile approaches in Section 1.2.2. The characteristic of MANETs are in Section 1.2.3. The types of Mobile Ad hoc network in Section 1.2.4. The traffic types in ad hoc networks which include the Infrastructure wireless LAN and ad hoc wireless LAN are presented in Section 1.2.5. In Section 1.2.6 highlight the relevant details about the ad hoc network routing protocol performance issues. The types of ad hoc protocols such as (Table-driven, On-demand and Hybrid) and Compare between Proactive versus Reactive and Clustering versus Hierarchical are in Section 1.2.7. And Section 1.2.8 respectively. The existing ad hoc protocols are presented in Section 1.2.9. The four important issues significant in MANET are Mobility, QoS Provisioning, Multicasting and Security is presented in Section 1.2.10. Furthermore, the practical application and the MANET layers are shown in Section 1.2.11 and Section 1.2.12 respectively. Finally, in Section 1.2.13 the summary of this chapter.

1.1 Overview of MANETs

The main concept of Wireless Local Area Networks (WLANs) refers to MANETs which are also called either infrastructure-based wireless networks or a single hop network

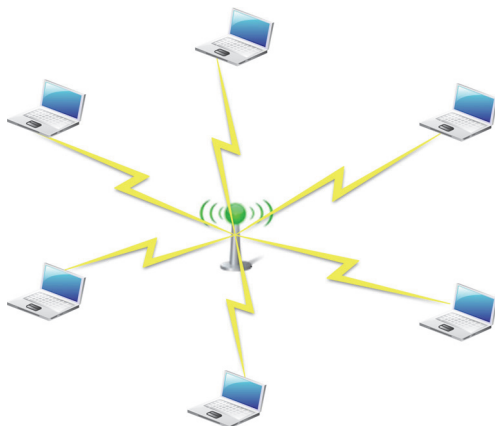


Fig. 1. Illustrates of a single hop WLAN with one AP

(Nicopolitidis et al., 2003) (Murthy and Mano, 2004). Inside a WLAN, the transmission is governed by at least one fixed Access Point (AP) between different mobile nodes. An existing network backbone and the stations contain a bridge as AP functions (Basagni et al., 2004). Both QoS and security issues are efficiently controlled by the AP within a particular network. Inside the network of WLAN, there is no need for different mobile nodes since the AP is the source that does communication through a single hop manner. Wireless network standards are included by the WLAN implementations and developed by Institute of Electrical and Electronics Engineers (IEEE) 802 project (IEEE 802.11, IEEE 802.11b, IEEE 802.11g, IEEE 802.11a, and IEEE 802.11n) and High Performance Radio Local Area Network Type 2 (HiperLAN2). In addition, the European Telecommunications Standardization Institute (ETSI) Broadband Radio Access Networks (BRAN) project (ETSI, 1999) developed the European version of IEEE 802.11a. A frequency of 2.4GHz runs for these standards. However, 5GHz runs for the IEEE 802.11a. For these standards, the transmission rates (bandwidths) are 2 Mbps where as for IEEE 802.11a and IEEE 802.11g, 54 Mbps is run. For IEEE 802.11b, 11 Mbps is run and for IEEE 802.11n, 100 Mbps is run. Note that a single hop WLAN with one AP is shown in Figure 1.

For mobile hosts, a new wireless networking paradigm indicates to a MANET. All sorts of fixed infrastructure are independent to MANET. In order to maintain a connection within the network, nodes (hosts) will rely on each other through a manner that is to be cooperative. Therefore, both computing and ubiquitous communication are considered to be two goals of mobile ad hoc networking. In the matter of fact, both of them are rapidly deployed in such a way they do not rely on a pre-existing infrastructure, for example, Base Station (BS) and Access point (AP) (Perkins et al., 2002). A peer to peer network refers to MANET which has the ability to allow a communication between each wireless client that relies on any infrastructure. MANET can also be defined as a mobile nodes collection of which a highly resource constrained network and a dynamic topology are formed by this collection (Mohapatra and Krishnamurthy, 2005) (Murthy and Mano, 2004). A single hop network refers to WLAN, Major functions within the network are being performed by the cooperation of the nodes. This process represents a multi-hop network that refers to the MANET. There are such problems entitled in MANETs. These comprise; security, QoS, routing and energy conversation. These problems came due to several reasons: high mobility, resource constrains such as power, storage, and bandwidth (Negi and Rajeswaran, 2004), its cooperative nature

and the dynamic topology of nodes operating in MANET's environment. In Defence Advanced Research Projects Agency (DARPA) Packet Radio projects (Jubin et al., 1987), ad hoc networking was initiated for military applications, specifically, for dynamic wireless networks since 1970s. Accordingly, this networking is not considered to be as a new concept. For MANET, a new networking group was formed within the Internet Engineering Task Force (IETF-manet) so that the standard Internet routing support could be developed for mobile IP autonomous segments. In addition, a framework for IP-based protocols in MANET will be developed as well. In the fields of mobile IP-based networks and wireless internet, the increasing improvement in the recent IEEE standards of 802 projects for wireless networks (Broch.J et al., 1998) has raised up. A MANET can be either heterogeneous or homogeneous depending on the type of mobile nodes being involved. When all mobile nodes are of the same type of a MANET, this is called a homogeneous MANET, whereas when different type of mobile nodes are involved, this is otherwise called a heterogeneous MANET. The homogeneous and heterogeneous mobile ad hoc network are shown in In Figure 2 and Figure 3 respectively. The same family of IEEE 802.11 standards is being used by MANETs. More

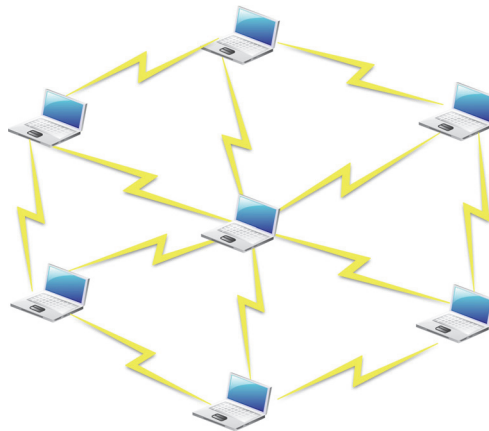


Fig. 2. Illustrates the homogeneous mobile ad hoc network

specifically, in Bluetooth and WLANs, these standards are being used (Morinaga et al., 2002). Table 1 shows a comparison between WLAN and MANET.

1.2 Mobil Ad Hoc Network

With the widespread rapid development of computers and the wireless communication, the mobile computing has already become the field of computer communications in high-profile link. MANET (Sarkar et al., 2008) is a completely wireless connectivity through the nodes constructed by the actions of the network, which usually has a dynamic shape and a limited bandwidth and other features, network members may be inside the laptop, Personal Digital Assistant (PDA), mobile phones and so on. On the Internet, the original mobility is the term used to denote actions hosts roaming in a different domain; they can retain their own fixed IP address, without need to constantly changing, which is Mobile IP technology.

Mobile IP nodes in the main action is to deal with IP address management, by home users and foreign users to the mobile node to packet tunneling, the routing and fixed networks are not different from the original. However, ad hoc network to be provided by mobility is a fully wireless, can be any mobile network infrastructure, without a base station, all

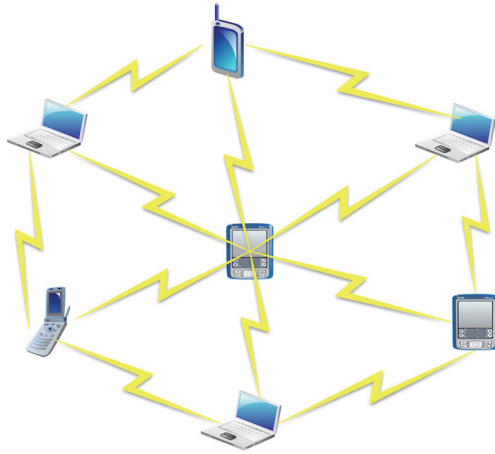


Fig. 3. Illustrates the heterogeneous mobile ad hoc network

Comparison Aspect	(WLAN)	(MANET)
Autonomous terminals	No	Yes
Self-configuration	No	Yes
Mobile host/router	No	Yes
Bandwidth constrained network	No	Yes
Infrastructure-based	Yes (APs/routers/Servers)	No
Power awareness	Does not matter	Yes
Security policy	Centralized	Distributed
Centralized/distributed operation	Centralized	Distributed
Routing	Easy	A bit difficult
Scalability	Easy	A bit difficult
Multicasting	Easy	A bit difficult
Static/ dynamic topology	Static	Dynamic
QoS guarantee	Can be guaranteed easily	A bit difficult
Typical applications	Home, enterprise network	Military/emergency
Single hope / multi hope	Single	multi
Communication mechanism	Base station type access	P2P

Table 1. Illustrates the comparison between WLAN and MANET

the nodes can contact each other at the same time take router work with the Mobile IP completely different levels of mobility. Early use of the military on the Mobile Packet Radio Networked (MPRN) in fact can be considered the predecessor of MANET, when the high-tech communication equipment, the size, weight continuously decreases, power consumption is getting low, Personal Communication System (Personal Communication System, PCs) concept evolved, from the past few years the rapid popularization of mobile phones can be seen to communicate with others at anytime and anywhere, get the latest information, or exchange the required information is no longer a dream. And we have gradually become an integral part of life. Military purposes, as is often considerable danger in field environment, some of the major basic communication facilities, such as base stations, may not be available,

in this case, different units, or if they want to communicate between the forces, they must rely on MANET networks infrastructure. In emergency relief, the mountain search and rescue operations at sea, or even have any infrastructure can not be expected to comply with the topographical constraints and the pressure of time under the pressure, ad hoc network completely wireless and can be any mobile feature is especially suited to disaster relief operations when personal communication devices and more powerful, some assembly occasions, if need to exchange large amounts of data, whether the transmission of computer files or applications that display. if can connect with a temporary network structure, then the data transmission will be more efficient without the need for large-scale projection equipment would not have point to point link equipment such as network line or transmission line. The current wireless LAN technology, Bluetooth is has attracted considerable attention as a development plan. Bluetooth's goal is to enable wireless devices to contact with each other, if sentence formation adding the design MANET.

1.2.1 History of Ad Hoc Network

Nowadays, the information technology will be mainly based on wireless technology, the conventional mobile network and cellular are still, in some sense, limited by their need for infrastructure for instance based station, routers and so on. For the MANET, this final limitation is eliminated. The ad hoc network are the key in the evolution of wireless network and the ad hoc network are typically composed of equal node which communication over wireless link without any central control. Although military tactical communication is still considered as the primary application for MANET and commercial interest in this type of networks continues to grow. And all the applications such as rescue mission in time of natural disasters, law enforcement operation, and commercial as rescue and in the sensor network are few commercial examples, but in this time it's become very important in our life and they become use it.

The MANET application is not new one and the original can be traced back to the Defence Advanced Research Projects Agency (DARPA), Packet Radio Networking (PRNET) project in 1972 (Freebersyser and Leiner, 2001, Jubin and Tornow, 1987) which evolved into the survivable adaptive radio networks (SURAN) program. Which was primarily inspired by the efficiency of the packet switching technology for instance the store/forward routing and then bandwidth sharing, it's possible application in the MANET environments. As well commercial rescue in the PRNET devises like repeaters and routers and so on, were all mobile although mobility was so limited in that time, theses advanced protocol was consider good in the 1970s. After few years advance in micro electronics technology and it's was possible to integrate all the nodes and also the network devices into a single unit called ad hoc nodes, and then the advance such as the flexibility, resilience also mobility and independence of fixed infrastructure, and in that time they so interesting to use it immediately among military battlefield, Ad hoc networks have played an important role in military applications and related research efforts. For example, the global mobile information systems (GloMo) simulator (Leiner et al.), the near-term digital radio (NTDR) program and also has been the increase in the police, commercial sector and rescue agencies in use of such networks under disorganized environments. Ad hoc network research stayed long time in the realm of the military. And in the middle of 1990s with advice of commercial radio technology and the wireless became aware of the great advantages of MANET outside the military battlefield domain, and then became so active research work on ad hoc network start in 1995 in the conference session of the Internet Engineering Task Force (IETF) (IETF-MANET). And then in 1996 this works had evolved into MANET, in that time focused to discussion centered in military satellite network, wearable computer network and tactical network with

specific concerns begin raised relative to adaptation of existing routing protocols to support IP network in dynamic environments, as well as they make the charter of the MANET Working Group (MANETWG) of the Internet Engineering Task Force (IETF) also the work inside the MANETs relies on other existing IETF standard such as Mobile IP and IP addressing. Most of the currently available solutions are not designed to scale to more than a few hundred nodes. Currently, the research in MANET became so active and vibrant area and the efforts this research community together with the current and future (MANET) enabling radio technology.

Recently, the Ad Hoc Wireless Network and computing consortium was established with the aim to coalescing the interests and efforts to use it anywhere such as academic area and industry and so on. And in order to apply this technology to application ranging for the Home Wireless (HW) to wide area peer to remote networking and communications. And it does will certainly pave the way for commercially viable MANETs and their new and exciting applications, which began to appear in all fields in this life. More recently, the computer has become spread significantly in the all the place and after a pervasive computing environment can be expected based on the recent progresses and advances in computing and communication technologies. Next generation of mobile communications will include both prestigious infrastructure wireless networks and novel infrastructureless MANETs.

1.2.2 Wireless Ad Hoc Networks

MANET is a collection of two or more devices or terminals with wireless communications and networking capability that communicate with each other without the aid of any centralized administrator also the wireless nodes that can dynamically form a network to exchange information without using any existing fixed network infrastructure. And it's an autonomous system in which mobile hosts connected by wireless links are free to be dynamically and some time act as routers at the same time. All nodes in a wireless ad hoc network act as a router and host as well as the network topology is in dynamically, because the connectivity between the nodes may vary with time due to some of the node departures and new node arrivals. The special features of MANET bring this technology great opportunity together with high challenges. All the nodes or devices responsible to organize themselves dynamically to communication between each other and to provide the necessary network functionality in the absence of fixed infrastructure or can call it ventral administration. It implies that maintenance, routing and management, etc, have to be done between all the nodes. This case called peer level multi hopping and that is the main building block for ad hoc network. In the end, conclude that the ad hoc nodes or devices are difficult and more complex than other wireless networks. Therefore, ad hoc networks form sort of clusters to the effective implementation of such a complex process. In Figure 4 shows some nodes forming ad hoc networks, and there are some nodes more randomly in different directions and different speeds.

In the past few years, the people became realized to use all the technology so widely and the people's future living environments are emerging, based on information resource provided by the connections of different communication networks for clients also have seen a rapid expansion in the field of mobile computing because the proliferation not expensive, widely available wireless devices. A new small devices such as personal communication like cell phones, laptops, Personal Digital Assistants (PDAs), handhelds, and there are a lot of traditional home appliances such as a digital cameras, cooking ovens, washing machines, refrigerators and thermostats, with computing and communicating powers attached. Expand this area to become a fully pervasive and so widely. With all of this, the technologies must be

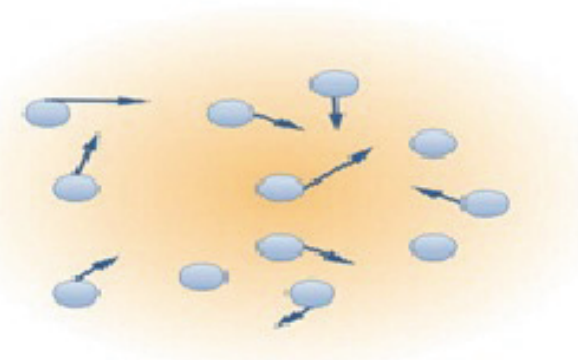


Fig. 4. Illustrates some of the nodes moves randomly in different direction and different speeds

formed the good and new standard of pervasive computing, that including the new standards, new tools, services, devices, protocols and a new architectures. As well as the people in this time, or the users of internet users in ad hoc network through increase in the use of its advantage is that not involve any connection link and the wiring needed to save space, and building low cost, and improve the use, and can be used in mobile phone, because of these advantage local wireless network architecture readily. And beads in these advantages the wireless network can be used in the local area network terminal part of the wireless (Liu and Chang, 2009).

1.2.2.1 Wireless mobile network approaches

The past decade, the mobile network is the only one much important computational techniques to support computing and widespread, also advances in both software techniques and the hardware techniques have resulted in mobile hosts and wireless networking common and miscellaneous. Now will discuss about to distinct approaches very important to enabling mobile wireless network or IEEE 802.11 to make a communication between each other (part-11, 1997) (part-12, 1999). Firstly infrastructure wireless networks and secondly, infrastructureless wireless networks (ad hoc networks) and will clarify both in bottom.

1.2.2.2 Infrastructure wireless networks

In this architecture that allow the wireless station to make a communication between each other through the Base Station (BS) as shown in Figure 5, and that will handover the offered traffic from the station to another, the same entity will regulate or organize the allocation of radio resources. When a source node likes to communicate with a destination node, the former notifies the BS. At this point, the communicating nodes do not need to know anything about the route from one to another. All that matters is that the both source and destination nodes are within the transmission range for the BS and then if there is any one loses this condition, the communication will frustration or abort.

1.2.2.3 Infrastructureless wireless networks

The mobile wireless network is known as Mobile Ad Hoc Network (MANET). As has been previously defined in the bidder is a collection of two or more devices or nodes or

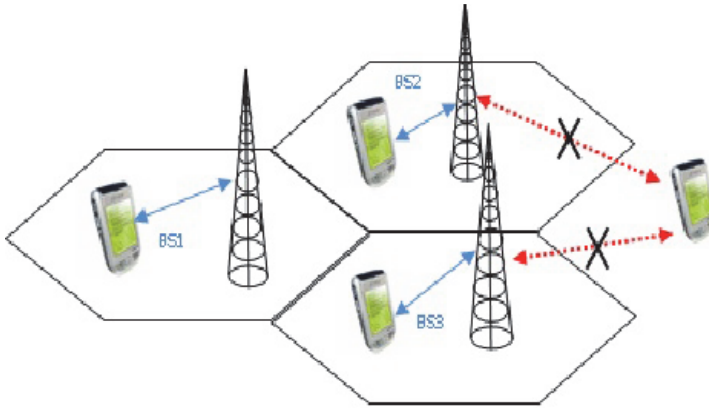


Fig. 5. Illustrates of the infrastructure network

terminals with wireless communications and networking capability that communicate with each other without the aid of any centralized administrator also the wireless nodes that can dynamically form a network to exchange information without using any existing fixed network infrastructure. And it's an autonomous system in which mobile hosts connected by wireless links are free to be dynamically and some time act as routers at the same time (Frodigh et al., 2000). The infrastructureless is important approaches in this technique to communication technology that supports truly pervasive computing widely due to there is a lot of context information need to exchange between mobile nodes but can not rely on the fixed network infrastructure, but in this time the communication wireless became develops very fast (IETF-manet). In Figure 6 shown a small example for the ad hoc networks, to explain how mobile ad hoc network working.

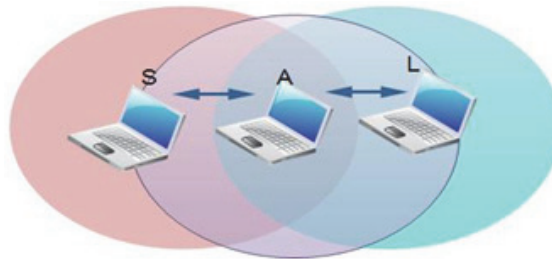


Fig. 6. Illustration of the infrastructureless networks (ad hoc networks)

This figure illustrates the modus operandi of ad hoc networks, there are a three ad hoc network nodes (S, A, L), the source node (S) need to make a communication with the destination node (L) and both of them (S, L) not in the same transmission range of each others, here both they must use the node (A) to send and receive or forewords the packets from source to the destination that means from node to another node. (R) is a node work as host and router in the same time. Additionally, the definition for the router is an entity that determines the path to be used, in order to forward a packet towards the last destination, and then the router chooses the next node to which a packet should be forwarded according to its current understanding of the state of the network.

1.2.3 Characteristics of MANET

Request For Comments (RFC) 2501 document (IETF, 1999) which is published by MANET working group within the IETF describes the main characteristics of MANET which differs from the characteristics of traditional wireless local area networks such as WLANs due to the dynamic and the infrastructureless natures of MANETs (Hekmat, 2006). According to the IETF RFC 2501, MANET has characteristics can be divided into the following:

1. A collection of autonomous terminals means that within a MANET, each mobile node performs its tasks as a router and a host.
2. It contains a dynamic topology which means there are a group of nodes into it that are moving and resulting to a random change rapidly at unpredictable times through the network topology.
3. A distributed operation is contained into it which means that the network's management and control is spread (distributed) in the nodes because of the infrastructure types' absence of which the central control of the network operations is supported. In a MANET, nodes must perform with each other. each node behaves as a router and a host simultaneously in order to have the network functions efficiently implemented, for example, routing and security.
4. It can be deployed as fast as it could be.
5. Pre-existing infrastructure is independent from it.
6. Bandwidth-constrained, variable capacity links compared with the wired network environment, the capacity of the wireless link itself is relatively small, but also susceptible to external noise, interference, and signal attenuation effects.
7. Self-adapts to the propagation patterns and connectivity.
8. Adapts to mobility patterns and traffic.
9. A limited physical security is contained into it, for example, in the absence of any centralized encryption or authentication. In order to reduce security threats, existing techniques of link security are at most applied into the WLANs and the wired networks.
10. It has an energy constrained operation a laptop or handheld computers are often used batteries to provide power, how to save electricity in the context of depletion of system design is also necessary to consider the point.

Mobile networking and MANETs are considered to be of good candidates due to many reasons: its simplicity for usage, robustness, speedy deployment and low cost. Its disadvantages comprise the complexity of routing due to the consistent move of nodes, mobility and dynamic topology, vulnerability of security due to the cooperation principle in MANETs, and the low computing power due to small devices used in MANETs.

1.2.4 Types of mobile ad hoc network

The wireless ad hoc network divided into three main types. Firstly, the quasi-static ad hoc network the nodes may be portable or static, because the power controls and link failures, the resulting network topology may be so active. The sensor network is an example for the quasi-static ad hoc network (Estrin et al., 1999). Secondly, the MANET the entire network may be mobile and the nodes may move fast relative to each other. thirdly, Vehicular Ad Hoc Networks (VANETs) are a kind of network useful for offering traffic information interchange in a collaborative way between vehicles.

1.2.4.1 Mobile Ad Hoc Networking (MANET)

MANET is a group of independent network mobile devices that are connected over various wireless links. It is relatively working on a constrained bandwidth. The network topologies are dynamic and may vary from time to time. Each device must act as a router for transferring any traffic among each other. This network can operate by itself or incorporate into large area network (LAN).

There are three types of MANET. It includes Vehicular Ad hoc Networks (VANETs), Intelligent Vehicular Ad hoc Networks (InVANETs) and Internet Based Mobile Ad hoc Networks (iMANET). The set of application for MANETs can be ranged from small, static networks that are limited by power sources, to large-scale, mobile, highly dynamic networks. On top of that, the design of network protocols for these types of networks is face with multifaceted issue. Apart from of the application, MANET need well-organized distributed algorithms to determine network organization, link scheduling, and routing. Conventional routing will not work in this distributed environment because this network topology can change at any point of time. Therefore, we need some sophisticated routing algorithms that take into consideration this important issue (mobile network topology) into account. While the shortest path (based on a given cost function) from a source to a destination in a static network is usually the optimal route, this idea is not easily far-reaching to MANET. Some of the factors that have become the core issues in routing include variable wireless link quality, propagation path loss, fading, interference; power consumed, and network topological changes. This kind of condition is being provoked in a military environment because, beside these issues in routing, we also need to guarantee assets security, latency, reliability, protection against intentional jamming, and recovery from failure. Failing to abide to of any of these requirements may downgrade the performance and the dependability of the network.

1.2.4.2 Mobile ad hoc sensor network

A mobile ad hoc sensor network follows a broader sequence of operational, and needs a less complex setup procedure compared to typical sensor networks, which communicate directly with the centralized controller. A mobile ad hoc sensor or hybrid ad hoc network includes a number of sensor spreads in a large geographical area. Each sensor is proficient in handling mobile communication and has some level of intelligence to process signals and to transmit data. In order to support routed communications between two mobile nodes, the routing protocol determines the node connectivity and routes packets accordingly. This condition has makes a mobile ad hoc sensor network highly flexible so that it can be deployed in almost all environments (Bakht, 2010). The wireless ad hoc sensor networks (Asif, 2009) are now getting in style to researchers. This is due to the new features of these networks were either unknown or at least not systematized in the past. There are many benefits of this network, it includes:

- Use to build a large-scale networks
- Implementing sophisticated protocols
- Reduce the amount of communication (wireless) required to perform tasks by distributed and/or local precipitations.
- Implementation of complex power saving modes of operation depending on the environment and the state of the network.

With the above-mentioned advances in sensor network technology, functional applications of wireless sensor networks increasingly continue to surface. Examples include the replacement

of existing detecting scheme for forest fires around the world. Using sensor networks, the detecting time can be reduced significantly. Secondly is the application in the large buildings that at present use various environmental sensors and complex control system to execute the wired sensor networks. In a mobile ad hoc sensor networks, each host may be equipped with a variety of sensors that can be organized to detect different local events. Besides, an ad hoc sensor network requires a low setup and administration costs (Akkaya and Younis, 2005) (Akyildiz et al., 2002).

1.2.4.3 Vehicular Ad Hoc Networks (VANETs)

Vehicular Ad Hoc Networks (VANETs) (Kosch et al., 2006) are a kind of network useful for offering traffic information interchange in a collaborative way between vehicles. They are foreseen to be a great revolution in the driving, providing new services such as Road safety, traffic management, Pollution reduction, Cost reduction in the vehicle security incorporation and public transport.

1.2.5 The traffic types in the ad hoc networks

The traffic types in the ad hoc networks are so different from the infrastructure wireless network. The traffic types are classified into three types (peer to peer, remote to remote and dynamic traffic) (Mbarushimana and Shahrabi, 2008). Firstly, peer to peer is a communication between two nodes in the same area, that means which are within one hop. Network traffic (in bits per second) is usually fixed. Secondly, remote to remote is a communication between two nodes beyond a single hop, but maintain a stable route between them. This may be the result of a number of nodes, to stay within the range of each other in one area or may move as a group. Movement it's a similar to the standard network traffic. Finally, dynamic traffic it will happen when the nodes are moving dynamically around and then the routers must be reconstructed. This results in a poor connectivity and network activity in short bursts. For example in IEEE 802.11 network and the basic structure divided into two types firstly infrastructures wireless LAN, the second structure ad hoc wireless LAN.

1.2.5.1 Infrastructure wireless LAN

In this kind of network as shown in the Figure 7, the network in any architecture will be an access point; its function is one or more of the wireless local area network and the existing cable network systems to link, so that stations within the wireless local area network and external nodes can connect with each other. It is characterized by a fixed and pre-positioning a good base station location, the static backbone network topology, a good environment and a stable connection, the base station that is doing a good job when you set up detailed plans (Li, 2006).

1.2.5.2 Ad hoc wireless LAN

The ad hoc wireless LAN is an infrastructures relies on infrastructures wireless local area network, which only targeted at local area network within the framework of each machine is able to be linked up into networks, regardless of whether the communication with the outside world, then such a structure, either one or two users can communicate directly with each other, and this structure is composed of at least composed of two or more workstations. Is characterized by no fixed base stations, network will be rapidly changing; dynamic network topology is vulnerable to interference, to automatically form a network without infrastructure

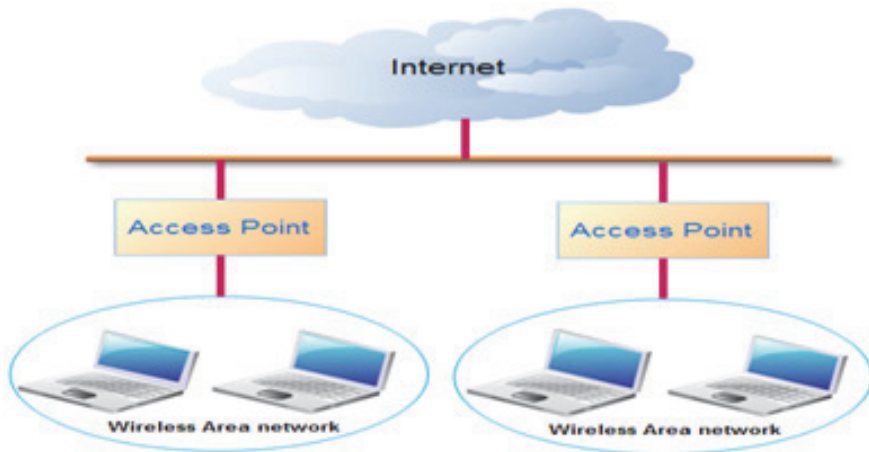


Fig. 7. Illustrates of the infrastructure wireless LAN architecture.

and adapt to topology changes. For more explain shows the Figure 8 for Ad Hoc Wireless network.



Fig. 8. Illustrates of the wireless ad hoc network

1.2.6 Ad hoc network routing protocol performance issues

The MANET with the traditional wired, fixed networks have many different characteristics, so to design a suitable routing protocol for MANET operating environment must also consider the different directions, the following sub-qualitative and quantitative aspects of the discussion:

1. On the qualitative aspects, can be divided into:
 - a) Distribution operation: Due to the existence of MANET where there is no prerequisite for the construction of the underlying network, so routing can not rely on a particular node to operate.

- b) Loop-freedom: All the routing protocol should be consistent with the characteristics; we must ensure the normal work in order to avoid waste of bandwidth.
 - c) Demand-based operation: In order to reduce the burden on each node, if the link is not so much the demand should be considered when using On-demand approach to the establishment of the path, and only when the need for a particular path query, the establishment of the path.
 - d) Proactive operation: With the On-demand concept of the contrary, if the network resources fairly adequate, proactive table-driven approach could speed up the path to the establishment of speed.
 - e) Security: Because it is the wireless environment, to how to ensure the security of the connection can not be ignored will be part of network security is also a MANET from theory to implementation of the key challenges.
 - f) Sleep Period operation: As the MANET nodes are generally smaller wireless devices, using the battery as a power supply, how to save power consumption, or for no work, the node goes into sleep mode, can operate more smoothly so that MANET. Also the nodes of a MANETs may stop transmitting or receiving or both, also even receiving requires power for arbitrary time periods and the routing protocol should be able to accommodate such sleep periods without overly adverse consequences. This property may require close coupling with the link-layer protocol through a standardized interface.
2. On the quantity, can be divided into:
- a) End-to-end data throughput and delay: Data transmission rate and delay in the case that every routing protocol must take into account the focus should be how to find the best path? Is the maximum bandwidth or minimum latency, or the link to the most stable? Considered more likely to make more complicated routing protocol, but it is possible to significantly improve the transmission quality.
 - b) Route Acquisition time: While the table-driven generally higher than on-demand performance good, but many of the former to pay the price, which, if properly designed, for example, there is more commonly used in the path cache, or a certain fixed path, can improve the path to the establishment of time.
 - c) Percentage Out-of-order delivery: Real-time data for this part of the more stringent requirements, and general information will not affect how and upper TCP cooperation is also IP routing work.
 - d) Efficiency: The simplest method, the smallest control overhead done the most complete, most powerful feature is a common goal for all routing protocol.

1.2.7 Types of ad hoc protocols

Ad hoc network routing protocols is divided to three type of routing protocols, which that depending on a different of routing protocols (Saleh Al-Omari and Putra Sumari, 2010).

1.2.7.1 Oriented routing table (table-driven)

The oriented routing table is an active routing environment in which the intervals between the wireless nodes will send medical information with more paths. Each wireless node is on the basis of information gathered recently to change its route table, when the network topology change makes the original path is invalid, or the establishment of any new path, all nodes will receive updates on the status path. The path will be continuously updated, so that the node in time of peace on its own routing tables is ready, and immediately available

when needed. However, such agreements must be periodically to broadcast messages, so a considerable waste of wireless bandwidth and wireless node power, but if you want to reduce the broadcast bandwidth consumption caused by a large number, we should lengthen the interval between each broadcast time, which in turn will result in the path table does not accurately reflect network topology changes.

1.2.7.2 Demand-driven (on-demand)

In the demand driven, When needed to send packets only it began to prepare to send the routing table. When a wireless node needs to send data to another wireless node, the source client node will call a path discovery process, and stored in the registers of this path. The path is not valid until the expiration or the occurrence of conditions of the agreement with the first phase of a ratio of such agreements in each node. A smaller amount of data needed, and do not need to save the entire network environment and the routing information. The main benefit of this agreement is that the use of a lower bandwidth, but the drawback is that not every wireless node that sends packets can always quickly find the path. The path discovery procedure can cause delays and the average delay time is longer (Liu and Chang, 2009).

1.2.7.3 Hybrid

Hybrid is an improvement of the above mentioned two, or the combination of other equipment, such as Global Positioning System (GPS) and other equipment, participate in the study of mechanisms to facilitate the routing of the quick search, and data transmission (Pandey et al., 2005) (Johnson and Maltz, 1999.). However, there are already more than 13 kinds of the above routing protocol have been proposed, following the more representative for several separate presentations, and to compare their individual differences lie.

1.2.8 Compare between proactive versus reactive and clustering versus hierarchical

1.2.8.1 Proactive versus Reactive Approaches

Ad hoc routing protocols can be classified into two types; proactive and On-Demand (reactive) base on each own strategy (Perkins, 2001). Proactive protocols demand nodes in a wireless ad hoc network to keep track of routes to all possible destinations. This is important because, whenever a packet requests to be forwarded, the route is beforehand identified and can be used straight away. Whenever there's modification in the topology, it will be disseminated throughout the entire network. Instances include "destination-sequenced distance-vector" (DSDV) routing (Perkins and Bhagwat, 1994), "wireless routing protocol" (WRP) (Murthy and Garcia-Luna-Aceves, 1996), "global state routing" (GSR) (Chen and Gerla, 1998), and "fisheye state routing" (FSR) (Iwata et al., 2002) and in next section will discuss about everyone.

On-demand (reactive) protocols will build the routes when required by the source node, in order for the network topology to be detected as needed (on-demand). When a node needs to send packets to several destinations but has no routes to the destination, it will start a route detection process within the network. When a route is recognized, it will be sustained by a route maintenance procedure until the destination becomes unreachable or till the route is not wanted anymore. Instances include "ad hoc on-demand distance vector routing" (AODV) (Perkins et al., 2003), "dynamic source routing" (DSR) (J. Broch et al., 2004), and "Cluster Based Routing protocol" (CBRP) (Jiang et al., 1999). Proactive protocols comprise the benefit that new communications with arbitrary destinations experience minimal delay, but experience the disadvantage of the extra control overhead to update routing information at all nodes. To overcome with this limitation, reactive protocols take on the opposite method by tracking down route to a destination only when required. Reactive protocols regularly utilize less

bandwidth compared to proactive protocols, however it is a time consuming process for any route tracking activity to a destination proceeding to the authentic communication. Whenever reactive routing protocols must relay route requests, it will create unnecessary traffic if route discovery is required regularly.

1.2.8.2 Clustering versus hierarchical approaches

Scalability is one of the major tribulations in ad hoc networking. The term scalability in ad hoc networks can be defined as the network's capability to provide an acceptable level of service to packets even in the presence of a great number of nodes in the network. If the number of nodes in the network multiply for proactive routing protocols, the number of topology control messages will increase nonlinearly and it will use up a large fraction of the available bandwidth. While in reactive routing protocols, if there are a large numbers of route requests propagated to the entire network, it may eventually become packet broadcast storms. Normally, whenever the network size expands beyond certain thresholds, the computation and storage requirements become infeasible. At a time whenever mobility is being taken into consideration, the regularity of routing information updates may be extensively enhanced, and will deteriorate the scalability issues. In order to overcome these obstacles and to generate scalable and resourceful solutions, the solution is to use hierarchical routing. Wireless hierarchical routing is based on the idea of systematizing nodes in groups and then assigns the nodes with different task within and outside a group. Both the routing table size and update packet size are decreased by comprising only a fraction of the network. For reactive protocols, restricting the scope of route request broadcasts can assist in improving the competency. The best method of building hierarchy is to gather all nodes geographically near to each other into groups. Every cluster has a principal node (cluster head) that corresponds with other nodes. Instances of hierarchical ad hoc routing protocols include "zone routing protocol" (ZRP) (Haas and Pearlman, 2000).

1.2.9 Existing ad hoc protocols

In the ad hoc network there are more than 13 kinds of the above routing protocol have been proposed such as DSDV, GSR, CGSR, WRP, FSR, AODV, DSR, TORA, CBRP, ABR, SSR, CEDAR and ZRP, for more dilates about existing ad hoc network protocols (Saleh Alomari and Putra Sumari, 2010). Further explanation for understanding some of the existing mobile ad hoc network are provided in Appendix A figure 10. The comparison between Table Driven, Demand Driven and Hybrid are shown in Table 2, and then show in Table 3 the Table Driven for three kind of protocols such as WRP, CGSR, DSDV and comparison between them, Demand Driven (On-Demand) with six type of protocols such as TORA, DSR, AODV, ABR, CEDAR and SSR and comparison between them shows in Table 4. Finally, shows compare the main characteristics of existing multipath routing protocols in Table 5.

* CEDAR, TORA itself, although it can not also be used in multicasting, but there have been constructed in the two above the multicast routing protocol was proposed.

1.2.10 Challenges and issues of MANETs

For ad hoc networking design and implementation, there lots of factors and challenges which are:

Scalability: in some applications, a MANET can grow to thousands of nodes, such as, battlefield deployments, urban vehicle grids and large environmental sensor fabrics. It is extremely hard to have the scalability handled in a MANET due to the random and unlimited mobility (Perkins et al., 2002).

	Table Driven(Proactive)	Demand Driven(Reactive)	Hybrid
Routing Protocols	DSDV,CGSR,WRP	AODV,DSR,TORA,ABR,SSR	ZRP
Route acquisition delay	Lower	Higher	Lower for Intra-zone; Higher for Inter-zone
Control overhead	High	Low	Medium
Power requirement	High	Low	Medium
Bandwidth requirement	High	Low	Medium

Table 2. Illustrates the comparison between Table Driven, Demand Driven and Hybrid

Table Driven	CGSR	WRP	DSDV
Routing philosophy	Hierarchical	Flat	Flat
Loop-free	Yes	Yes, but not instantaneous	Yes
Number of required tables	2	4	2
Frequency of update transmissions	Periodically	Periodically and as needed	Periodically and as needed
Updates transmitted to	Neighbors and cluster head	Neighbors	Neighbors
Utilize hello messages	No	Yes	Yes
Critical nodes	Cluster head	No	No
Communication complexity	$O(x = N)$	$O(x = N)$	$O(x = N)$

Table 3. Shows the Table-Driven for the three kinds of protocols and comparison between them

Mobility is at most the first designer’s enemy of MANET (Murthy and Mano, 2004).

Energy conservation most ad hoc nodes, such as Personal Digital Assistants (PDAs), sensors and Laptops are often power supplied using batteries which have limited power. Therefore, for MANET, energy conservation is considered to be an enormous challenge.

Application/Market penetration: multi-hop technology is not commercial at present. More clearly, the short coverage area’s limitation of the wireless products can be justified in its belonging to the standard of IEEE 802.11.

Design/Implementation: manageable, secure, reliable and survivable implementation and design must act for MANET since a bandwidth-constrained operation and a limited physical security are contained in MANETs.

Limited wireless transmission range depends on the wireless technology’s capabilities.

Operational/Business-related how to have the network managed and how to bill for services.

The main key issues that affect the design, deployment, and performance of an ad hoc wireless system are summarized as following: scalability, security, energy management, QoS provisioning, deployment considerations, self organization, multicasting, pricing scheme, medium access scheme, routing, transport layer protocols, addressing and service discovery.

On-Demand	TORA	DSR	AODV	ABR	CEDAR	SSR
Overall complexity	High	Medium	Medium	High	High	High
Overhead	Medium	Medium	Low	High	High	High
Routing philosophy	Flat	Flat	Flat	Flat	Core-Extracted	Flat
Loop Free	Yes	Yes	P	Yes	Yes	Yes
Multicast capability	No*	No	Yes	No	No*	No
Beaconing requirements	No	No	No	Yes	Yes	Yes
Multiple route support	Yes	Yes	No	No	No	No
Routes maintained in	Route table	Route cache	Route table	Route table	Route table	Route table
Route reconfiguration methodology	Link reversal	Erase route	Erase route	Localized broadcast query	Dynamic route re-compute	Erase route

Table 4. Shows the Demand Driven (On-Demand) with six types of protocols and comparison between them

	AODV	DSR	CBRP	DSDV	WRP	GSR	FSR
Routing Category	Reactive	Reactive	Reactive	Proactive	Proactive	Proactive	Proactive
TTL Limitation	Yes	Yes	Yes	No	No	No	No
Flood Control	No	No	No	Yes	Yes	Yes	Yes
QoS Support	Yes	Yes	P	Yes	Yes	Yes	Yes
Periodic Update	No	No	No	No	No	No	No
Power Management	No	No	No	No	No	No	No
Multicast Support	Yes	No	No	No	No	No	No
Beaconing	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Security Support	No	No	No	No	No	No	No

Table 5. Shows the comparison of the main characteristics of existing multipath routing protocols

The four important issues significant in MANET are Mobility, QoS Provisioning, Multicasting and Security.

1.2.10.1 Mobility

The mobile user can freely move anywhere and are free to join and move away from the network at anytime. The mobile client can explore the area and can form groups or teams to create a taskforce. In the ad hoc network, the mobile client can have individual random and group mobility and the mobility model can have major impact on the selection of a routing scheme and this directly influences the performance. The mobile clients in MANETs have no physical boundary and their location changes as they move around. This movement of mobile nodes makes the network topology highly dynamic as well as causing the intercommunication patterns between nodes to change frequently in an unpredictable manner (Frodigh et al., 2000), (Satyanarayanan, 2001). Thus, an ongoing communication session suffers frequent path breaks. As a result, broadcasting protocols for MANETs must handle mobility management efficiently (Basagni et al., 1998).

1.2.10.2 QoS provisioning

A network or a service provider offers the QoS to be the performance level of services the user in terms of many performance metrics of QoS such as packet delivery, the average end-to-end delay, and available bandwidth. Between the network and the host, negotiation is mostly needed when providing QoS (i.e. QoS provision). More specifically, this demand is based on the call admission control, resource reservation schemes and priority scheduling. Therefore, when different levels of QoS are provided in a highly changeable environment, an important issue takes place for this provision. (Chakrabarti and Mishra, 2001).

In MANETs, the provision of QoS is made to be more difficult than providing it in fixed wired networks. This difficulty is due to a high change in network topology, the presence of additional bandwidth, and medium and linked constraints. Static constraints such as memory, processing power and bandwidth, will be only taken into account (Basagni et al., 2004). An implementation must be performed for an adaptive QoS within the traditional resource reservation techniques (Ilyas, 2003), in order that multimedia services in MANETs could be efficiently supported.

1.2.10.3 Security

Security attacks consider Ad hoc networks to be highly vulnerable to it. In the matter fact, this is taken into account to be as the main challenges of the developers of MANET. Particular security problems are involved in a MANET. This is referred to several reasons, such as insecure operating environment, shared broadcast radio channel, malicious attacks of a neighbor node, lack of central authority, limited availability of resources, lack of association among nodes, and physical vulnerability. Integrity, availability, confidentiality, non-repudiation and authentication are the most common attributes of MANETs security system (Ilyas, 2003) (Makki et al., 2007).

Survivability of network services despite the denial of service attacks is ensured by the Availability. Certain information is never disclosed to unauthorized entities. This is ensured by confidentiality. A corruption is never happened for a message being sent. This is ensured by Integrity. In order to ensure the identity of the peer node for communications, a node is enabled by authentication. Finally, the message being sent cannot be denied by the origin of a message. This is guaranteed by non-repudiation (Buttayan and Hubaux, 2007). The major security threats that are available in MANETs are denial of service, passive eavesdropping, signaling attacks, resource of service, host impersonation and information disclosure.

1.2.10.4 Multicasting

Multicast is another significant issue of MANETs because the multicast tree is not static in MANETs due to the random movement of nodes in the network. Multiple hops are potentially contained by routes of each pair of nodes. The single hop communication type is less complex than this type of communication. When multicast packets should be sent to groups in several networks, multicast routing becomes essentially. In MANETs, a vital role is played by multicasting through several applications such as in emergency, military operations and rescue operations. Node mobility with the power and bandwidth constraints make multicast routing very challenging in MANETs (Ritvanen, 2004).

1.2.11 Application of MANETs

Mobile ad hoc networks (MANETs) are very flexible networks and suitable for a lot of types of potential applications applied on the Ad hoc networks, as they allow the

Applications	The Possible Service of Ad Hoc Networks
Tactical networks	1)Military communication. 2)Military operations. 3)In the battlefields.
Emergency services	1)Search and rescue operations in the desert and in the mountain and so on. 2)Replacement of fixed infrastructure in case of environmental disasters.3)Policing.4)fire fighting.Supporting doctors and nurses in hospitals.
Coverage extension	1)Extending cellular network access. 2)Linking up with the Internet and so on.
Sensor networks	1)Inside the home: smart sensors and actuators embedded in consumer electronics. 2) Body area networks (BAN). 3) Data tracking of environmental conditions such as animal movements, chemical/biological detection.
Education	1)Classrooms. 2)Ad hoc Network when they make a meetings or lectures. 3) Multi-user games. 4) Wireless P2P networking. 5) Outdoor Internet access Robotic pets. 6)Theme parks.
Home and enterpriser	1)Using the wireless networking in Home or office.2) Conferences, 3)meeting rooms. 4)Personal area networks Personal networks.
Context aware services	1)Follow-on services: call-forwarding, mobile workspace. 2)Information services: location specific services, time dependent services. 3)Infotainment: tourists information.
Commercial and civilian environments	1)E-commerce: electronic payments anytime and anywhere. 2)Business: dynamic database access, mobile offices. 3)Vehicular services: road or accident guidance, transmission of road and weather conditions, taxi cab network, inter-vehicle networks. 4)Sports stadiums, trade fairs, shopping malls and so on. 5)Networks of visitors inside the airports

Table 6. Illustrates some of the application for the ad hoc networks

establishment of temporary communication without any pre-installed infrastructure, the application such as the European telecommunications standard institute (ETSI) also the HIPERLAN/2 standard (Masella, 2001) (Habetha et al., 2001), IEEE 802.11 wireless LAN standard family (Crow. B et al., 1997) and Bluetooth (Bluetooth, 2001) the ad hoc network are very important area in this time and very useful for the military (battlefield) and for the disasters (flood, fire and earthquake and so on), meetings or conventions in which people wish to quickly share information (Chlamtac et al., 2003). And then use it in the emergency search-and-rescue operations, recovery, home networking etc. Nowadays, ad hoc network became so important in our circle life, because can be applied anywhere where there is little or without communication infrastructure or may be the existing infrastructure is expensive to use. The ad hoc networking allows to nodes or devices to keep the connections to the network for as long as it's easy to add and to remove to the end of the network. And there are a lot of varieties of applications for the mobile ad hoc networks, ranging large scale such as dynamic network and mobile and small fixed-constrained energy sources. As well as legacy applications that move from the traditional environment to the Ad Hoc infrastructure environments, a great deal of new services can and will be generated for the new environment, finally as the result the mobile Ad Hoc Network is the important technique for the future and to became for the fourth generation (4G), and the main goals for that to provide propagation the computer environments, that support the users to achieved the tasks to get the information and communicate at anytime, anyplace and from any nodes or devices. And now will present some of these practical applications has been arranged in Table 6.

These are many applications on ad hoc networks as we mentioned above and in Table 6 provide an overview of present and future MANET applications. However, the following is a summary of the major applications in MANETs such as tactical networks (military battlefield), home and enterprise network (personal area network) etc.

- Military battlefield, Military equipment currently is equipped with the state of the art computer equipment. Ad hoc networking help the military with the commonplace network technology to maintain information network between military personnel's, vehicles, and military information head quarters. The basic techniques of ad hoc network originated from this field.
- Commercial sector, ad hoc network can be applied in emergency or rescue operations for disaster relief efforts for example in fire, flood, or earthquake and so on. Emergency rescue operations will go to places where communications are impermissible. Therefore proper infrastructure and rapid deployment of a communication network is badly needed. Information is relayed from one rescue team member to another over a small handheld device. Other commercial application includes for instance ship to ship ad hoc mobile communication and so on.
- Local level, ad hoc networks can autonomously link immediate and temporary multimedia network by using notebook or palmtop computers to distribute and allocate information among conference or classroom participants. Besides, it can also be applied for home networks where devices can be link; other examples include taxicab, sports stadium, boat and small aircraft.
- Personal Area Network (PAN), short-range MANET can simplify the intercommunication between a lot of mobile devices such as a PDA, a laptop, and a cellular phone and there are a lot of new devices in this for MANETs. Wired cables can easily be replaced with wireless connections. Ad hoc network enhances the access to the Internet or other networks by means of Wireless LAN (WLAN), GPRS, and UMTS. The PAN is an upcoming application field of MANET for the future computing technology.
- Personal communications (i.e. cell phones, laptops and ear phone).
- Cooperative environments (i.e. meeting rooms, sports stadiums, boats etc.).
- Conferencing (i.e. using mobile nodes).
- Home Network (almost used for PANs).
- Wireless Mesh Networks (very reliable networks that are closely related to MANETs, the nodes of a mesh network generally are not mobile).
- Hybrid Wireless Networks (the goal is to cost savings, enhanced resilience to failures and performance improvements).
- Wireless Sensor Networks (a very active research area of ad hoc networking which includes fixed networks or mobile sensors (Sarkar et al., 2008).

1.2.12 MANET layers

The network architecture can be described using a reference the model. More obviously, the layers of software and hardware are described by this model so that data could be sent among two points, besides, to make it capable for interpellating of multiple devices/applications in a network. In order to increase compatibility in the network between different components from different manufacturers, reference models are required for so (White, 2002). Seven layers are contained in the International Organization for Standardization (ISO/IEC, 2003) which proposed the Open Systems Interconnection (OSI) reference model. In the matter of fact, these layers are ordered from the lowest to the highest layer. The lowest layer represents layer one whereas the highest layer represents layer seven as shown in Figure 9. In other words, these layers are respectively ordered as: application layer, presentation layer, session layer, transport layer, network layer, data link layer and physical layer (from the highest to the

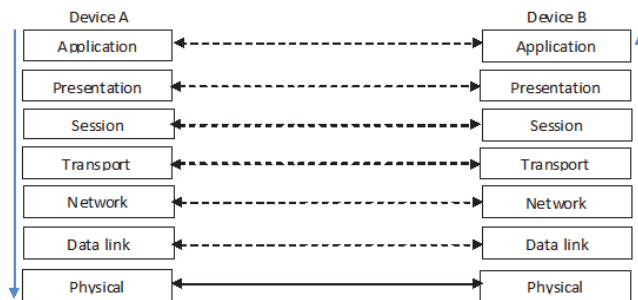


Fig. 9. Illustrates of the original International Organizations for Standardization (ISO) and Open Systems Interconnection (OSI) reference model.

lowest). The transmission of bits is handled by the physical layer through a communications channel. In addition, other physical specifications are taken into account. Such specifications comprise; modulation techniques, connectors and media choice. The access of multiple nodes is coordinated by the functions of data link layer along to a shared medium, control and address information, error detection code, flow control, Medium Access Control (MAC) addressing and so on. Network layer is responsible for creating, maintaining and ending network connection. It transfers a data packet from node to node within the network. In other words, it is responsible for congestion control, IP addressing, and internet working. The transport layer provides an end to end error-free network connection, and makes sure the data arrives at the destination exactly as it left the source. In order to establish sessions between users, the session layer is the layer that controls such a process. At the same time, a series of functions necessary for presenting the data package properly to the sender or receiver are performed by the presentation layer, for example, such as compression and encryption. The application layer is considered to be as the highest layer that provides the user the ability to efficiently access the network. Frequent reconnection and disconnection with peer applications are handled by this layer as a main role of it. Another role of it is to have services and data transmission among users supported, such as, electronic mail and remote file transfer.

1.2.13 Summary

In this chapter, described the necessary an overview for the current literature of Mobile Ad Hoc Network (MANET), covering the main concepts of MANET and the existing wireless mobile network approaches, wireless ad hoc networks, wireless mobile approaches, characteristic, applications, challenges, MANET layers and MANET issues. In particular, mobile ad hoc networks have been classified into two types, MANET and mobile ad hoc sensor network. The traffic types in ad hoc networks which include the Infrastructure wireless LAN and ad hoc wireless LAN are presented in Section 1.2.5. In Section 1.2.6 highlight the relevant details about the ad hoc network routing protocol performance issues. The types of ad hoc protocols such as (Table-driven, On-demand and Hybrid) and Compare between Proactive versus Reactive and Clustering versus Hierarchical are in Section 1.2.7. And Section 1.2.8 respectively. The existing ad hoc protocols are presented in Section 1.2.9. The four important issues significant in MANET are Mobility, QoS Provisioning, Multicasting and Security is presented in Section 1.2.10. Furthermore, in Section 1.2.11 and Section 1.2.12 shows the practical application and the layers of the MANET.

2. Appendix A

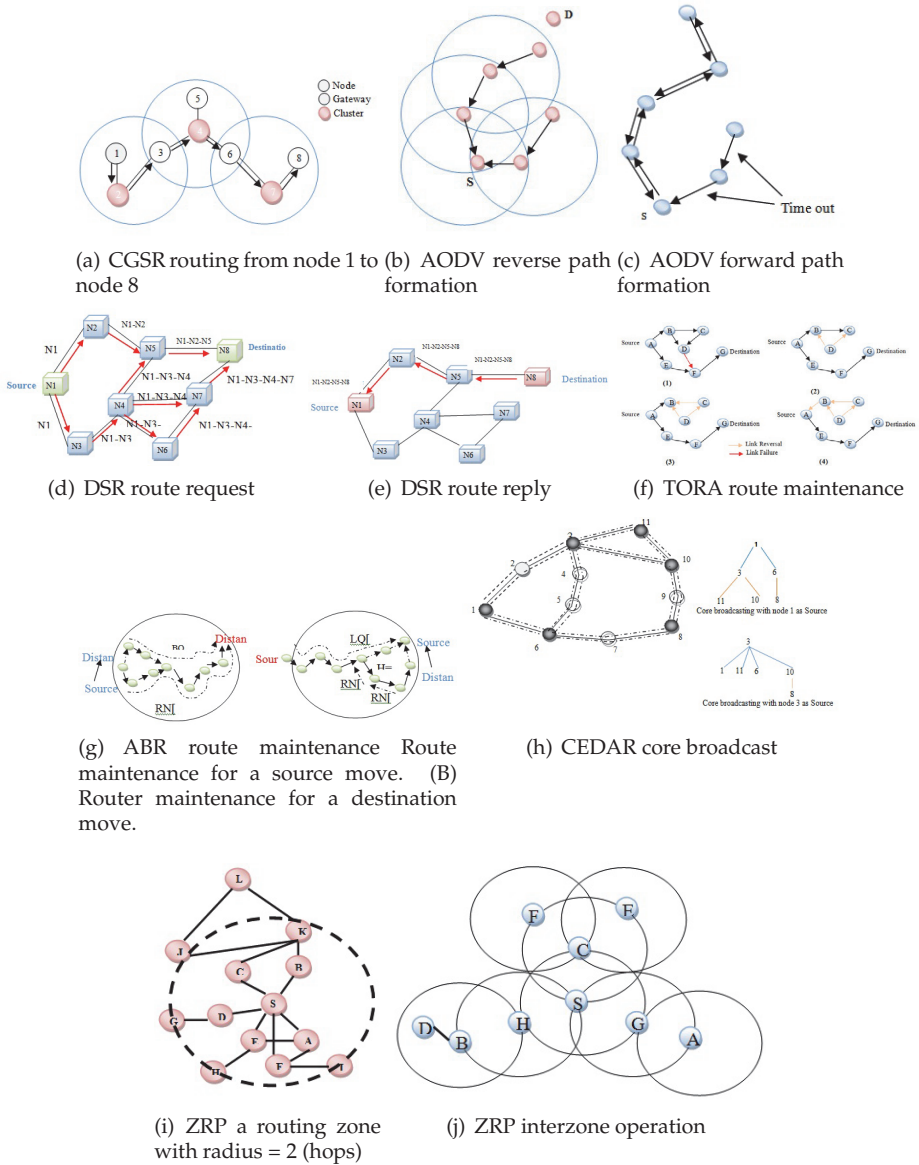


Fig. 10. illustrates the exiting Mobile ad Hoc network Protocols

3. References

- [1] Akkaya, K. and Younis, M. (2005) A survey on routing protocols for wireless sensor networks. In *Ad-hoc Networks* (2005). Vol.3, N0.2, pp. 325-349.
- [2] Akyildiz, I. F., Su, W., Sankarasubramanian, Y. and Cayirci, E. (2002) A Survey on Sensor Networks. *IEEE Communications Magazine* (August 2002), pp.102-114.
- [3] Asif, H. M. (2009) <http://mobius.cs.uiuc.edu/publications/SECON04.pdf>. Computer Engineering Department Saudi Arabia King Fahd University of Petroleum and Minerals.
- [4] Crow, B., Widjaja, I. and Sakai, P. (1997) Investigation of the IEEE 802.11 medium access control (MAC) sublayer functions. In *Proceedings of the INFOCOM '97, Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies.*, pp. 66-43.
- [5] Bakht, H. (2010) WIRELESS INFRASTRUCTURE, Sensor networks and ad-hoc networking. <http://www.computingunplugged.com/issues/issue200410/00001398001.html>, (accessed on 4 Feb 2011).
- [6] Basagni, S., Conti, M., Giordano, S. and Stojmenovic, J. (2004) *Mobile Ad Hoc Networking*, A John Wiley and Sons, Inc., Publication.
- [7] Basagni, S., Chlamtac, I., Syrotiuk, V. R. and Woodward, B. (1998) A distance routing effect algorithm for mobility (DREAM) for wireless networks. *Proceedings 4th Annual ACM/IEEE International Conference on Mobile Computing Networking (MobiCom)*, pp. 76-84.
- [8] Bluetooth (2001) Specification of the Bluetooth system, Core, v1.1. Bluetooth SIG. (<http://www.bluetooth.com>)
- [9] Broch, J., David, A. and David, B. (1998) "A Performance comparison of multi-hop wireless ad hoc network routing protocols". *Proc. IEEE/ACM MOBICOM'98*, pp.85-97.
- [10] Buttyan, L. and Hubaux, J.-P. (2007) "Security and Cooperation in Wireless Networks, Thwarting Malicious and Selfish Behavior in the Age of Ubiquitous Computing", A Graduate Textbook, Available on <http://secowinet.epfl.ch> under a permission from Cambridge University Press, Draft Version 1.3, Feb. 2007.
- [11] Charabarti, S. and Mishra, A. (2001) QoS Issues in Ad Hoc Wireless Networks. *IEEE Communications Magazine*, February 2001.
- [12] Chen, T. and Gerla, M. (1998) Global state routing: A new routing scheme for ad hoc wireless networks. in *Proceedings of IEEE ICC'98*, Vol. 1, No. 7-11, pp.171 - 175.
- [13] Chlamtac, I., Conti, M. and Liu, J. J (2003) Mobile ad hoc networking: imperatives and challenges. *Ad Hoc Networks*, Vol. 1, No. 1, pp. 13-64.
- [14] Estrin, D., Govindan, R. and Hedemann, J. (1999) New Century Challenges: Scalable Coordination in Sensor Networks. *ACM, Mobicom*, 1999.
- [15] Etsi, E. T. (1999) High Performance Radio Local Area Network Type 2 (Hiperlan2), Broadband Radio Access Networks (BRAN) project, 1999. On the URL: <http://portal.etsi.org/archived/radio/hiperlan/hiperlan.asp>.
- [16] Freerbersyser, J. A. and Leiner, B. (2001) A DoD perspective on mobile ad hoc networks. In: Perkins, C. (Ed.) *Ad Hoc Networking*, Addison Wesley, Reading, MA, 2001, pp. 29-51.
- [17] Frodigh, M., Johansson, P. and Larsson, P. (2000) Wireless ad hoc networking: the art of networking without a network *Ericsson Review*. Vol.5, No.4, pp. 248-263.
- [18] Haas, Z. J. and Pearlman, M. R. (2000) The Zone Routing Protocol (ZRP) for Ad Hoc Networks. Internet draft, <http://www.ics.uci.edu/atm/adhoc/papercollection/haas-draft-ietf-manet-zone-zrp-00.txt>.
- [19] Habetha, J., Mangold, S. and Wiegert, J. (2001) 802.11 versus HiperLAN/2 - a comparison of decentralized and centralized MAC protocols for multihop ad hoc radio

- networks. in Proceedings of 5th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, USA, pp. 33-40.
- [20] Hekmat, R. (2006) Ad-hoc Networks: Fundamental Properties and Network Topologies, A book published by Springer.
- [21] IEEE (2004) Std 802.16-2004 TM, IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems.
- [22] IEEE (2005a) Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems. IEEE 806.16e, IEEE P802.16e/D12.
- [23] IEEE (2005b) Std 802.16e-2005TM, IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, Feb 2006.
- [24] IEEE (2011) 802 Working group. [Online] http://www.ieee802.org/IEEE_802_LAN/MAN_Standards_Committee.
- [25] IETF-MANET IETF MANET Working Group. <http://www.ietf.org/html.charters/manetcharter.html>.
- [26] IETF-MANET IETF Working Group: Mobile Adhoc Networks (manet). <http://www.ietf.org/html.charters/manetcharter.html>.
- [27] IETF (1999) RFC 2501 - Mobile Ad hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Considerations. <http://www.faqs.org/rfcs/rfc-sid-x26.html>.
- [28] Ilyas, M. (2003) The Handbook of Ad Hoc Wireless Networks (Electrical Engineering Handbook) [Hardcover], CRC Press 2003.
- [29] ISO/IEC (2003) Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC). Joint Video Team of ITU-T and ISO/IEC JTC 1, JVT G050r1.
- [30] Iwata, A., Chiang, C.-C., PEL, G., Gerla, M. and Chen, T.-W. (2002) Scalable Routing Strategies for Ad Hoc Wireless Networks. IEEE Journal on Selected Areas in Communications, Special Issue on Ad-Hoc Networks, Vol. 17, Issue. 8, pp.1369 - 1379.
- [31] Broch, J, Johnson, D and D. Maltz, T. (2004) The Dynamic Source Protocol for Mobile Ad hoc Networks. <http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-10.txt>, IETF Internet draft, 19 July 2004.
- [32] Jiang, M., Li, J. and Tay, Y. C. (1999) Cluster Based Routing Protocol. August 1999 IETF Draft. <http://www.ietf.org/internet-drafts/draft-ietf-manet-cbrp-spec-01.txt>.
- [33] Johnson, J. B. D. B. and Maltz, D. A. (1999.) The dynamic source routing protocol for mobile ad hoc networks. IETF MANET Working Group, Internet-Draft, October 1999.
- [34] Jubin, J. and Tornow, J. D. (1987) The DARPA Packet Radio Network Protocols. proceedings of the IEEE, January, 1987, vol. 75, no. 1, pp.21-32.
- [35] Saleh Ali AL-OMARI, and Putra Sumari. (2010) AN OVERVIEW OF MOBILE AD HOC NETWORKS FOR THE EXISTING PROTOCOLS AND APPLICATIONS International journal on applications of graph theory in wireless ad hoc networks and sensor networks, Vol. 2, No.1, pp. 87-110.
- [36] LEHR, W. and MCKNIGHT, L. W. (2003) Wireless Internet Access: 3G vs. WiFi? Telecommunication Policy, Vol. 27, No.5, pp. 351-370.
- [37] Leiner, B., Ruth, R. and Sastry. A. R (1996)" Goals and challenges of the DARPA GloMo program. IEEE Personal Communications, December 1996., Vol. 3, No. 6, pp. 34-43.
- [38] Li, X. (2006) Multipath Routing and QoS Provisioning in Mobile Ad hoc Networks. Department of Electronic Engineering, Queen Mary, University of London, PhD thesis.
- [39] Liu, C.-H. and Chang, S.-S. (2009) The study of effectiveness for ad-hoc wireless network. Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Vol. 403, No.6, pp.412-417.

- [40] Makki, S. K., Reiher, P. and Makki, K. (2007) *Mobile and Wireless Network Security and Privacy*, Springer Science and Business Media, LLC, 2007.
- [41] Masella, A. K. A. A. (2001) Serving IP quality of service with Hiper-LAN/2. *Computer Networks: The International Journal of Computer and Telecommunications Networking - Wireless networking*, Vol. 37, issue. 1, pp. 17-24.
- [42] Mbarushimana, C. and Shahrabi, A. (2008) Type of service aware routing protocol in mixed traffic Mobile Ad Hoc Networks. *IEEE International Symposium On Wireless Communication Systems (ISWCS '08)*, Reykjavik pp.677-681.
- [43] Mohapatra, P. and Krishnamurthy, S. V. (2005) *Ad Hoc Networks Technologies and Protocols*, Springer, 2005.
- [44] Morinaga, N., Kohno, R. and Sampei, S. (2002) *Wireless Communication Technologies New Multimedia Systems*, Kluwer Academic Publishers, 2002.
- [45] Murthy, C. S. R. and Mano, B. (2004) *Ad Hoc Wireless Networks: Architectures and Protocols*, Prentice Hall PTR.
- [46] Murthy, S. and Garcia-luna-aceves, J. J. (1996) An efficient routing protocol for wireless networks. *ACM Mobile Networks and Applications Journal*, pp.183-197.
- [47] Maltz, J.B. and D. Johunson, (2005). "Lessons from a full-Scale multi-hop wireless ad hoc network test bed". *IEEE Personal communications magazine*.
- [48] Nicopolitidis, P., Obaidat, M. S., Papadimitriou, G. I. and Pomportsis, A. S. (2003) *Wireless Networks*. John Wiley and Sons, Ltd.
- [49] Pandey, A. K., and Fujinoki, H. (2005). "Study of MANET routing protocols by GlomoSim simulator. *International Journal of Network Management*", November 2005, Vol 15, pp.393 -410.
- [50] Part-11 (1997) IEEE Computer Society LAN MAN Standards Committee, *Wireless LAN medium access control(MAC) and physical layer (PHY) specifications*, IEEE standard 802.11, 1997. The Institute of Electrical and Electronics Engineers, New York, NY.
- [51] Part-12 (1999) IEEE Computer Society. *IEEE standard for information technology telecommunications and information exchange between systems - local and metropolitan networks - specific requirements*.
- [52] Part-16 (2004) IEEE Standard for Local and metropolitan area networks Part 16: *Air Interface for Fixed Broadband Wireless Access Systems*, IEEE Std 802.16-2004, 2004.
- [53] Perkins, C. E., M, E., belding-royer and Das. R. S. (2003) *Ad Hoc On-Demand Distance Vector (AODV) Routing*. <http://www.ietf.org/internetdrafts/draft-ietf-manet-aodv-13.txt>, IETF Internet draft, RFC.
- [54] Perkins, C. E. (2001) *Ad hoc networking: an introduction*, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA 12001, pp. 1 - 28, ISBN: 0-201-30976-9.
- [55] Perkins, C. E. and Bhagwat, P. (1994) Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers. *n Proceedings of ACM SIGCOMM*, pp.234-244.
- [56] Perkins, D. D., Hughes, H. D. and Owen, C. B. (2002) *Factors A ecting the Performance of Ad Hoc Networks*. *IEEE internet computing*, East Lansing, 2002, MI 48824-1226.
- [57] Park, V. and S. Corson. (2001). "Temporally-ordered Routing algorithm (TROA)". *Internet Draft*, draft.ietf-manet-tora-spec-04.txt. July, 2001.
- [58] Ritvanen, K. (2004) *Multicast Routing and Addressing*. Helsinki University of Technology Department of Computer Science and Engineering, A Seminar on Internetworking.
- [59] Sarkar, S. K., Basavaraju, T. G. and Puttamadappa, C. (2008) *Ad Hoc Mobile Wireless Networks: Principles, Protocols, and Applications*, Auerbach Publications Taylor and Francis Group.

- [60] Sinha .P , R.Sivakumar, V. Bharghavan,"CEDAR: a Core-Extraction Distributed Ad hoc Routing algorithm". IEEE INFOCOM'99, Vol 4, No.2, pp. 120-127.
- [61] Satyanarayanan, M. (2001) IEEE Pervasive Computing: Vision and Challenges. Personal Communication, Vol. 8, No. 2, pp. 10 -17.
- [62] White, C. (2002) Data Communications and Computer Networks, Published by Thomson, Third Edition, 2002.
- [63] Wu, S.-L., Yu-chee and TSENG (2007) Wireless Ad Hoc Networking. Auerbach Publications -Taylor and Francis Group.